

Gradient boosting for extreme quantile regression

Jasper Velthoen * Clément Dombry † Juan-Juan Cai ‡
 Sebastian Engelke §

December 22, 2022

Abstract

Extreme quantile regression provides estimates of conditional quantiles outside the range of the data. Classical quantile regression performs poorly in such cases since data in the tail region are too scarce. Extreme value theory is used for extrapolation beyond the range of observed values and estimation of conditional extreme quantiles. Based on the peaks-over-threshold approach, the conditional distribution above a high threshold is approximated by a generalized Pareto distribution with covariate dependent parameters. We propose a gradient boosting procedure to estimate a conditional generalized Pareto distribution by minimizing its deviance. Cross-validation is used for the choice of tuning parameters such as the number of trees and the tree depths. We discuss diagnostic plots such as variable importance and partial dependence plots, which help to interpret the fitted models. In simulation studies we show that our gradient boosting procedure outperforms classical methods from quantile regression and extreme value theory, especially for high-dimensional predictor spaces and complex parameter response surfaces. An application to statistical post-processing of weather forecasts with precipitation data in the Netherlands is proposed.

Keywords: extreme quantile regression; gradient boosting; generalized Pareto distribution; extreme value theory; tree-based methods.

*Department of Applied Mathematics, Delft University of Technology, Mekelweg 4 2628 CD Delft.
 E-mail: j.j.velthoen@tudelft.nl

†Université Bourgogne Franche-Comté, Laboratoire de Mathématiques de Besançon, CNRS UMR 6623,
 F-25000 Besançon, France. E-mail: clement.dombry@univ-fcomte.fr

‡Department of Econometrics and Data Science, Vrije Universiteit Amsterdam, De Boelelaan 1105,
 1081HV, Amsterdam, the Netherlands. E-mail: j.cai@vu.nl

§Research Center for Statistics, University of Geneva, Boulevard du Pont d'Arve 40, 1205 Geneva,
 Switzerland. E-mail: sebastian.engelke@unige.ch

1 Introduction

In a regression setup the distribution of a quantitative response Y depends on a set of covariates (or predictors) $\mathbf{X} \in \mathbb{R}^d$. These predictors are typically easily available and can be used to predict conditional properties of the response variable Y . Machine learning offers a continuously growing set of tools to perform prediction tasks based on a sample $(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_n, Y_n)$ of independent copies of a random vector (\mathbf{X}, Y) . The main objective is usually to predict the conditional mean $\mathbb{E}(Y | \mathbf{X} = \mathbf{x})$, which corresponds to minimizing the squared error prediction loss. While the mean summarizes the behavior of Y in the center of its distribution, applications in the field of risk assessment require knowledge of the distributional tail. For a probability level $\tau \in (0, 1)$, an important quantity is the conditional quantile

$$Q_{\mathbf{x}}(\tau) = F_Y^{-1}(\tau | \mathbf{X} = \mathbf{x}), \quad (1)$$

where $F_Y^{-1}(\cdot | \mathbf{X} = \mathbf{x})$ is the generalized inverse of the conditional distribution function of Y given $\mathbf{X} = \mathbf{x}$. There has been extensive research in statistics and machine learning to adapt mean prediction methods to other loss functions than squared error. For instance, quantile regression relies on minimizing the conditional quantile loss, which is based on the quantile check function [22]. This has been extended to more flexible regression functions such as the quantile regression forest [24] and the gradient forest [1], which both build on the original random forest [4]. Another popular tree-based method in machine learning is gradient boosting by [13]. This versatile method aims at optimizing an objective function with a recursive procedure akin to gradient descent.

