On the inescapable bias in random forests: sources, manifestations, and corrections

Matthew Berkowitz

mcberko@gmail.com

Simon Fraser University

Rachel MacKay Altman

Simon Fraser University

Thomas M. Loughin

Simon Fraser University

Research Article

Keywords: random forest, quantile regression forest, bias correction, quantile estimation

Posted Date: April 22nd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9431439/v1

Additional Declarations: No competing interests reported.

On the inescapable bias in random forests: sources, manifestations, and corrections

Matthew Berkowitz[1*], Rachel MacKay Altman[1†] and Thomas M. Loughin[1†]

[1*]Statistics and Actuarial Science, Simon Fraser University, 8888 University Dr W, Burnaby, V5A 1S6, BC, Canada.

*Corresponding author(s). E-mail(s): mcberko@gmail.com;
Contributing authors: rachelm@sfu.ca; tloughin@sfu.ca;
†These authors contributed equally to this work.

Abstract

In regression settings, random forests (RFs) often produce unavoidably biased estimates and predictions. We explain sources of bias in terms of RF-based estimated conditional distribution functions (ECDFs). For given covariate values, the RF ECDF is typically based on observations that are not identically distributed, which can produce bias in the ECDF and in mean or quantile estimates. Bias is especially pronounced in sparsely populated regions and when tail quantiles are estimated, as with prediction intervals. We distinguish distal and proximal sources of bias, show how they manifest differently, and explain how tuning parameters and data complexity contribute to ECDF bias. We propose a two-stage bias-correction procedure to reduce bias in the ECDF and in estimates derived from it, including means and quantiles. Using an estimate of the relationship between the RF ECDF and the covariates, we develop a bias adjustment for the entire ECDF and derived estimates. Compared with other procedures, ours was, in the settings considered, more effective at reducing conditional bias in 0.5-quantile estimates while maintaining or reducing MSE. We also show settings where its conditional bias adjustment yields prediction intervals valid over a larger region and/or with less coverage error than other methods.

Keywords: random forest, quantile regression forest, bias correction, quantile estimation

1 Introduction

Random forests (RFs; [1]) are an ensemble method used for regression and classification in which B trees are each grown on a bootstrap resample or subsample of the training data. Each tree is built by recursively splitting the data into ever-smaller subsets, or "nodes", thus partitioning the covariate space. At each potential split, a random subset of the p covariates is drawn as splitting candidates. In the regression setting, the split that minimizes a chosen loss function (typically squared-error) is selected. Splitting continues until terminal nodes are reached. The classical random forest algorithm computes the sample mean of the responses within each terminal node and defines the estimate of the conditional mean at any covariate value as the average of the means from all trees' terminal nodes associated with this value. Equivalently, one can view all the observations in these terminal nodes as representing a "neighbourhood" around the chosen covariate value [2]. The conditional mean estimate is then just a weighted average of all the response values in the neighbourhood. Two key tuning parameters govern forest behaviour: the number of covariates randomly sampled as split candidates at each node, often referred to as `mtry`, and the size of trees to be built, often implemented as the minimum number of observations a node must contain before it can be split, referred to as `nodesize`.

A useful feature of RFs is their built-in validation set defined by out-of-bag (OOB) samples. Because each tree is trained on a bootstrap resample or subsample, a fraction of the training observations are left out of that tree. These OOB observations provide a hold-out set for error estimation, tuning, and, most relevant for this paper, bias correction.

While early RF work focused exclusively on estimating means, [3] observed that, for a given covariate value, one can extract the empirical distribution function (EDF) of the responses in the associated terminal node of each tree. The average of these EDFs then becomes the estimated conditional distribution function (ECDF) for that covariate value. Under certain restrictive regularity conditions, [3] proved that this ECDF consistently estimates the true conditional distribution. He then used the ECDF as the basis for quantile estimation, referring to the technique as a quantile regression forest (QRF). [4] later extended these ideas by growing trees to optimize loss functions beyond squared error while still producing an ECDF that is consistent under regularity conditions.

However, due to the way the ECDF is constructed, RFs produce inescapably biased estimates of various conditional quantities when the sample is finite. This bias is inherent to their construction: observations from regions of covariate space with potentially heterogeneous response distributions are pooled together, causing the ECDF to systematically deviate from the true conditional distribution. As a result, location estimates are pulled toward observations from nearby regions with higher data density, and the dispersion of the true conditional distribution is typically over-estimated. In this paper, our first goal is to explain the fundamental ("distal") source of bias in the ECDF; in other words, what is the root cause of bias in the ECDF, and how is it distinct from more proximal sources of bias? Our second goal is to offer a novel bias-correction procedure that reduces finite-sample bias in RF-based ECDFs.

There is a sparse literature on the finite-sample bias of RF estimates. To our knowledge, four papers have proposed bias-correction methods for RFs — [5–8] — each of which briefly documents bias in RF-based estimates of the mean response conditional on covariates. These papers focus on bias correction and do not go into deep explanations about the sources of bias. Furthermore, bias correction has received little attention in the case where quantiles are estimated using a RF-based ECDF.

Relatedly, various methods have been proposed for producing RF prediction intervals [3, 4, 9, 10]. Prediction intervals based on quantiles of the ECDF [3] have been shown to be inferior to other methods in terms of both validity and width when default tuning parameters are used or when tuned conventionally [8–10], but recent work on tuning has demonstrated that ECDF-based prediction intervals are competitive with these other methods if the RF is appropriately tuned [11]. While tuning can reduce the bias of the *marginal* coverage probability of quantile estimates and associated prediction intervals, to our knowledge, no tuning method has been proposed that corrects the bias of the coverage probability of quantile estimates *conditional* on covariates. The only approach that even gestures at this sort of bias correction is that of [8], who first correct the conditional bias associated with initial RF mean estimates and then construct prediction intervals from these corrected values.

We develop a bias correction procedure that exploits our understanding of the distal source of bias. More specifically, our procedure is based on the probability integral transform and uses the ECDF to derive an estimate of the cumulative probability associated with each response value given any covariate value in the covariate space. Next, we use a second-stage model to predict these cumulative probabilities as a function of the covariates. From this model and the original ECDF, we obtain quantile estimates associated with each of these cumulative probability predictions.

We begin in Section 2 by illustrating and explaining the distal source of bias in the ECDF produced by RFs, then proceed to differentiate between this distal source of bias and more proximal sources of bias that manifest in various ways. In Section 3, we describe in detail our two-stage bias-correction procedure—bias-adjusted random forest (BARF)—to reduce the bias of the ECDF and therefore of quantile estimates. In Section 4, we describe our simulation study that we conducted to illustrate the performance of BARF compared to other bias-correction methods. We also discuss the results. In Section 5, we offer explanations for why BARF excels across diverse conditions, highlight the importance of choosing a suitable second-stage model, and explain why BARF's performance may differ from other methods. We conclude by summarizing our procedure's strengths and limitations and providing avenues for future research opportunities.

2 Finite-Sample Bias of Random Forests

We assume the standard regression framework. Let $\mathbf{x}$ be a realization of $\mathbf{X} = (X_1, \ldots, X_p)$, a p-dimensional random vector denoting the covariates. We assume that $\mathbf{X}$ is defined on a covariate space $\mathcal{X}$, which may consist of both continuous and discrete variables. Conditional on $\mathbf{X} = \mathbf{x}$, we assume $T = g(\mathbf{x}) + \epsilon$, where $\epsilon \sim D$ and

$g(\mathbf{x}) = E[T \mid \mathbf{X} = \mathbf{x}]$. We assume throughout that all observations from this model are drawn independently.

2.1 Illustration of Conditional Bias in the ECDF

We start with a toy example to explain why random forests produce biased ECDFs in most settings. The key point is that the ECDFs are usually based on observations that are not identically distributed.

Consider a single numerical covariate, X, taking on values $\{1, 2, 3, 4, 5\}$. Suppose that the mean of T conditional on $X = x$ is simply $g(x) = x$ and that, given $X = x$, T is observed without error (i.e., $\epsilon = 0$). Now consider a dataset of size $n = 5$ with observations $(t_i, x_i) = (i, i)$, $i = 1, \ldots, 5$. Suppose we fit a RF to these data with a subsampling rate of 0.8 and a minimum terminal node size of 2. Each tree is trained on only 4 out of the 5 observations. Because the minimum terminal node size is 2, each tree partitions its four observations into two leaf nodes, each of size 2.

In this simple setting, we are able to compute the expected value of both the predicted response and the ECDF for any value of x and therefore show the bias in these quantities. Specifically, we can enumerate all five possible trees that arise by letting each of the five observations be OOB once. Table 1 shows these five trees, indicating (i) which observation is OOB, (ii) which observations end up in Node 1 vs. Node 2, (iii) the predicted response for each of $x = 1, 2, 3, 4, 5$ from that tree, and (iv) the mean prediction across the five trees. For $x = 3$, the predicted value can differ depending on whether the split sorts observations according to < 3 or ≤ 3 (denoted 3 and 3^* in the table). Because the trees occur with equal probability in any random forest, the mean predictions represent the expected values of the predicted response for each value of x. The mean predictions differ from the true expected responses and are therefore biased.

OOB	**Nodes**		**Split Rule**	**Tree Prediction**					
obs	Node 1	Node 2	**for Node 1**	$x = 1$	$x = 2$	$x = 3$	$x = 3^*$	$x = 4$	$x = 5$
5	{1,2}	{3,4}	$x < 2.5$	1.5	1.5	3.5	3.5	3.5	**3.5**
4	{1,2}	{3,5}	$x < 2.5$	1.5	1.5	4.0	4.0	**4.0**	4.0
3	{1,2}	{4,5}	$x \leq 3$ or $x < 3$	1.5	1.5	**1.5**	**4.5**	4.5	4.5
2	{1,3}	{4,5}	$x < 3.5$	2.0	**2.0**	2.0	2.0	4.5	4.5
1	{2,3}	{4,5}	$x < 3.5$	**2.5**	2.5	2.5	2.5	4.5	4.5
Mean Prediction				1.8	1.8	2.7	3.3	4.2	4.2

Table 1 The five unique trees that can arise in our toy scenario. Each tree partitions its in-bag observations into two nodes of size 2. The final row shows the mean predictions for each x, while bold entries in the table are the OOB estimates of the expected responses.