Let n denote the sample size and $\tau = \tau_n$ the quantile level. The existing quantile regression methodology works well in the case of a fixed quantile level, or in the case of

a quantile that is only moderately high, that is, $\tau_n \rightarrow 1$ and $n(1 - \tau_n) \rightarrow \infty$ as $n \rightarrow \infty$, meaning that there are sufficient observations above the τ_n level. For more extreme quantiles with $n(1 - \tau_n) \rightarrow c \in [0, \infty)$, the quantile loss function is no longer useful because observations become scarce at that level and extrapolation beyond the range of observed values is needed. Extreme value theory provides the statistical tools for a sensible extrapolation into the tail of the variable of interest Y . For a large threshold u close to the upper endpoint of the distribution of Y , the distribution of the threshold exceedance $Y - u \mid Y > u$ can be approximated by the generalized Pareto distribution (GPD)

$$H_{\gamma, \sigma}(y) = 1 - (1 + \gamma y / \sigma)_+^{-1/\gamma}, \quad y \geq 0, \quad (2)$$

where for $a \in \mathbb{R}$, $a_+ = \max(0, a)$, and $\gamma \in \mathbb{R}$ and $\sigma > 0$ are the shape and scale parameters, respectively.

There are three main streams in the literature focusing on the estimation of covariate dependent extreme quantiles. First, a parametric form (e.g. linear) can be assumed for the conditional quantile function (1) and estimators for extreme quantiles can be derived and studied as in [7]. The second stream uses GPD modeling of exceedances above a high threshold and assumes that the parameters $\sigma(\mathbf{x})$ and $\gamma(\mathbf{x})$ depend on the covariates via parametric or semi-parametric models [9, 32] or generalized additive models [6, 34]. The third stream is a fully non-parametric local approach where local smoothing estimation techniques for the conditional quantile at moderately high levels are applied and then extrapolated to the extreme level. For example, [8] and [30] apply kernel smooting estimation for the conditional tail distribution and the conditional quantile, respectively, and [16] considers a covariate dependent adaption of Weissman's estimation for heavy-tailed data. While linear or additive models are restricted in their modeling flexibility, local smoothing methods are known to be sensitive to the curse of dimensionality and work well

only for a low-dimensional predictor space. To bypass these issues for modern applications with complex data, tree-based methods are attractive due to their modelling flexibility and robustness in higher dimensions. A first contribution to the use of tree-based models in extreme value theory is the generalized Pareto regression tree [12], but a single tree is used resulting with a model with limited flexibility and predictive performance.

Our goal is to estimate the extreme conditional quantile $Q_{\mathbf{x}}(\tau)$ in (1), where the dimension of covariates d is large and the response surface allows for complex non-linear effects. To this end, we build a bridge between the predictive power of tree-based ensemble methods from machine learning and the theory of extrapolation from extreme value theory. Following the second stream of research mentioned above, we model the tail of the conditional distribution of Y given $\mathbf{X} = \mathbf{x}$ using a GPD distribution in (2) with covariate dependent parameters $\gamma(\mathbf{x})$ and $\sigma(\mathbf{x})$. We propose **gbex**, a gradient boosting algorithm to optimize the deviance (negative log-likelihood) of the GPD model, to estimate $\gamma(\mathbf{x})$ and $\sigma(\mathbf{x})$. In each boosting iteration, these parameters are updated based on an approximation of the deviance gradient by regression trees. The resulting model includes many trees and is flexible enough to account for a complex non-linear response surface. The boosting algorithm has several tuning parameters, the most important ones being the number of trees and the tree depth. We show how they can be chosen effectively using cross-validation.

In two numerical experiments we illustrate that, for the task of extremal quantile estimation, our methodology outperforms quantile regression approaches that do not use tail extrapolation [24, 1] and methods from extreme value theory that assume simple forms for $\gamma(\mathbf{x})$ and $\sigma(\mathbf{x})$ such as generalized additive models [34]. As a result, to the best of our knowledge, our gradient boosting is the first method that reliably estimates extreme quantiles in the case of complex predictor spaces and in the presence of possibly high-dimensional

noise variables.

We apply the developed method to forecast the extreme quantiles of daily precipitation in the Netherlands using the output of numerical weather prediction models as covariates. Our diagnostic tools, namely variable importance score and partial dependence plots, are able to identify changes in the tail heaviness of precipitation as seasonality patterns in the shape parameter estimates $\gamma(\mathbf{x})$. We further investigate the contribution of weather prediction model outputs of neighbouring stations to forecasting the extreme precipitation of a specific location.