When a given x is dropped down each tree in the forest, it lands in a terminal node containing two training observations. The average relative frequency with which each training observation joins x in a terminal node across all trees in the forest comprises the RF estimate of the conditional distribution function, the ECDF, for our example.

Table 2 displays the probability mass that the ECDF assigns to each response value for each x. For example, the ECDF for $x = 1$ puts 0.4 mass on $t = 1$, 0.4 mass on $t = 2$, and 0.2 mass on $t = 3$. This distribution is very different from the true conditional distribution at $x = 1$, which places mass 1 on $t = 1$. This exercise shows that the ECDF is a systematically biased estimate of the true conditional distribution of T and that the direction and severity of this bias depend on the value of x at which the distribution is estimated.

t	$x = 1$	$x = 2$	$x = 3$	$x = 3^*$	$x = 4$	$x = 5$
1	0.4	0.4	0.2	0.1	0.0	0.0
2	0.4	0.4	0.2	0.1	0.0	0.0
3	0.2	0.2	0.4	0.4	0.2	0.2
4	0.0	0.0	0.1	0.2	0.4	0.4
5	0.0	0.0	0.1	0.2	0.4	0.4

Table 2 Probability mass assigned by the ECDF to each response value for each x.

The estimated mean and distributions from the forest at a given $\mathbf{x}$ are biased because terminal nodes contain observations whose conditional distributions differ from that at $\mathbf{x}$. This phenomenon is starkly apparent in our zero-noise scenario, where the true distribution at each x is a point mass at $t = x$. However, the same phenomenon persists in any RF where the ECDF mixes responses corresponding to different values of $\mathbf{x}$, either within terminal nodes or across trees. In that case, each observation's ECDF is an estimate of a mixture of conditional distributions, with weights determined, in part, by the observed values of the covariates and response variables. This phenomenon—where coarse partitions in the covariate space lead to the mixing of responses from different underlying conditional distributions—is the cause of what we refer to as the *distal bias* of RFs. Only when the conditional distributions that contribute to a given ECDF are identical can this mixture result in an unbiased estimate of the true distribution at each x.

We now give a more realistic scenario to illustrate possible manifestations of the conditional bias in the ECDF, where we define conditional bias as the expected difference between an estimator of a population quantity and the true quantity, conditional on the covariates. We simulated 100 datasets of size 200, each comprising independent observations, where $g(\mathbf{x}) = \mathbf{x}^\top \boldsymbol{\beta}$, $\mathbf{X} \sim U[0,1]^{10}$, $\epsilon \sim \mathcal{N}(0,1)$, $\boldsymbol{\beta}$ is chosen so that $R^2 \approx 0.5$ on average across the datasets, and $\mathbf{X}^\top \boldsymbol{\beta}$ has variance 1. We fit RFs using default tuning parameter values on all training sets and evaluated them on a fixed test set of size 1000—simulated with the same distribution as above and kept constant across the 100 datasets.

Figure 1 consists of plots of the RF-estimated densities with the true densities overlaid for the three values of $\mathbf{x}$ in the test set (specifically, those whose respective $g(\mathbf{x})$ values were closest to 2, 0, and -2). See Appendix A for details on how the RF-estimated density was obtained using `R`. The plots reveal that the RF-estimated

conditional densities tend to have greater dispersion than the true densities, highlighting the conditional bias in the ECDFs that stems from mixing different distributions. In addition, the means of the ECDFs for $g(\mathbf{x}) \approx -2$ and $g(\mathbf{x}) \approx 2$ are noticeably biased toward 0, the overall mean value. The reason is that the corresponding values of $\mathbf{x}$ occur in regions with data from distributions that are heterogeneous in an asymmetric manner; i.e., there are more nearby observations from distributions with means closer to 0 than from distributions with means farther away. The phenomena illustrated in Figure 1 occur any time the RF ECDF is forced to aggregate observations that come from different conditional distributions, which is common in the case of continuous covariates.

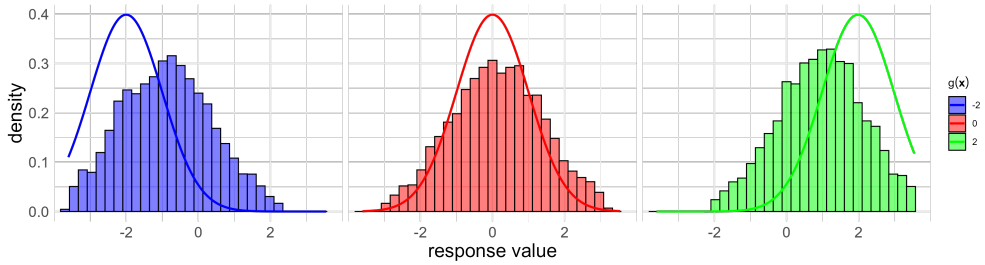

Fig. 1 Random forest density estimates and true densities for $g(\mathbf{x}) \in \{-2, 0, 2\}$

2.2 Distal Bias

For RF-based estimates of the CDF, we distinguish between distal sources of bias—the ultimate, inherent cause of bias—and proximal sources of bias that induce bias through some mechanism that triggers the distal source. We posit that the distal source of RF bias occurs when forests cannot partition the covariate space, $\mathcal{X}$, into neighbourhoods with homogeneous conditional response distributions. This can occur when observations with different covariate values share the same terminal node. It can also occur when observations from terminal nodes of different trees are aggregated across trees in the forest. Hence, this bias can occur even with a terminal node size of 1. For any covariate value $\mathbf{x}_0 \in \mathcal{X}$, the RF-based ECDF puts mass on observed values of T in an adaptively determined neighbourhood of $\mathbf{x}_0$, as described by [2]. This neighbourhood typically comprises covariate values with differing conditional distributions. Hence, the RF-based ECDF at $\mathbf{x}_0$ is an empirical mixture of those differing distributions and is thus biased. This bias is evident in each of the examples above.

The bias of the ECDF at $\mathbf{x}_0$ is influenced by the proximity and distribution of observations in $\mathcal{X}$, the shape of the mean surface in the region around $\mathbf{x}_0$, the amount of noise in the underlying data-generating mechanism, and the sample size. In particular, bias tends to be largest where $g(\mathbf{x})$ changes quickly relative to noise and where there are few nearby observations; conversely, bias tends to be low when $g(\mathbf{x})$ is locally flat and there are many nearby observations. Location bias, in particular, is higher where the

neighbourhood of $\mathbf{x}_0$ is asymmetric in its heterogeneity—i.e., nearby observations have means that lie predominantly above or below $g(\mathbf{x})$—and lower when the distribution of means in the neighbourhood is symmetric. For example, high location bias is evident in the left-hand and right-hand plots of Figure 1, where neighbourhoods nearer the boundary of covariate space contain disproportionately many observations closer to the centre, shifting the ECDFs toward the mean responses associated with more central covariate values. On the other hand, the middle plot shows that, with uniform covariate distributions and a linear response surface, ECDFs associated with central covariate values show no location bias. In addition, the ECDFs are excessively variable, may have the wrong skewness and shape, and so forth. The bias in the ECDFs can lead to bias in other quantities derived from the ECDF. However, the relationship is not always simple. For example, because quantiles are nonlinearly related to the CDF, it is possible to have bias in the quantile coverage probability (and thus prediction interval coverage) computed from the ECDF but comparatively little in the corresponding quantile estimate, or vice versa. Consequently, bias measured on the probability scale can differ in magnitude and even sign from bias measured on the quantile scale.

2.3 Proximal Bias

We next explain and demonstrate RF biases more thoroughly as we break down the various proximal sources of ECDF bias that contribute to the distal source.

2.3.1 Data structure complexity

We operationalize data structure complexity as the degree of heterogeneity among the conditional distributions in regions of covariate space. Several aspects of the data structure drive complexity.

First, adding covariates increases the dimension of $\mathcal{X}$. For a fixed sample size, the expected distance between pairs of covariate points increases as dimension increases, which increases the potential for heterogeneity in their respective distributions. In particular, because different trees are built from different subsets of the data, the nearest observations to $\mathbf{x}_0$ in some subsets might be quite distant in different directions, yet they are pooled into a single ECDF.

Second, the structure of the covariates matters. If many observations have the same covariate value $\mathbf{x}_0$, as in the case of a dataset with categorical covariates and few unique combinations of levels, then the neighbourhood of $\mathbf{x}_0$ will contain just these observations and the ECDF for $\mathbf{x}_0$ will be unbiased. But continuous covariates frequently lead to neighbourhoods containing observations with non-identical conditional distributions, in which case bias is likely unavoidable.

Third, for a fixed sample size, greater variability in $g(\mathbf{x})$ near $\mathbf{x}_0$ results in greater heterogeneity in the conditional distributions of the observations in the neighbourhood of $\mathbf{x}_0$, thus increasing bias in the ECDF.

Overall, a high-complexity data structure is one that leads to substantial local heterogeneity in the conditional distributions of the observations in the neighbourhood of at least one value $\mathbf{x}_0$, whereas a low-complexity data structure is one that leads to

relative homogeneity in the conditional distributions of the observations within every neighbourhood.