Our main contribution is methodological and demonstrates that the tree-based modeling of extremes initiated in [12] with a single tree can be extended to a powerful ensemble method thanks to boosting. Algorithm 1 is an adaptation of Friedman’s gradient boosting [13, 14] to the GPD model, with an extra clipping gradient step introduced for numerical stability. The methodology and resulting algorithm are explained in detail for the sake of completeness and pedagogy. Beyond this, the overall procedure in Algorithm 2 combines extreme value theory and machine learning in two ways: we propose an adaptive covariate dependent threshold to define the exceedances that are the input of Algorithm 1 and we introduce the extreme conditional quantile estimator. A general issue with gradient boosting is selection of hyperparameter such as the number of trees and the parameters governing the tree structure. We propose adaptive hyperparameters selection with deviance-based cross-validation. This is not straightforward since the quantity of interest, namely, the extreme conditional quantile, has no clear relationship with deviance. Our choice is due to the fact that the more natural pinball loss used in quantile regression degenerates in the extreme regime $\tau_n \rightarrow 0$. Our simulation studies reveals that deviance-based cross-validation performs well also for extreme quantile estimation (see Figure 2). The asymptotic analysis

of our `gbex` algorithm is challenging and beyond the scope of this paper, mainly due to two issues: the GPD deviance is non convex while all of the existing theory on gradient boosting considers convex loss functions; model misspecification has to be taken into account since the GPD model is only an approximation for the threshold exceedances.

There has been active research on machine learning methods for extremes in parallel to this paper. Extremal random forests [17] are another proposal for tree-based GPD modelling where the localizing weights of a generalized random forest [1] are used. Extreme quantile regression via neural networks is considered in [18, 19]. [17] and [18] provide comparative simulation studies of the different approaches. As pointed out by a referee, another line of research for extremes in complex high-dimensional models consists in dimension reduction techniques as in the single index model for extreme quantile estimation [15].

The paper is organized as follows. Section 2 introduces our methodology and algorithms for extreme quantile regression based on GPD modeling with gradient boosting. Practical questions such as parameter tuning and model interpretation are discussed in Section 3, while Section 4 is devoted to assessing the performance of our method in two simulation studies. The application to statistical post-processing of weather forecasts with precipitation data in the Netherlands is given in Section 5. We conclude the paper with a summary and a discussion of future research directions.

The gradient boosting method is implemented in an R package and can be downloaded from GitHub at <https://github.com/JVelthoen/gbex/>

2 Extreme quantile regression with gradient boosting

2.1 Background on extreme quantile estimation

Extreme value theory provides the asymptotic results for extrapolating beyond the range of the data and statistical methodology has been developed to accurately estimate extreme quantiles. In the simplest case with no covariate, a sample of n independent copies Y_1, \dots, Y_n of the response Y is observed and the goal is to estimate a quantile $Q(\tau_n)$ of Y at an extreme probability level $\tau_n \in (0, 1)$. Here, extreme means that $\tau_n \rightarrow 1$ and $n(1 - \tau_n) \rightarrow c \geq 0$ as $n \rightarrow \infty$, that is, the expected number of observations that exceed $Q(\tau_n)$ does not go to infinity as $n \rightarrow \infty$. Empirical estimation then becomes strongly biased and extrapolation beyond observations is needed. The usual strategy is to use the empirical quantile $Y_{k-n:n}$ as a threshold and to consider exceedances above this threshold. Asymptotic theory assumes that $k = k(n)$, the number of observations above threshold satisfies $k \rightarrow \infty$ and $k/n \rightarrow 0$. Stated differently, $Y_{k-n:n}$ is the empirical quantile at level $\tau_{0,n} = 1 - k/n$. The level $\tau_{0,n}$ is said intermediate as it satisfies $\tau_{0,n} \rightarrow 1$ and $n(1 - \tau_{0,n}) \rightarrow \infty$. These distinctions are particularly important for the asymptotic theory of estimators in extreme value theory [20].