Figure 2 illustrates the effect that differing numbers and types of covariates can have on the bias of the ECDF. We consider three settings of differing complexity. For each of the three settings, we simulated 100 datasets of size 1,000 using $g(\mathbf{X}) = \mathbf{X}^\top \boldsymbol{\beta}$ and $\epsilon \sim \mathcal{N}(0, 1)$. The value of $\boldsymbol{\beta}$ was chosen so that $R^2 \approx 0.75$ on average across the datasets within each setting, providing ample opportunity for bias due to the heterogeneity of the response surface within neighbourhoods built by a RF (see Appendix B for more details). In the low-complexity covariate setting, all five covariates had a Bernoulli(0.5) distribution; in the medium- and high-complexity settings, $\mathbf{X} \sim U[0, 1]^5$ and $\mathbf{X} \sim U[0, 1]^{10}$, respectively. Using results in [12] and [11], we set $\texttt{mtry} = \lceil p \cdot R^2 \rceil$ and $\texttt{nodesize} = 3$ to fit RFs. We obtained quantile estimates for coverage probabilities $\tau \in \{0.1, 0.5, 0.9\}$ for each observation in a single test set of size 1,000 where the observations were drawn from the same distributions as those in the training data. For each $\mathbf{x}$ in the test set, we computed the conditional bias estimate for each quantile as the difference between the average estimated and true quantile at that τ. We computed confidence intervals for the biases and plotted LOESS-smoothed curves for these bias estimates and confidence bounds as functions of $g(\mathbf{x})$ (standardized to have mean 0 and variance 1 across all three settings combined) to aid in identifying trends. Figure 2 shows these curves for each complexity level. The horizontal lines represent the estimated marginal bias (average of all estimated conditional biases) for each τ in each setting. Note that the low complexity setting has a more restricted range of $g(\mathbf{x})$ values, hence the truncated LOESS curves.

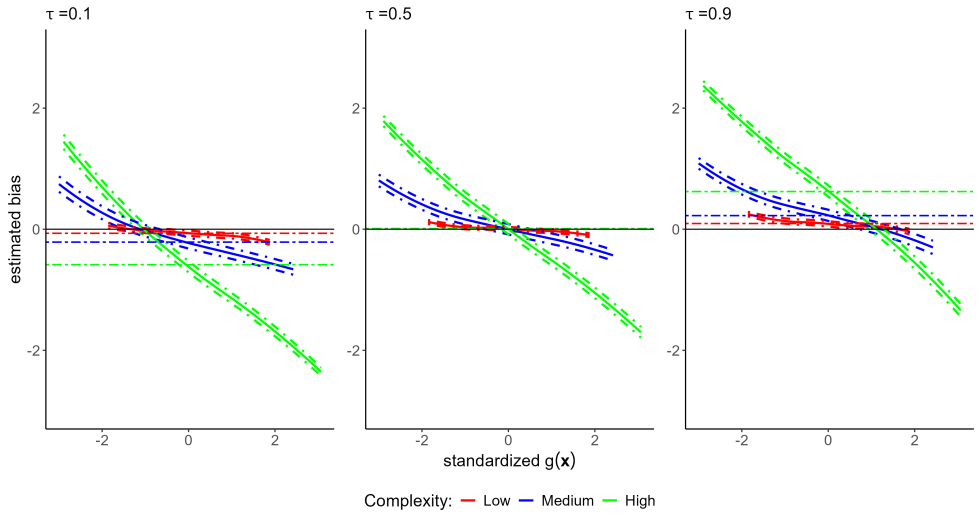

Fig. 2 Estimated bias of RF quantile estimates from three settings, accompanied by 95 % confidence intervals. Horizontal lines represent the estimated marginal bias for each setting.

The plots reveal two key results. First, the biases illustrated in Figure 1 become more pronounced as the covariate structure becomes more complex. Specifically, an increase in complexity leads to more severely biased estimates of quantiles as $\mathbf{x}_0$ moves farther away from the centre of $\mathcal{X}$. In the low-complexity setting, trees can be fully grown because there are only $2^4 = 16$ unique means and sufficient data to be able to separate observations according to their means, thus producing homogeneity within the terminal nodes. Hence, the ECDFs should be unbiased estimates of the true conditional distributions, although the resulting quantile estimates may exhibit slight bias because they are computed from discrete distributions. Second, there is marginal bias in the 0.1- and 0.9-quantile estimates, most pronounced in the high-complexity setting. On average, the lower and upper quantiles are underestimated and overestimated, respectively, because of the excess variability in responses that is captured by ECDFs. For the 0.5-quantile estimates, our symmetric mean structure causes the positive and negative conditional biases to average out, thus preventing marginal bias. The excess variability has little effect on central quantiles in this setting. Both patterns—the relationship between the bias of the estimated quantile and $g(\mathbf{x})$ and the relationship between the marginal bias and τ—are indicative of the distal bias we have articulated. This bias is triggered or exacerbated by the data complexity.

2.3.2 Sample size

Smaller sample sizes can increase the heterogeneity of the conditional distributions in the RF-neighbourhood of $\mathbf{x}_0$. More specifically, for a given distribution of the covariates, smaller sample sizes, in general, cause the neighbourhood around $\mathbf{x}_0$ to encompass observations whose covariate values lie farther from $\mathbf{x}_0$. In other words, all else held constant, increasing the sample size typically leads to less-biased ECDFs. But in our experience, very high sample sizes are often needed to observe meaningful decreases in the bias of quantile estimates and point predictions.

2.3.3 Tuning parameters

Tuning in the RF context typically focuses on `mtry` and `nodesize`. Previous research has indicated a need to tune these parameters according to some loss function to obtain less biased, and/or less variable, mean or quantile estimates [12, 13]. Increasing the value of `nodesize`, in general, results in smaller trees and ECDFs that aggregate more observations with different underlying distributions, potentially leading to more bias in ECDFs. A low `mtry` value can lead to splits on relatively unimportant covariates. Because the sample size limits the number of splits each tree can make, splits on less important covariates leave fewer splits available for more important covariates, resulting in terminal nodes whose observations exhibit greater heterogeneity in their conditional distributions. Setting `mtry` too low can also lead to premature split termination, resulting in shallower trees. With categorical (especially binary) covariates, setting `mtry` $< p$ can lead to the selection of only "pure" covariates—covariates that cannot be split on because all observations in the node contain identical values of that covariate. Many RF algorithms (such as those used in the `ranger` and `randomForestSRC` packages in `R`) terminate in this event, thus preventing further splitting on possibly important covariates. One may question the programming/algorithmic

choice to allow pure covariates to be selected as candidates upon which to make splits. The `randomForestSRC` package authors cite computational burden as one reason for this choice (personal communication).

Tuning `mtry` and `nodesize` can reduce the marginal bias most commonly observed at more extreme quantiles [11] but does not in general reduce the conditional bias of the ECDF at different locations in $\mathcal{X}$. Our proposed bias correction procedure is designed to reduce this conditional bias.

3 Bias Correction Procedure

In this section, we describe our modification to the standard RF, the bias-adjusted random forest (BARF). Our modification is primarily a location shift applied to the quantile function through the ECDF. We describe the BARF procedure in the context of quantile estimation but, in fact, the procedure could be applied to the estimation of any function of the conditional distribution (e.g., the mean).

Let $(t_i, \mathbf{x}_i)$, $i = 1, \ldots, n$ be a random sample from the joint distribution of $(T, \mathbf{X})$. The conditional distribution function of T given covariates $\mathbf{X} = \mathbf{x}$ is denoted by $F_T(t \mid \mathbf{x})$. We obtain the ECDF, $\hat{F}_T$, an estimate of F_T, using RFs as per [3]. For any $\tau \in (0, 1)$ the conditional quantile function for the τ-quantile, given covariates $\mathbf{x}$, is denoted by $q_\tau(\mathbf{x}) = F_T^{-1}(\tau \mid \mathbf{x})$. We define $\hat{q}_\tau^{RF}(\mathbf{x}) = \hat{F}_T^{-1}(\tau \mid \mathbf{x}) = \inf\{t : \hat{F}_T(t|\mathbf{x}) \geq \tau\}$ as the estimate of the τ quantile (the OOB estimate in the case where $\mathbf{x}$ represents an observation from the training set).

BARF directly models the relationship between the coverage probabilities of RF-based quantile estimates and the covariates. It then uses this fitted model as a basis for correcting the ECDF and hence, the quantile estimates. The motivation for BARF is driven by the fact that, by the probability integral transform, $F_T(T|\mathbf{x}_0)$ is uniformly distributed on $(0, 1)$. Thus, if the ECDF at covariate value $\mathbf{x}_0$ is unbiased, its distribution can be approximated by a uniform distribution. Systematic departures from the uniform distribution therefore provide an indication of bias in the ECDF.

To estimate potential bias in the ECDF at a given covariate value $\mathbf{x}_0$, we use the OOB ECDFs associated with the observations in the training set to create pseudo-data $v(t_i, \mathbf{x}_i) = \hat{F}_T(t_i|\mathbf{x}_i)$, i.e., we compute estimated cumulative probabilities corresponding to the observed responses. We then fit a second-stage model to these cumulative probabilities. We primarily consider quantile regression models for this purpose. A key point is that for the uniform distribution, the ζ-quantile equals ζ. Therefore, if the ECDF is unbiased at $\mathbf{x}_0$, then the estimated mean cumulative probability at $\mathbf{x}_0$ based on the fitted second-stage model, $\hat{v}(\mathbf{x}_0)$, will lie at τ. We use the discrepancy from τ as the bias correction to the ECDF at $\mathbf{x}_0$ (on the probability scale). In other words, we define the bias-corrected ECDF at $\mathbf{x}_0$ as the $\hat{v}(\mathbf{x}_0)$-quantile of the original ECDF. For example, if we wanted to estimate the 0.9-quantile of the response at $\mathbf{x}_0$ and $\hat{v}(\mathbf{x}_0) = 0.95$, we would use the 0.95-quantile of the original ECDF at $\mathbf{x}_0$. We also consider second-stage models that describe the *mean* of the cumulative probabilities (e.g., beta regression models). In these cases, we rely on the fact that, if the ECDF is unbiased at $\mathbf{x}_0$, then the mean estimated cumulative probability at $\mathbf{x}_0$ will lie at 0.5. Since the mean and median of a uniform random variable are equal, we can determine

a bias correction for the median estimated cumulative probability at $\mathbf{x}_0$ (as described above) and then apply it to the mean. The benefit of using a quantile regression model over a model for the mean cumulative probability is that the former yields a τ-specific bias correction for each τ of interest, whereas the latter yields a single bias correction (on the probability scale) based on the 0.5-quantile. As a result, the bias adjustments to the ECDF based on quantile regression may be more accurate.

The BARF procedure for estimating a τ-quantile is as follows:

(I) Train the models.
   i. Fit a first-stage model: Train a forest on $\{(t_i, \mathbf{x}_i)\}, i = 1, \ldots, n$.
   ii. For each $\mathbf{x}_i$, compute $v(t_i, \mathbf{x}_i) = \hat{F}_T(t_i|\mathbf{x}_i)$.
   iii. Fit a second-stage model using $v(t_i, \mathbf{x}_i)$, $i = 1, \ldots, n$ as responses. Choose a suitable model (e.g., quantile regression for a specific τ, beta regression, fractional logit, multiple linear regression, random forest, etc.) to describe an attribute of $v(T, \mathbf{x})$ as a function of $\mathbf{x}$.

(II) Compute and apply the bias adjustment for any covariate value $\mathbf{x}_0$.
   i. Use the fitted second-stage model to obtain the estimated cumulative probability $\hat{v}_\tau(\mathbf{x}_0)$.
   ii. Use the fitted first-stage RF to obtain the associated estimated quantile at $\mathbf{x}_0$ corresponding to probability $\hat{v}_\tau(\mathbf{x}_0)$, $\hat{q}_{\hat{v}_\tau(\mathbf{x}_0)}(\mathbf{x}_0)$.
   iii. The final BARF estimate of the τ quantile associated with $\mathbf{x}_0$ is $\hat{q}_\tau^{BARF}(\mathbf{x}_0) = \hat{q}_{\hat{v}_\tau(\mathbf{x}_0)}(\mathbf{x}_0)$. The bias correction is therefore $\hat{b}_\tau(\mathbf{x}_0) = \hat{q}_\tau^{RF}(\mathbf{x}_0) - \hat{q}_{\hat{v}_\tau(\mathbf{x}_0)}(\mathbf{x}_0)$.

Our procedure has several distinguishing features. First, BARF derives each observation's correction by using the OOB ECDFs to estimate bias directly on the probability scale rather than on the response scale. Because this probability-based correction is bounded between 0 and 1, the resulting adjustments vary gradually over the covariate space, mitigating the influence of outliers. Second, our second-stage model is flexible, allowing the use of any model—including parametric models—that seems well-suited to the data. The choice of model can be determined through, for example, knowledge of the underlying data structure or exploratory analysis. (We discuss this issue further in Section 4.5.) In our simulations, we have found that quantile regression is usually a suitable choice due to its ability to tailor the bias correction according to each quantile of interest.

In applications where the RF is used for estimating a parameter other than a quantile (e.g., the mean response), a single correction to the ECDF may suffice. In that case, we could use a second-stage model that describes the mean cumulative probability—which is equivalent to the median if $F_T(T|\mathbf{x}_0)$ is uniformly distributed on $(0, 1)$—to obtain a bias correction for the median that could then be applied to *each* τ-quantile of the ECDF at $\mathbf{x}_0$. Alternatively, we could consider a grid of τ values and use quantile regression at the second stage to obtain τ-specific bias corrections to the ECDF at $\mathbf{x}_0$. However, the latter could result in an adjusted function that is not a valid ECDF, e.g., a function that is not monotonically non-decreasing, and may be undesirable for that reason.

4 Simulation Study

In this section, we describe our simulation study, which we designed to compare bias-adjusted quantile estimates produced by BARF both to the unadjusted estimates ("Default") and to those produced by two other RF-based bias correction methods. We focus on quantile estimation because it allows us to examine biases that impact the ECDF in different ways, as shown in Figure 2. The first of the existing methods, which we call residual forest ("RES"), is a bias-correction technique that uses a two-stage approach applied to residuals [5–7]. The second, which we call "ED", was advanced in a more recent paper and uses an estimate of the conditional prediction error distribution [8].

4.1 Comparator methods

The specific version of RES [5] that we consider requires that a random forest first be trained on the data to produce OOB estimates of the conditional mean responses, denoted by $\hat{t}(\mathbf{x}_i)$, for each training observation $\mathbf{x}_i$, $i = 1, \ldots, n$. Using these estimates, the corresponding OOB residuals (or, more generally, realized prediction errors) are calculated as

$$r(\mathbf{x}_i) = t_i - \hat{t}(\mathbf{x}_i).$$

A second random forest is then fit to these residuals to capture any systematic deviation in the initial predictions. The estimated bias of the conditional mean response for any $\mathbf{x}_0$ is $\hat{b}^{RES}(\mathbf{x}_0) = -\hat{r}(\mathbf{x}_0)$, where $\hat{r}(\mathbf{x}_0)$ is a prediction of the residual error based on the second-stage forest. The bias-corrected estimate of the conditional mean response is then defined as

$$\hat{\phi}^{\text{RES}}(\mathbf{x}_0) = \hat{t}(\mathbf{x}_0) - \hat{b}^{RES}(\mathbf{x}_0).$$

The authors do not propose a specific correction for quantile estimates, so we apply $\hat{b}^{RES}$ to the initial quantile estimates. That is, we use $\hat{q}_\tau^{RES}(\mathbf{x}) = \hat{q}_\tau^{RF}(\mathbf{x}) - \hat{b}^{RES}(\mathbf{x})$.

[8] propose a correction that makes use of the same OOB residuals in a much different way. They define a conditional error distribution given $\mathbf{x}_0$ consisting of probability masses on $r(\mathbf{x}_i)$, $i = 1, \ldots, n$, where the masses depend on $\mathbf{x}_0$. Specifically, they define the estimated conditional prediction error distribution at $\mathbf{x}_0$ as

$$\hat{F}_R(r \mid \mathbf{x}_0) = \sum_{i=1}^{n} w_i(\mathbf{x}_0)\, \mathbb{1}\big(r(\mathbf{x}_i) \leq r\big).$$

The weights $w_i(\mathbf{x}_0)$, $i = 1, \ldots, n$, are the same RF weights used to form the ECDF, determined by the frequency with which $\mathbf{x}_0$ and x_i occur together in the same terminal nodes. The resulting error distribution estimate captures the behaviour of prediction errors for covariate values that are similar to $\mathbf{x}_0$.

The bias of the estimated conditional mean response is estimated from the expected value of the residual at $\mathbf{x}_0$:

$$\hat{b}^{ED}(\mathbf{x}_0) = -\int r \, d\hat{F}_R(r \mid \mathbf{x}_0) = -\sum_{i=1}^{n} w_i(\mathbf{x}_0) \left(t_i - \hat{t}(\mathbf{x}_i)\right).$$

The bias-corrected estimated conditional mean response is given as

$$\hat{\phi}^{ED}(\mathbf{x}_0) = \hat{t}(\mathbf{x}_0) - \hat{b}^{ED}(\mathbf{x}_0),$$

and the bias-corrected estimate of the τ quantile is

$$\hat{q}_\tau^{ED}(\mathbf{x}_0) = \hat{t}(\mathbf{x}_0) + \hat{Q}_E^\tau(\mathbf{x}_0),$$

where $\hat{Q}_E^\tau(\mathbf{x}_0) = \inf\{r : \hat{F}_R(r|\mathbf{x}_0) \geq \tau\}$.

These two comparator methods differ from BARF in the mechanism by which their bias corrections are derived. RES involves fitting a second forest to the OOB residuals and then subtracting predicted residuals from the estimated mean responses. ED involves reusing the RF weights to estimate a conditional error distribution and then obtaining bias estimates for desired quantities. In contrast, BARF involves using each training observation's OOB ECDF to estimate that observation's location on its ECDF , fitting a second-stage model to estimate a quantile of those cumulative probabilities from the covariates, and then inverting the original ECDF at those estimated cumulative probabilities to derive a covariate-specific shift in the estimated quantile function that can be applied to quantile or other estimates.

In our study, we compare the performance of these three bias-correction methods when estimating quantiles.

4.2 Data-generating mechanisms

We tested each method on datasets simulated from 11 different data structures. For all 11 settings, we used the same framework mentioned in Section 2.1, namely

$$T = g(\mathbf{x}) + \epsilon, \quad \epsilon \overset{\text{iid}}{\sim} \mathcal{N}(0, 1),$$

where $\mathbf{x} = (x_1, \ldots, x_p)$ were sampled as $\mathbf{X} \overset{\text{iid}}{\sim} U[0,1]^p$ with $p = 10$ unless otherwise stated. Settings 1–3 are included because they are identical to the homoscedastic additive-error settings tested by [8]; settings 4–11 represent additional, realistic settings that probe different aspects of the RF distal and proximal biases.

Because the signal-to-noise ratio (SNR) was very low for some settings used by [8], we used a higher SNR for settings 4–11 (approximately 1) so that differences among values of $g(\mathbf{x})$ are more clearly expressed relative to noise; keeping the SNR approximately constant across settings 4–11 reduces confounding from unequal noise levels when comparing performance.

The settings were as follows:

1. Linear 1: $g(\mathbf{x}) = x_1$.
2. Step: $g(\mathbf{x}) = 10 \cdot \mathbb{1}(x_1 > 1/2)$.
3. Friedman [14]: $g(\mathbf{x}) = 10\sin(\pi x_1 x_2) + 20(x_3 - \frac{1}{2})^2 + 10x_4 + 5x_5$.
4. Binary: $g(\mathbf{x}) = -x_1 - 0.8x_2 - 0.5x_3 - 0.3x_4 - 0.2x_5 + 0.2x_6 + 0.3x_7 + 0.5x_8 + 0.8x_9 + x_{10}$, where $\mathbf{X} \overset{\text{iid}}{\sim} \text{Bernoulli}[0.5]^p$.
5. Linear 4: $g(\mathbf{x}) = -2x_1 - 1.4x_2 + 1.4x_3 + 2x_4$, where $p = 4$.
6. Linear 10: $g(\mathbf{x}) = c(-1.5x_1 - 1.25x_2 - x_3 - 0.75x_4 - 0.5x_5 + 0.5x_6 + 0.75x_7 + x_8 + 1.25x_9 + 1.5x_{10})$, with $c = 1.03$ to meet the target SNR.
7. Linear 10/5: $g(\mathbf{x}) = c(-1.5x_1 - x_2 - 0.5x_3 + 0.75x_4 + 1.25x_5)$, with $c = 1.47$ to meet the target SNR.
8. Linear 10E: $g(\mathbf{x}) = c(-1.5x_1 - 1.25x_2 - x_3 - 0.75x_4 - 0.5x_5 + 0.5x_6 + 0.75x_7 + x_8 + 1.25x_9 + 1.5x_{10})$, with $c = 0.3$, where $\mathbf{X} \overset{\text{iid}}{\sim} \text{Exp}[1]^p$.
9. Elbow: $g(\mathbf{x}) = 7x_1$.
10. Quadratic: $g(\mathbf{x}) = c(x_1 - 0.5)^2$, with $c = 13.5$.
11. Exponential: $g(\mathbf{x}) = \exp(3 * x_1)/c$, with $c = 5.15$.