One of the main results for extrapolation in the univariate case is the Pickands–de Haan–Balkema theorem [2, 26], which states that under mild regularity conditions on the tail of the distribution of Y , the rescaled distribution of exceedances over a high threshold converges to the generalized Pareto distribution. More precisely, if y^* denotes the upper endpoint of the distribution of Y then there exist a normalizing function $\sigma(u) > 0$ such that

$$\lim_{u \uparrow y^*} \mathbb{P} \left(\frac{Y - u}{\sigma(u)} > y \mid Y > u \right) = 1 - H_{\gamma,1}(y), \quad y \geq 0, \quad (3)$$

where H is defined in (2), with the convention $H_{0,\sigma}(y) = 1 - \exp(-y/\sigma)$, $y \geq 0$. The shape parameter $\gamma \in \mathbb{R}$ indicates the heaviness of the upper tail of Y , where $\gamma < 0$, $\gamma = 0$ and $\gamma > 0$ correspond to distributions respectively with short tails (e.g., uniform), light tails (e.g., Gaussian, exponential) and power tails (e.g., Student's t).

Moreover, the GPD is the only non-degenerate distribution that can arise as the limit of threshold exceedances as in (3), and therefore it is an asymptotically motivated model for tail extrapolation and high quantile estimation. By the limit relation in (3), for a large threshold u , the conditional distribution of $Y - u$ given $Y > u$ can be approximated by $H_{\gamma,\sigma}$ with $\sigma = \sigma(u)$. The threshold u can be chosen as the quantile $Q(\tau_0)$ of Y for some intermediate probability level $\tau_0 \in (0, 1)$. Inverting the distribution function in (2) provides an approximation of the quantile for probability level $\tau > \tau_0$ by

$$Q(\tau) \approx Q(\tau_0) + \sigma \frac{\left(\frac{1-\tau}{1-\tau_0}\right)^{-\gamma} - 1}{\gamma}. \quad (4)$$

2.2 Setup for extreme quantile regression

We consider here the setting where the response $Y_i \in \mathbb{R}$ depends on covariates $\mathbf{X}_i \in \mathbb{R}^d$ and our goal is to develop an estimator for the conditional quantile $Q_{\mathbf{x}}(\tau)$ defined by (1) at an extreme quantile level $\tau = \tau_n$. For this purpose, exceedances above an intermediate quantile $\tau_0 = \tau_{0,n}$ will be considered; see the beginning of Section 2.1 for a discussion on extreme and intermediate quantiles. Recall that $(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_n, Y_n)$ denote independent copies of the random vector (\mathbf{X}, Y) with $\mathbf{X} \in \mathbb{R}^d$ and $Y \in \mathbb{R}$.

In this setup, the intermediate threshold $Q(\tau_0)$, shape parameter γ and scale parameter σ in (4) may depend on covariates. We therefore assume that the GPD approximation in (3) holds pointwise for any $\mathbf{x} \in \mathbb{R}^d$ with $u(\mathbf{x}) = Q_{\mathbf{x}}(\tau_0)$, $\gamma(\mathbf{x})$ and $\sigma(\mathbf{x})$, where for the scale we

omit the dependence on the intermediate level $u(\mathbf{x})$ for simplicity. The approximation for the extreme conditional quantile becomes

$$Q_{\mathbf{x}}(\tau) \approx Q_{\mathbf{x}}(\tau_0) + \sigma(\mathbf{x}) \frac{\left(\frac{1-\tau}{1-\tau_0}\right)^{-\gamma(\mathbf{x})} - 1}{\gamma(\mathbf{x})}, \quad \tau > \tau_0. \quad (5)$$

The triple $(Q_{\mathbf{x}}(\tau_0), \sigma(\mathbf{x}), \gamma(\mathbf{x}))$ provides a model for the tail (that is above the probability level τ_0) of the conditional law of Y given $\mathbf{X} = \mathbf{x}$. An estimator of conditional extreme quantiles $\hat{Q}_{\mathbf{x}}(\tau)$ is obtained by plugging in estimators $(\hat{Q}_{\mathbf{x}}(\tau_0), \hat{\sigma}(\mathbf{x}), \hat{\gamma}(\mathbf{x}))$ in Equation (5).

In the following we propose estimators for these three quantities. Our main contribution is a gradient boosting procedure for estimation of the GPD parameters $(\sigma(\mathbf{x}), \gamma(\mathbf{x}))$ that allows flexible regression functions with possibly many covariates. For estimation of the intermediate quantile $Q_{\mathbf{x}}(\tau_0)$, any method for (non-extreme) quantile regression can be used and we outline in Section 2.4 how the existing method of quantile random forests can be applied.

2.3 GPD modeling with gradient boosting

In this section we propose the `gbex` algorithm to estimate the GPD parameters $(\sigma(\mathbf{x}), \gamma(\mathbf{x}))$ using gradient boosting to build an ensemble of tree predictors. The algorithm is the standard Friedman's boosting algorithm [13, 14] applied with objective function given by the GPD negative log-likelihood (also called deviance in the machine learning literature). Since the GPD has two parameters, two sequences of trees are needed; this is similar to the strategy for multiclass classification where several sequences of trees are trained to learn the different class probabilities.

Based on Equation (3), the Peaks-over-Threshold approach assumes that, given $\mathbf{X} = \mathbf{x}$, the excess of Y above the threshold $Q_{\mathbf{x}}(\tau_0)$ follows approximately a GPD. In order to

compute the sample of exceedances, we rely on a (non-extreme) quantile regression method providing an estimation of the intermediate quantile function $\mathbf{x} \mapsto \hat{Q}_{\mathbf{x}}(\tau_0)$. Applying this estimated function at the predictor values $\mathbf{X}_1, \dots, \mathbf{X}_n$, we obtain the exceedances above the intermediate threshold defined as

$$Z_i = \left(Y_i - \hat{Q}_{\mathbf{X}_i}(\tau_0) \right)_+, \quad i = 1, \dots, n. \quad (6)$$

Note that $Z_i = 0$ whenever the value Y_i is below threshold. We assume that the intermediate threshold is high enough so that the exceedances can be modeled by the generalized Pareto distribution and the approximation of conditional quantiles (5) is good. Our aim is to learn the conditional parameter $\theta(\mathbf{x}) = (\sigma(\mathbf{x}), \gamma(\mathbf{x}))$ based on the sample of exceedances above the threshold. We apply tree-base gradient boosting [13, 14] and use the GPD deviance (negative log-likelihood) as the objective function to minimize.

In absence of covariates, a standard way of estimating the GPD parameters $\theta = (\sigma, \gamma)$ is the maximum likelihood method [29], which provides asymptotically normal estimators in the unconditional case with $\gamma > -1/2$. The negative log-likelihood, or deviance, for an exceedance Z_i from a GPD distribution with parameters $\theta(\mathbf{X}_i) = (\sigma(\mathbf{X}_i), \gamma(\mathbf{X}_i))$ is given by

$$\ell_{Z_i}(\theta(\mathbf{X}_i)) = \left[(1 + 1/\gamma(\mathbf{X}_i)) \log \left(1 + \gamma(\mathbf{X}_i) \frac{Z_i}{\sigma(\mathbf{X}_i)} \right) + \log \sigma(\mathbf{X}_i) \right] \mathbb{1}_{Z_i > 0}. \quad (7)$$

The gradient boosting algorithm starts with an initial estimate, which is given by the unconditional maximum likelihood estimator, that is

$$\theta_0(\mathbf{x}) \equiv \theta_0 = \operatorname{argmin}_{\theta} \sum_{i=1}^n \ell_{Z_i}(\theta). \quad (8)$$

This initially constant model is then gradually improved in an additive way. The big picture is the following. Starting from the initial constant model $\theta_0 = (\sigma_0, \gamma_0)$, we sequentially

construct a sequence of B pairs of trees. At step b , the goal is to improve the current model $\theta_{b-1}(\mathbf{x}) = (\sigma_{b-1}(\mathbf{x}), \gamma_{b-1}(\mathbf{x}))$ by adding a pair of gradient trees $(T_b^\sigma(\mathbf{x}), T_b^\gamma(\mathbf{x}))$. For $b = 1, \dots, B$:

- i) a subsample is randomly drawn from the set of exceedances $(\mathbf{X}_i, Z_i)_{1 \leq i \leq n}$;
- ii) on this subsample, the residuals (deviance derivatives) with respect to σ and γ are computed;
- iii) two regression trees are fitted on these residuals; they provide two partitions of the feature space into different leaves; on the different leaves, the tree values are modified by line search approximation so as to minimize the deviance;
- iv) the model is updated by adding a shrunken version of these trees.