4.3 Error metrics

We simulated 1,000 training datasets per simulation setting. Each dataset consisted of $n = 200$ observations, following both [5] and [8]. For each setting, we applied all methods to a single test set of size 1,000. The covariate values in the test set, denoted by $\{\mathbf{x}_i : i = 201, \ldots, 1200\}$, were generated once per setting (using the covariate distribution for that setting) and kept fixed across all 1,000 training datasets.

For each setting, we computed four error metrics: the mean absolute conditional bias (MAB) and mean squared error of conditional biases (MSE) for both the quantile estimates and their coverage probabilities. To obtain coverage probabilities for a given setting, we first computed the estimated conditional bias associated with the τ-quantile estimate for each observation i in the test set, $\hat{q}_\tau(\mathbf{x}_i) - q_\tau(\mathbf{x}_i)$. Let $\tilde{\tau}_i = F_T(\hat{q}_\tau(\mathbf{x}_i)|\mathbf{x}_i)$ denote the true conditional coverage probability associated with the method's τ-quantile estimate. Then $\tilde{\tau}_i - \tau$ is the bias of the conditional coverage probability associated with the quantile estimate for covariates $\mathbf{x}_i$. Estimates of quantile and coverage-probability bias were computed for each x_i in the test set, and their absolute or squared values were averaged across the test set to compute MAB or MSE, respectively. For each prediction interval method we tested, at each x_i in the test set, we first computed the true conditional coverage probability between the method's estimated lower and upper interval bounds. We then calculated the proportion of these interval coverage probabilities that were at least as great as the nominal level.

4.4 Implementation

All simulations were carried out using R, version 4.4.2. We fit RFs using version 0.17.0 of the `ranger` package in `R`. We used the package's default tuning parameter values, `mtry` $= \sqrt{p}$ and `nodesize`=5. We used the `forestError` package to implement the ED bias correction procedure proposed by [8]. We wrote our own R code to implement the BARF and RES procedures.

The default model for the second stage of our procedure was a quantile regression model using the `quantreg` package, resulting in a procedure we refer to as BARF-QR. We fit four additional second-stage models—a generalized additive model (GAM), a RF, a fractional logit (FL) model, and a multiple linear regression (MLR) model—to one dataset from each setting to determine whether another model gave much better results. For the beta regression, FL, and MLR models, we included only first-order terms; for the GAMs, we included only spline terms (no linear terms) for all numerical covariates. We used logit link functions in the beta regression model, FL model, and GAM.

Because of the nonlinear relationship between bias on the quantile scale and bias on the coverage scale, quantile estimates that minimize the former may not minimize the latter. Therefore, in the cases where our goal was to produce quantile estimates with accurate *coverage*—including for prediction intervals—we used a version of BARF-QR that is based on adjusted values of τ [15]. Specifically, we used the value $\tau_{adj} = (\tau - p/2n)(1 - p/n)$ suggested by [15] to correct the bias in the coverage probabilities and obtain better-calibrated prediction intervals. [16] suggested some alternative bias corrections that could be incorporated into BARF-QR, but in practice, we found that the aforementioned transformation produced good estimates without the added computational cost of these alternatives.

4.5 Results

Figures 3–7 display, for select settings, the average estimated conditional biases of the τ-quantile estimate for $\tau \in \{0.1, 0.5, 0.9\}$, along with 95% confidence intervals for the true conditional bias. We computed these quantities by fitting RFs to each setting's simulated training sets. When $g(\mathbf{x})$ is a function of only one covariate, we plot the estimated bias against that covariate. Otherwise, we plot estimated bias against standardized $g(\mathbf{x})$ (a surrogate for relative position in $\mathcal{X}$). For all settings, we use LOESS smoothing of both the estimated bias and the upper and lower confidence limits, thereby producing smooth curves that emphasize trends across standardized $g(\mathbf{x})$. In Setting 1 (Linear 1), we found that using a beta regression model as the second-stage model led to less biased estimates and used it in our BARF models. For all other settings—including all the settings for which we display results—we used a quantile regression model because it led to the least biased estimates. (See Appendix C for one example of the results of this process.) Although our focus is conditional bias, to summarize our results, we also report both MAB and MSE of the quantile estimates and the coverage probabilities produced by each method in each setting. See Tables 3 and 4 for MAB and MSE associated with the 0.5-quantile estimates; see Table D2, Table D3, Table D4, and Table D5 in Appendix D for those associated with the 0.1- and 0.9-quantile estimates.

Across the full set of conditional bias plots (select examples shown here), BARF usually produced 0.5-quantile estimates with the least absolute conditional bias. The associated plots on the coverage probability scale (using BARF-QR with adjusted values of τ) are very similar. For the 0.1- and 0.9-quantile estimates, BARF produced the least biased estimates in the majority of the eleven settings and the second-least biased estimates in the other three settings (where ED produced the least biased

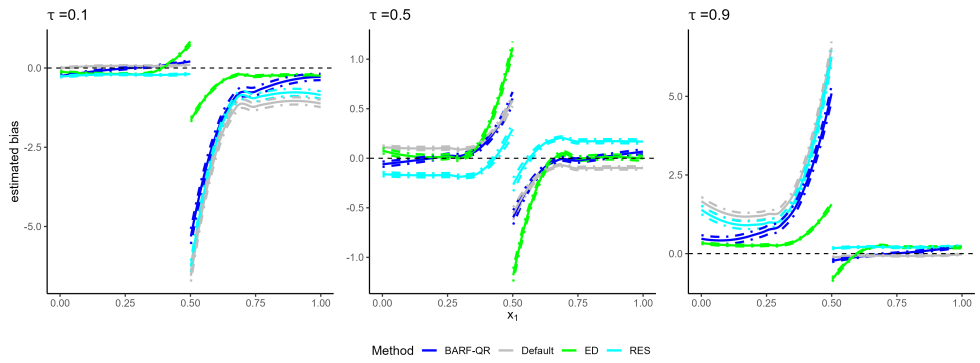

Fig. 3 Setting 2 (Step): Estimated conditional biases of bias-corrected and uncorrected RF quantile estimates

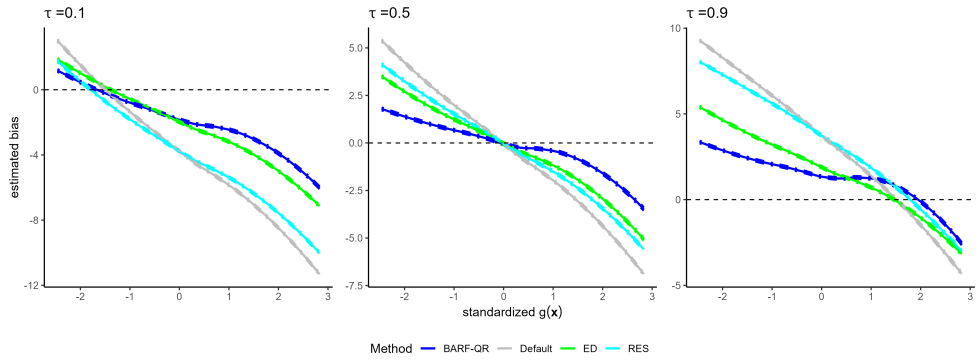

Fig. 4 Setting 3 (Friedman): Estimated conditional biases of bias-corrected and uncorrected RF quantile estimates

estimates). On the coverage probability scale, BARF tail quantile estimates were the least biased in most linear settings (4–8) and was only slightly less biased than the best-performing method (usually either ED or RES) in the other settings. In all linear settings, BARF bias curves for all three values of τ lie close to zero, reflecting relatively accurate quantile estimates over the range of $g(\mathbf{x})$. Ultimately, BARF substantially corrected the pronounced over-estimation of quantiles at the lowest values of $g(\mathbf{x})$ and under-estimation of quantiles at the highest values of $g(\mathbf{x})$.

Table 3 shows that, overall, BARF achieved lower MAB for the 0.5-quantile estimates than did all other estimators in all but one simulation setting (Quadratic), often quite substantially so. In seven settings, BARF also produced either the lowest or second-lowest MSE for the 0.5-quantile estimates. ED produced slightly less biased estimates of the 0.5-quantile estimates than did RES in most settings. In the Step setting, overall, ED and RES actually led to *more* biased 0.5-quantile estimates than did

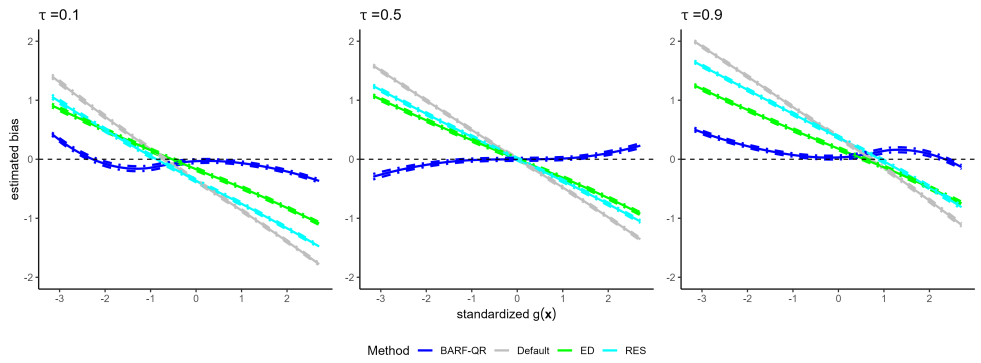

Fig. 5 Setting 6 (Linear 10): Estimated conditional biases of bias-corrected and uncorrected RF quantile estimates