We now provide mathematical details for each of the different steps. We assume that the reader is familiar with the standard regression tree based on square loss minimization (CART algorithm as in [5] or [21]). For $b = 1, \dots, B$:

- i) A random subset $S_b \subset \{1, \dots, n\}$ of size $[sn]$ is randomly drawn, where the parameter $s \in (0, 1]$ is called the *subsampling fraction*.
- ii) The model residuals are computed on the subsample S_b by

$$r_{b,i}^\sigma = \frac{\partial \ell_{Z_i}}{\partial \sigma}(\theta_{b-1}(\mathbf{X}_i)) \quad \text{and} \quad r_{b,i}^\gamma = \frac{\partial \ell_{Z_i}}{\partial \gamma}(\theta_{b-1}(\mathbf{X}_i)), \quad i \in S_b.$$

The deviance derivatives are provided in Appendix A.

- iii) A pair of regression trees $(T_b^\sigma(\mathbf{x}), T_b^\gamma(\mathbf{x}))$ are fitted, respectively on $(\mathbf{X}_i, r_{b,i}^\sigma)_{i \in S_b}$ and $(\mathbf{X}_i, r_{b,i}^\gamma)_{i \in S_b}$. The tree construction uses the standard CART algorithm [5] based on

recursive binary splitting and provides a partition of the feature space into several rectangles called leaves. Several parameters are involved in the stopping rule: the maximal depths D^σ, D^γ (i.e., the maximum number of splits between the root and a leaf in the tree) and the minimal leaf sizes $L_{\min}^\sigma, L_{\min}^\gamma$ (minimum number of observations in each leaf). The leaves of T_b^σ (resp. T_b^γ) are denoted by $(L_{b,j}^\sigma)_{1 \leq j \leq J_b^\sigma}$ (resp. $(L_{b,j}^\gamma)_{1 \leq j \leq J_b^\gamma}$). Following [13], the regression trees are then modified: the partitions are kept unchanged but the tree values are chosen so as to minimize the deviance. This is done by line search, that is, the updated value $\xi_{b,j}^\sigma$ in leaf $L_{b,j}^\sigma$ is obtained by solving the minimization problem

$$\xi_{b,j}^\sigma = \underset{\xi}{\operatorname{argmin}} \sum_{\mathbf{x}_i \in L_{b,j}^\sigma} \ell_{Z_i}(\theta_{b-1}(\mathbf{X}_i) + \xi e_\sigma), \quad j = 1, \dots, J_b^\sigma, \quad (9)$$

where $e_\sigma = (1, 0)$ gives the direction of the line search corresponding to σ . For the parameter γ , the line search is performed in direction $e_\gamma = (0, 1)$, yielding the value $\xi_{b,j}^\gamma$ in the leaf $L_{b,j}^\gamma$ (same equation with σ replaced by γ everywhere). In practice the line search (9) can be computationally expensive and an approximation is used instead, i.e., a Newton–Raphson step resulting in

$$\tilde{\xi}_{b,j}^\sigma = - \frac{\sum_{\mathbf{x}_i \in L_{b,j}^\sigma} \frac{\partial \ell_{Z_i}}{\partial \sigma}(\theta_{b-1}(\mathbf{X}_i))}{\sum_{\mathbf{x}_i \in L_{b,j}^\sigma} \frac{\partial^2 \ell_{Z_i}}{\partial \sigma^2}(\theta_{b-1}(\mathbf{X}_i))}.$$

For the parameter γ , the line search approximation in leaf $L_{b,j}^\gamma$ yields the value $\tilde{\xi}_{b,j}^\gamma$ (same equation with σ replaced by γ everywhere). The gradient trees are given by

$$T_b^\sigma(\mathbf{x}) = \sum_{j=1}^{J_b^\sigma} \tilde{\xi}_{b,j}^\sigma \mathbb{1}_{\{\mathbf{x} \in L_{b,j}^\sigma\}} \quad \text{and} \quad T_b^\gamma(\mathbf{x}) = \sum_{j=1}^{J_b^\gamma} \tilde{\xi}_{b,j}^\gamma \mathbb{1}_{\{\mathbf{x} \in L_{b,j}^\gamma\}}. \quad (10)$$

iv) The model $\theta_{b-1}(\mathbf{x}) = (\sigma_{b-1}(\mathbf{x}), \gamma_{b-1}(\mathbf{x}))$ is finally updated by

$$\begin{aligned}\theta_b(\mathbf{x}) &= (\sigma_b(\mathbf{x}), \gamma_b(\mathbf{x})) \\ &= (\sigma_{b-1}(\mathbf{x}) + \lambda^\sigma T_b^\sigma(\mathbf{x}), \gamma_{b-1}(\mathbf{x}) + \lambda^\gamma T_b^\gamma(\mathbf{x})),\end{aligned}\quad (11)$$

where the shrinkage parameters $\lambda^\sigma, \lambda^\gamma \in (0, 1)$ are called learning rates. They are used to slow down the dynamics since only a shrunken version of the trees is added to the current model.

The final output for the estimated parameters is the gradient boosting model

$$\hat{\sigma}(\mathbf{x}) = \sigma_0 + \lambda^\sigma \sum_{b=1}^B T_b^\sigma(\mathbf{x}), \quad \hat{\gamma}(\mathbf{x}) = \gamma_0 + \lambda^\gamma \sum_{b=1}^B T_b^\gamma(\mathbf{x}). \quad (12)$$

The algorithm as described above is an adaptation of Friedman's gradient boosting algorithm [13, 14] to the conditional GPD model for exceedances above a threshold. In a first implementation of this algorithm, we could observe a numerical instability due to the fact that the GPD negative log-likelihood is not Lipschitz and that its gradient may explode. For this reason, we introduce *gradient clipping*, a standard trick in machine learning to avoid gradient explosion [27, 23]. This means that we bound the absolute value of the Newton–Raphson step by 1 in order to mitigate the strong influence of extreme observations, leading to

$$T_b^\sigma(\mathbf{x}) = \sum_{j=1}^{J_b^\sigma} \text{sign}(\tilde{\xi}_{b,j}^\sigma) \min(|\tilde{\xi}_{b,j}^\sigma|, 1) \mathbb{1}_{\{\mathbf{x} \in L_{b,j}^\sigma\}} \quad (13)$$

and similarly for $T_b^\gamma(\mathbf{x})$. We observe in practice that gradient clipping results in a more stable algorithm with better performance.

Algorithm 1 summarizes the procedure for GPD modeling of exceedances. In practice, the number of iterations B is an important parameter and its choice corresponds to a

trade-off between bias and variance. The procedure is prone to overfitting as $B \rightarrow \infty$ and cross-validation is used to prevent this by early stopping; see Section 3.1 where we discuss the interpretation of the different tuning parameters and their selection in practice.

Algorithm 1 gbex boosting algorithm for GPD modeling

Input:

- θ_0 : the initial values of the parameters with default value as in (8);
- $(\mathbf{X}_i, Z_i)_{1 \leq i \leq n}$: data sample of exceedances above threshold;
- B : number of gradient trees;
- D^σ, D^γ : maximum tree depth for the gradient trees;
- $\lambda^\sigma, \lambda^\gamma$: learning rates for the update of the GPD parameters σ and γ respectively;
- s : subsampling fraction;
- $L_{\min}^\sigma, L_{\min}^\gamma$: minimum leaf size of the nodes in the trees.

Algorithm: For $b = 1, \dots, B$:

1. Draw a random subsample $S_b \subset \{1, \dots, n\}$ of size $[sn]$.

2. Compute the deviance derivatives on the subsample S_b :

$$r_{b,i}^\sigma = \frac{\partial \ell_{Z_i}}{\partial \sigma}(\theta_{b-1}(\mathbf{X}_i)) \quad \text{and} \quad r_{b,i}^\gamma = \frac{\partial \ell_{Z_i}}{\partial \gamma}(\theta_{b-1}(\mathbf{X}_i)), \quad i \in S_b.$$

3. Fit regression trees T_b^σ, T_b^γ that predict the gradients $r_{b,i}^\sigma$ and $r_{b,i}^\gamma$ as functions of the covariates \mathbf{X}_i on the sample $i \in S_b$; the trees are built with maximal depth (D^σ, D^γ) and minimal leaf size $(L_{\min}^\sigma, L_{\min}^\gamma)$; for the tree values, use the line search approximation with gradient clipping (13).