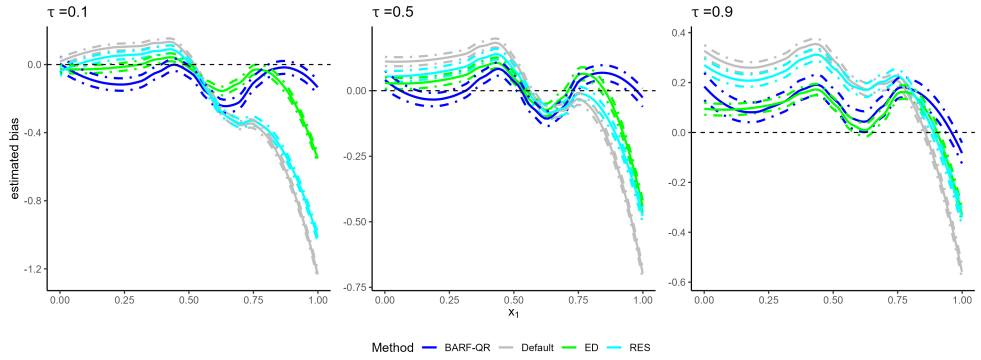

Fig. 6 Setting 9 (Elbow): Estimated conditional biases of bias-corrected and uncorrected RF quantile estimates

the default RFs; the ED estimates in the neighbourhood of $x_1 = 0.5$ were particularly poor. However, ED was much more likely to produce the least biased estimates *away* from the neighbourhood of $x_1 = 0.5$. In all other settings, all bias-correction methods improved upon the default 0.5-quantile estimates. On the coverage scale, the results are very similar. Table 4 reveals that BARF attained the lowest MAB in all but one setting, with values close to 0 for all linear settings.

For the 0.1-quantile estimates, BARF produced the lowest MAB in eight settings (Table D2). For the 0.9-quantile estimate, it produced the lowest MAB in six settings (Table D3). In the remaining settings, the 0.1- and 0.9-quantile BARF estimates were the second-least biased. ED produced the least-biased estimates in these other settings—typically only slightly lower than those produced by BARF. ED also was most likely to produce 0.1- and 0.9-quantile estimates with the lowest MSE, while

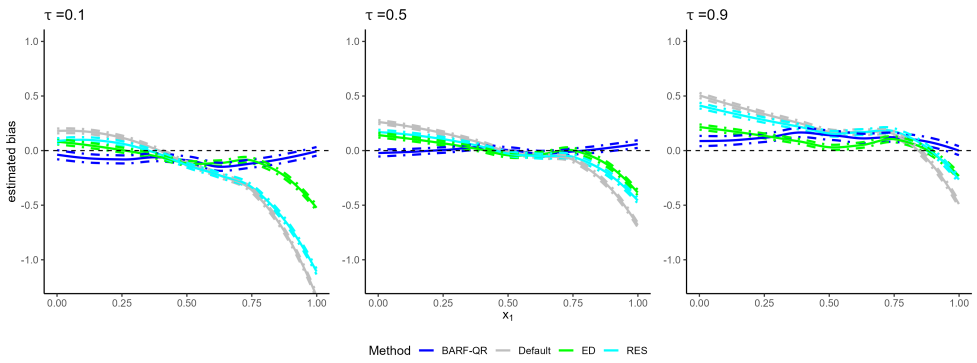

Fig. 7 Setting 11 (Curvature 2): Estimated conditional biases of bias-corrected and uncorrected RF quantile estimates

	MAB				MSE			
Setting	Default	RES	ED	BARF	Default	RES	ED	BARF
Linear 1	0.081	0.063	0.046	0.010	0.080	0.108	0.109	0.117
Step	0.145	0.185	0.215	0.095	0.314	0.368	0.477	0.412
Friedman	1.822	1.475	1.329	1.274	6.249	4.621	3.985	4.062
Binary	0.266	0.159	0.150	0.062	0.291	0.314	0.255	0.261
Linear 4	0.169	0.110	0.094	0.033	0.237	0.267	0.203	0.215
Linear 10	0.394	0.303	0.262	0.031	0.362	0.295	0.272	0.197
Linear 10/5	0.320	0.240	0.189	0.043	0.276	0.238	0.213	0.197
Linear 10E	0.380	0.298	0.265	0.043	0.404	0.344	0.333	0.212
Elbow	0.171	0.116	0.099	0.061	0.147	0.151	0.148	0.209
Curvature 1	0.784	0.783	0.787	0.787	1.026	1.107	1.150	1.123
Curvature 2	0.168	0.114	0.093	0.035	0.153	0.153	0.148	0.200

Table 3 Estimated MAB and MSE for the $\tau = 0.5$ quantile estimates produced by default RFs and the three bias-corrected RFs. For BARF, QR was used as the second-stage model for all settings but Linear 1, for which beta regression was used. For each setting, the minimum error is highlighted red.

BARF led to estimates with MSEs that were either less (or only slightly greater) than those produced by Default.

Prediction Intervals

Figures 8 and 9 display LOESS-smoothed curves for the estimated coverage of two-sided, 80% prediction intervals produced by six methods for eight settings. The first three methods (Default, RES, and BARF) all use QRFs [3] to obtain 0.1- and 0.9-quantile estimates, which we use to form the interval bounds. The fourth, ED, uses quantiles from the estimated conditional prediction error distribution, as discussed in Section 4.1. The fifth and sixth are two additional residual-based alternatives: the

	MAB				MSE			
Setting	Default	RES	ED	BARF	Default	RES	ED	BARF
Linear 1	0.031	0.024	0.017	0.004	0.012	0.015	0.016	0.017
Step	0.048	0.062	0.064	0.026	0.014	0.021	0.032	0.023
Friedman	0.322	0.285	0.269	0.248	0.166	0.153	0.145	0.144
Binary	0.095	0.056	0.054	0.007	0.036	0.038	0.033	0.032
Linear 4	0.061	0.039	0.034	0.012	0.031	0.034	0.027	0.028
Linear 10	0.139	0.108	0.094	0.010	0.043	0.037	0.034	0.026
Linear 10/5	0.116	0.087	0.069	0.016	0.035	0.031	0.028	0.026
Linear 10E	0.129	0.102	0.091	0.015	0.043	0.038	0.037	0.026
Elbow	0.064	0.043	0.037	0.020	0.020	0.021	0.020	0.027
Quadratic	0.243	0.236	0.234	0.229	0.084	0.084	0.084	0.089
Exponential	0.062	0.042	0.035	0.013	0.021	0.021	0.020	0.026

Table 4 Estimated MAB and MSE for the coverage probabilities of the 0.5 quantile estimates produced by default RFs and the three bias-corrected RFs. For BARF, QR with an adjustment to τ (see Section 4.4) was used as the second-stage model for all settings but Linear 1, for which beta regression was used. For each setting, the minimum error is highlighted red.

out-of-bag interval (Res-OOB) method of [9] and the split conformal interval (Res-SC) method of [4]. [10] suggested some additional methods in the context of achieving accurate marginal coverage, but we focused on methods that have been studied in the context of achieving accurate conditional coverage for the purpose of comparison with BARF. We estimated coverage probability by averaging the true coverage probability of each interval across all replicates.

These plots demonstrate that BARF intervals are clearly superior in the linear settings (settings 4–8), the Step setting, and the Friedman setting. In all these settings, the greatest improvements are seen at extreme values of $g(\mathbf{x})$ or x_1, representing locations closer to the boundaries of covariate space. In the other two settings shown, BARF produced intervals with slight under-coverage at extreme covariate values but well-calibrated intervals elsewhere. In the Step setting, BARF intervals had the most accurate coverage over the range of $g(\mathbf{x})$. Moreover, only the QRF-based intervals (BARF, RES, and Default) intervals were valid across the full range of $g(\mathbf{x})$, while other methods produced under-covered intervals, often drastically so at values of $\mathbf{x}$ close to the discontinuity at $x_1 = 0.5$. In the Friedman setting, BARF intervals maintained validity over the widest range of $g(\mathbf{x})$ values but was most conservative (reaching coverage probabilities close to 1) for standardized values of $g(\mathbf{x})$ close to 0. In the Elbow and Exponential settings, most methods produced valid intervals at moderate values of x_1 but produced intervals with under-coverage at more extreme values. In the linear settings (settings 4–8), BARF again produced intervals that maintained validity over the widest range of $g(\mathbf{x})$ values. ED produced intervals with under-coverage across the range of $g(\mathbf{x})$ values in the Binary setting, and both ED and RES produced intervals with under-coverage across the range of $g(\mathbf{x})$ values in the Linear 4 setting.

In setting 1 (Linear 1; not shown in the figures), Res-SC intervals were valid across most of the range of $g(\mathbf{x})$, while most other methods produced intervals with slight under-coverage. In setting 10 (Quadratic; also not shown here), all methods produced

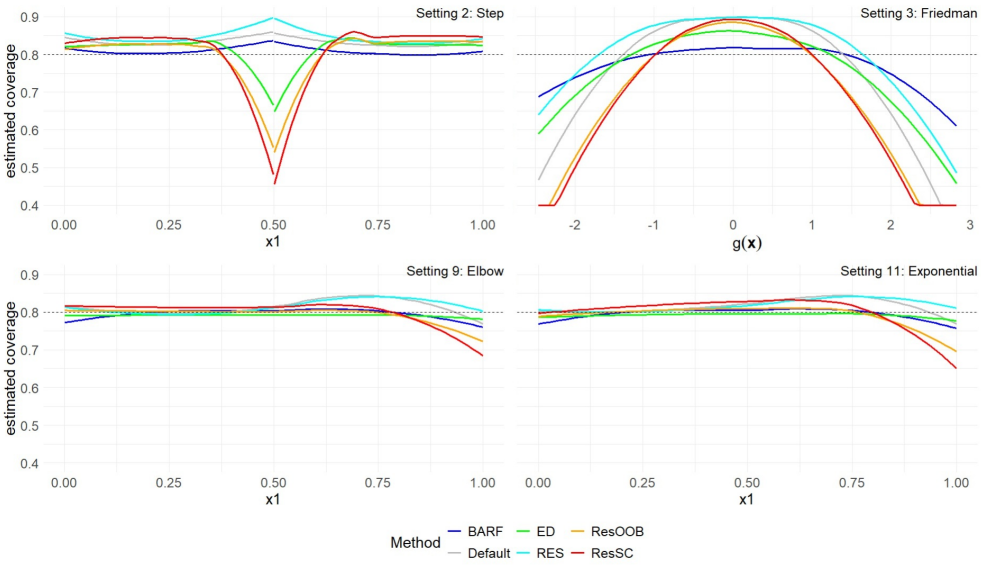

Fig. 8 Estimated conditional coverage probability of 80% prediction intervals using six methods in settings 2, 3, 9, and 11. We again use LOESS-smoothed curves to emphasize trends.

intervals that were valid for only a small proportion of x_1 values, with severe under-coverage for $x_1 < 0.4$ and $x_1 > 0.8$.

5 Discussion

BARF was consistently successful at producing RF quantile estimates with the least conditional bias for all three values of τ in most settings we considered. Moreover, on the coverage probability scale, BARF produced the least-biased 0.5-quantile estimates in nearly every setting. BARF also produced the least biased tail quantile estimates in nearly every linear setting and either the least or second-least biased tail quantile estimates in the other settings. The BARF conditional bias adjustment frequently led to prediction intervals that were valid across the largest range of $g(\mathbf{x})$. Where under-coverage persisted, BARF was usually among the methods that produced prediction intervals with greatest coverage probabilities.

One reason for BARF's relative success is that its bias adjustment at each covariate value involves combining information on both the quantile and coverage probability scales, exploiting the probability integral transform to produce an estimate of bias of the whole ECDF. In contrast, both RES and ED compute bias adjustment entirely on the response scale. Neither uses information about the true coverage probability associated with each covariate's OOB quantile estimate. BARF's use of the coverage probability scale may also be part of why it generally produced close-to-unbiased quantile estimates on the coverage probability scale, especially in the linear settings and in general for $\tau = 0.5$.

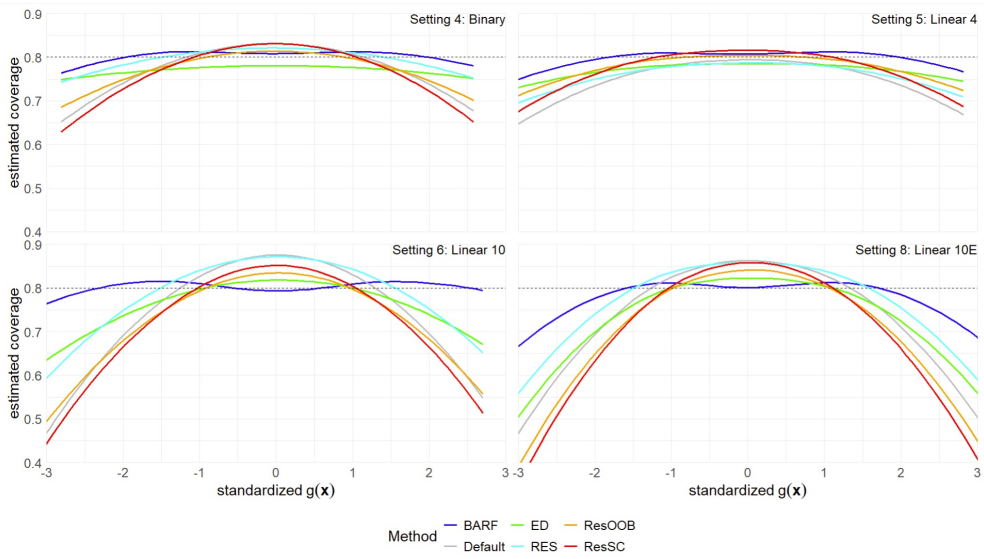

Fig. 9 Estimated conditional coverage probability of 80% prediction intervals using six methods in settings 4, 5, 6, and 8. We again use LOESS-smoothed curves to emphasize trends.

BARF's success hinges on choosing a suitable second-stage model. Failure in this regard—for example, by restricting the second-stage model to another RF when an alternative model may be superior—can lead to biased estimates of the ECDF bias and therefore quantiles. Indeed, RES was initially developed by [5], who used a RF to model the residuals in the second stage. [6] later developed a very similar method using linear regression rather than a RF for the second-stage model. Our exploratory work revealed that quantile regression models with linear effects of covariates are a better default choice for the second-stage model. However, with more complicated data structures that have non-linear and/or interaction terms (like the Friedman function), we have found that fitting a second-stage model that includes spline terms for covariates with apparent non-linear effects produces even better bias-corrected estimates. Given the complexity of the data-generating process in such settings, a more flexible second-stage model is likely more appropriate than the simpler models we tried. See Appendix C for a comparison of second-stage models in one setting. Though the flexibility in choosing the second-stage model proved beneficial in our theoretical setting, choosing an ideal second-stage model in practice could be difficult, particularly if the underlying data structure is very complex.

Other forms of bias have been discussed in the RF context. For example, [17] and [18] noted that the OOB sample may lead to biased estimates of error in classification settings. [11] studied this issue in the regression context, where they developed an estimator of a loss function to quantify error in quantile coverage. They found that the bias of OOB estimates of the loss function was minimal.

Limitations of our work include the modest number of settings in which we tested BARF. Moreover, BARF's tail quantile estimates sometimes exhibited persistent bias

in settings with more complicated data structures; in such settings, finding a suitable second-stage model can be challenging. This latter result presents a future research opportunity: an amended procedure that identifies an adequate second-stage model that ensures that BARF produces close-to-unbiased quantile estimates.

Another avenue for future research involves understanding sources of bias in classification or censored data settings and developing techniques to reduce the overall bias.

In summary, we have proposed a novel bias-adjustment procedure that shows promise for correcting the conditional bias of quantile estimates and improving the conditional coverage probability of prediction intervals.

Declarations

- **Funding:** This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC), grant numbers RGPIN-04304-2018 and RGPIN-2024-05146.
- **Competing interests:** The authors have no relevant financial or non-financial interests to disclose.
- **Ethics approval and consent to participate:** Not applicable.
- **Consent for publication:** Not applicable.
- **Data availability:** No external dataset was analysed in this study. All results were obtained from simulated data generated by the authors' code. The code required to generate the simulated data and reproduce all analyses will be made publicly available upon publication at https://github.com/mcberko.
- **Materials availability:** Not applicable.
- **Code availability:** The code required to reproduce the simulations, analyses, and figures in this paper will be made publicly available upon publication at https://github.com/mcberko.
- **Author contributions:** Matthew berkowitz conceived the study, developed the methodology, carried out the simulations and analyses, and drafted the manuscript. Thomas M. Loughin contributed to the methodological development and manuscript revision. Rachel MacKay Altman contributed to the study design, interpretation of results, and manuscript revision. All authors read and approved the final manuscript.

Appendix A Extracting and using the ECDF

Because the `ranger` package allows the user to extract estimated quantiles (but not the probability masses that form the ECDF) from a fitted forest object, we had to write code to construct the ECDF by inverting the estimated quantile function. To do so, for each x_i in the training data, we created a fine grid of OOB quantile estimates using $\tau_s = \{0.01, 0.02, \ldots, 0.99\}$, thus yielding an ECDF. Then we used this ECDF to obtain the estimated cumulative probabilities $\hat{v}(t_i, \mathbf{x}_i)$. We used these probability values as the response in the second-stage model in the BARF procedure.

To implement the BARF procedure, we wrote a second interpolation function that similarly uses a fine grid of quantile estimates to determine $\hat{q}_{\tilde{v}(\mathbf{x}_i)}(\mathbf{x}_i)$, the quantile corresponding to $\tilde{v}(t_i, \mathbf{x}_i)$ for each observation in the test set.

To produce Figure 1, we selected the test observations with covariate values closest to the three $g(\mathbf{x})$ targets and extracted their corresponding averaged conditional quantile estimates. For each observation in the test set, we first obtained estimated conditional quantiles from each of the 100 RFs on the same fine grid of probabilities as above. We then averaged those quantile estimates across the 100 RFs to form a single averaged conditional quantile function per test observation. An approximate inverse cumulative distribution function was then constructed by interpolating these quantile estimates, yielding a continuous function with which to estimate the full distribution. To visualize the RF-based density estimates, 10,000 observations were generated from the estimated quantile function using the generalized inverse transform method.

Appendix B Linear predictors used in Section 2.3.1

The three complexity levels in Figure 2 are defined by the following linear predictors in Table B1.

Complexity	$\boldsymbol{\beta}$ vector
Low	$\boldsymbol{\beta}_l = 0.48 \times (5,\ 4,\ -2.5,\ -2.3)$
Medium	$\boldsymbol{\beta}_m = 0.83 \times (5,\ 4,\ 2,\ -1.5,\ -2.5)$
High	$\boldsymbol{\beta}_h = 0.56 \times (5,\ 4,\ -2.5,\ -4,\ -5,\ 0.5,\ 1.5,\ -3,\ 3,\ 2.5)$

Table B1 Coefficient vectors for linear predictors used in Section 2.3.1 to produce Figure 2.

Appendix C Comparison of second-stage models

Figure C1 compares second-stage models for computing bias-adjusted quantiles using BARF in Setting 7: Linear 10/5 for $\tau \in \{0.1, 0.5, 0.9\}$. The five models considered are quantile regression (QR), beta regression, generalized additive model (GAM), RF, fractional logit (FL), and multiple linear regression (MLR). For the beta regression, QR, FL, and MLR models, we included only first-order terms; for the GAMs, we included only spline terms for all covariates. We used logit link functions for the beta regression models, FL models, and GAMs. We then computed the estimated conditional biases of the τ-quantile estimates produced by the five models—as well as that of the default (uncorrected) RF estimates.

The plot demonstrates that quantile regression led to the least bias, followed by beta regression and FL. The other models were not competitive. The other linear settings led to similar results. Quantile regression models were also the best-performing second-stage models in all other settings except the Linear 1 setting, in which beta regression led to the least biased bias-corrected quantile estimates.

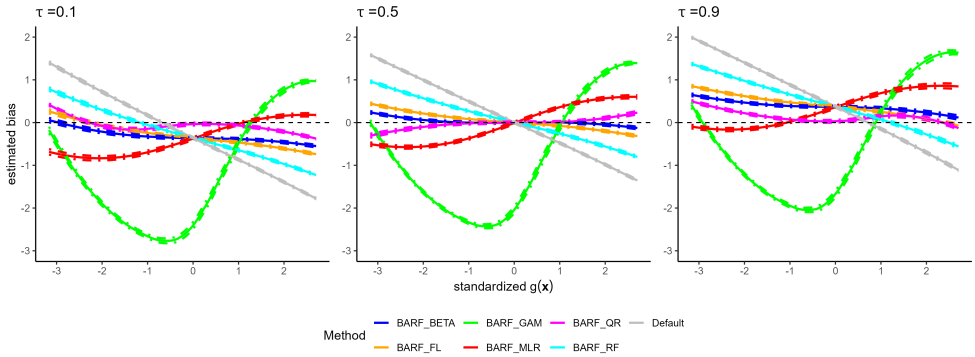

Fig. C1 Comparison of the estimated conditional biases of bias-corrected RF quantile estimates in Setting 7: Linear 10/5 using six different second-stage models: beta regression, QR, GAM, RF, FL, MLR. We also include 95% confidence intervals for the true bias. As in Figures 3–7, we use LOESS smoothing of both the bias and its upper and lower confidence limits.

Appendix D Additional results

	MAB				MSE			
Setting	Default	RES	ED	BARF	Default	RES	ED	BARF
Linear 1	0.101	0.085	0.063	0.054	0.114	0.141	0.136	0.164
Step	1.011	0.946	0.322	0.714	6.564	6.076	0.607	5.649
Friedman	3.809	3.716	2.140	2.017	21.033	19.121	8.010	7.944
Binary	0.340	0.254	0.175	0.127	0.382	0.370	0.336	0.383
Linear 4	0.223	0.175	0.129	0.152	0.314	0.328	0.272	0.313
Linear 10	0.510	0.445	0.294	0.107	0.557	0.481	0.326	0.340
Linear 10/5	0.445	0.387	0.221	0.117	0.450	0.401	0.259	0.343
Linear 10E	0.478	0.421	0.310	0.134	0.580	0.518	0.395	0.378
Elbow	0.333	0.284	0.112	0.119	0.364	0.333	0.188	0.392
Quadratic	0.792	0.798	0.801	0.785	0.999	1.030	1.130	1.296
Exponential	0.316	0.267	0.116	0.101	0.339	0.309	0.189	0.373

Table D2 Estimated MAB and MSE for the $\tau = 0.1$ quantile estimates produced by default RFs and the three bias-corrected RFs. For BARF, QR with adjustment for τ (see Section 4.4) was used as the second-stage model for all but Linear 1, for which beta regression was used. For each setting, the minimum error is highlighted red.

References

[1] Breiman, L.: Classification and regression random forests. Machine Learning **45**(1), 5–32 (2001)

	MAB				MSE			
Setting	Default	RES	ED	BARF	Default	RES	ED	BARF
Linear 1	0.109	0.093	0.061	0.054	0.118	0.145	0.136	0.165
Step	0.979	0.921	0.333	0.653	6.247	5.778	0.585	5.048
Friedman	3.881	3.771	2.138	1.838	21.282	19.332	7.769	7.021
Binary	0.330	0.246	0.169	0.133	0.384	0.373	0.339	0.394
Linear 4	0.216	0.166	0.122	0.148	0.312	0.327	0.272	0.316
Linear 10	0.527	0.459	0.304	0.106	0.568	0.491	0.330	0.335
Linear 10/5	0.437	0.381	0.218	0.113	0.441	0.393	0.259	0.342
Linear 10E	0.509	0.451	0.329	0.136	0.635	0.564	0.415	0.369
Elbow	0.292	0.244	0.126	0.141	0.264	0.266	0.183	0.501
Quadratic	0.843	0.820	0.779	0.831	1.332	1.408	1.275	1.554
Exponential	0.270	0.228	0.122	0.130	0.250	0.252	0.182	0.466

Table D3 Estimated MAB and MSE for the $\tau = 0.9$ quantile estimates produced by default RFs and the three bias-corrected RFs. For BARF, QR with adjustment for τ (see Section 4.4) was used as the second-stage model for all but Linear 1, for which beta regression was used. For each setting, the minimum error is highlighted red.

	MAB				MSE			
Setting	Default	RES	ED	BARF	Default	RES	ED	BARF
Linear 1	0.018	0.017	0.008	0.013	0.003	0.004	0.004	0.005
Step	0.040	0.027	0.032	0.021	0.005	0.004	0.010	0.007
Friedman	0.110	0.099	0.104	0.102	0.028	0.019	0.036	0.046
Binary	0.048	0.031	0.028	0.007	0.009	0.008	0.010	0.008
Linear 4	0.034	0.026	0.017	0.009	0.009	0.009	0.007	0.007
Linear 10	0.062	0.051	0.042	0.010	0.010	0.008	0.009	0.008
Linear 10/5	0.054	0.044	0.031	0.010	0.008	0.007	0.007	0.007
Linear 10E	0.059	0.049	0.043	0.018	0.011	0.009	0.012	0.009
Elbow	0.040	0.034	0.018	0.012	0.005	0.005	0.005	0.013
Quadratic	0.131	0.136	0.165	0.167	0.038	0.045	0.075	0.089
Exponential	0.038	0.032	0.017	0.009	0.005	0.005	0.005	0.008

Table D4 Estimated MAB and MSE for the coverage probabilities of the 0.1 quantile estimates produced by default RFs and the three bias-corrected RFs. For BARF, QR with adjustment for τ (see Section 4.4) was used as the second-stage model for all but Linear 1, for which beta regression was used. For each setting, the minimum error is highlighted red.

[2] Lin, Y., Jeon, Y.: Random forests and adaptive nearest neighbors. Journal of the American Statistical Association **101**(474), 578–590 (2006)

[3] Meinshausen, N.: Quantile regression forests. Journal of Machine Learning Research **7**(35), 983–999 (2006)

[4] Athey, S., Tibshirani, J., Wager, S.: Generalized random forests. The Annals of Statistics **47**(2), 1148–1178 (2019)

[5] Zhang, G., Lu, Y.: Bias-corrected random forests in regression. Journal of Applied

	MAB				MSE			
Setting	Default	RES	ED	BARF	Default	RES	ED	BARF
Linear 1	0.020	0.018	0.009	0.014	0.003	0.004	0.004	0.005
Step	0.040	0.027	0.036	0.021	0.005	0.004	0.012	0.008
Friedman	0.113	0.102	0.107	0.105	0.031	0.021	0.039	0.046
Binary	0.047	0.031	0.028	0.007	0.010	0.009	0.010	0.008
Linear 4	0.034	0.026	0.018	0.009	0.009	0.009	0.008	0.007
Linear 10	0.064	0.052	0.042	0.010	0.010	0.008	0.008	0.008
Linear 10/5	0.053	0.044	0.031	0.010	0.008	0.007	0.007	0.008
Linear 10E	0.058	0.049	0.041	0.017	0.010	0.009	0.010	0.009
Elbow	0.037	0.027	0.017	0.013	0.006	0.005	0.005	0.009
Quadratic	0.089	0.084	0.095	0.097	0.015	0.014	0.019	0.025
Exponential	0.035	0.026	0.016	0.011	0.006	0.005	0.005	0.009

Table D5 Estimated MAB and MSE for the coverage probabilities of the 0.9 quantile estimates produced by default RFs and the three bias-corrected RFs. For BARF, QR with adjustment for τ (see Section 4.4) was used as the second-stage model for all but Linear 1, for which beta regression was used. For each setting, the minimum error is highlighted red.

Statistics **39**(1), 151–160 (2012)

[6] Song, J.: Bias corrections for random forest in regression using residual rotation. Journal of the Korean Statistical Society **44**(2), 321–326 (2015)

[7] Ghosal, I., Hooker, G.: Boosting random forests to reduce bias; one-step boosted forest and its variance estimate. Journal of Computational and Graphical Statistics **30**(2), 493–502 (2020)

[8] Lu, B., Hardin, J.: A unified framework for random forest prediction error estimation. Journal of Machine Learning Research **22**(8), 1–41 (2021)

[9] Zhang, H., Zimmerman, J., Nettleton, D., Nordman, D.J.: Random forest prediction intervals. The American Statistician **74**(4), 392–406 (2020)

[10] Roy, H., Larocque, D.: Prediction intervals with random forests. Statistical Methods in Medical Research **29**(1), 205–229 (2020)

[11] Berkowitz, M., Altman, R.M., Loughin, T.M.: Targeted tuning of random forests for quantile estimation and prediction intervals (2025). https://arxiv.org/abs/2507.01430

[12] Mentch, L., Zhou, S.: Randomization as regularization: A degrees of freedom explanation for random forest success. Journal of Machine Learning Research **21**(171), 1–36 (2020)

[13] Berkowitz, M., Altman, R.M., Loughin, T.M.: Random forests for survival data: which methods work best and under what conditionss. The International Journal of Biostatistics **20**(2), 315–345 (2024)

[14] Friedman, J.H.: Multivariate adaptive regression splines. Annals of Statistics **19**(1), 1–67 (1991)

[15] Bai, Y., Mei, S., Wang, H., Xiong, C.: Understanding the under-coverage bias in uncertainty estimation. Advances in Neural Information Processing Systems **34**, 18307–18319 (2021)

[16] Gibbs, I., Cherian, J.J., Candès, E.J.: Correcting the coverage bias of quantile regression. arXiv preprint arXiv:2511.00820 (2025)

[17] Mitchell, M.: Bias of the random forest out-of-bag (oob) error for certain input parameters. Open Journal of Statistics **1**(3), 205–211 (2011)

[18] Janitza, S., Hornung, R.: On the overestimation of random forest's out-of-bag error. PLoS ONE **13**(8), 0201904 (2017)