4. Update the GPD parameters $\theta_b(\mathbf{x}) = (\hat{\sigma}_b(\mathbf{x}), \hat{\gamma}_b(\mathbf{x}))$ with learning rates $(\lambda^\sigma, \lambda^\gamma)$, i.e.,

$$\hat{\sigma}_b(\mathbf{x}) = \hat{\sigma}_{b-1}(\mathbf{x}) + \lambda^\sigma T_b^\sigma(\mathbf{x}) \quad \text{and} \quad \hat{\gamma}_b(\mathbf{x}) = \hat{\gamma}_{b-1}(\mathbf{x}) + \lambda^\gamma T_b^\gamma(\mathbf{x}).$$

Output: Conditional GPD parameters $(\hat{\sigma}(\mathbf{x}), \hat{\gamma}(\mathbf{x})) = (\hat{\sigma}_B(\mathbf{x}), \hat{\gamma}_B(\mathbf{x}))$.

2.4 Extreme quantile regression

The input of Algorithm 1 is the sample of exceedances Z_i defined by (6). The conditional intermediate quantile $\hat{Q}_{\mathbf{X}_i}(\tau_0)$ used in this definition generally also depends on the covariate vector \mathbf{X}_i and needs to be modeled first. For this task, any method for (non-extreme) quantile regression can be used, but we note that the quality of the approximation (5) of the extreme quantile will also depend on the accuracy of the intermediate quantile estimate. Together with the gradient boosting procedure for the GPD parameters in Section 2.3, we obtain an algorithm for extreme quantile prediction. We refer to this algorithm as the `gbex` method. It combines the flexibility of gradient boosting with the extrapolation technique from extreme value theory.

While in principle any quantile regression method can be used for estimation of the conditional intermediate quantiles $\hat{Q}_{\mathbf{X}_i}(\tau_0)$, we propose to use a quantile random forest. The reason for this is three-fold: first it requires no parametric assumptions on the quantile functions; secondly it exhibits good performance for high dimensional predictor spaces; finally it requires minimal tuning for good results. Quantile regression forests were first proposed by [24] using the weights from a standard random forest [4]. The drawback of this method is that the criterion used in recursive binary splitting to build the trees of the random forest is not tailored to quantile regression. [31] therefore define a generalized random forest with splitting rule designed for that specific task, where the splitting criterion is related to the quantile loss function. In our case, we require the estimator of $Q_{\mathbf{x}}(\tau_0)$ at the sample points $\mathbf{x} \in \{\mathbf{X}_1, \dots, \mathbf{X}_n\}$ and we recommend the use of out-of-bag estimation $\hat{Q}_{\mathbf{X}_i}(\tau_0) = \hat{Q}_{\mathbf{X}_i}^{oob}(\tau_0)$. This means that only the trees for which the i th observation is out-of-bag are kept for the quantile estimation at $\mathbf{x} = \mathbf{X}_i$, that is, trees based on sub-samples not containing the i th observation. This is necessary to avoid giving too much weight to

Algorithm 2 gbex algorithm for extreme quantile prediction

Input:

- $(\mathbf{X}_i, Y_i)_{1 \leq i \leq n}$: data sample;
- τ_0 : probability level for the threshold;
- τ : probability level for the prediction such that $\tau > \tau_0$;
- parameters of the gbex boosting algorithm for GPD modeling of exceedances (Algorithm 1).

Algorithm:

1. Fit a quantile regression to the sample $(\mathbf{X}_i, Y_i)_{1 \leq i \leq n}$ that provides estimates $\hat{Q}_{\mathbf{x}}(\tau_0)$ of the conditional quantiles of order τ_0 .
2. Compute the exceedances $Z_i = (Y_i - \hat{Q}_{\mathbf{x}_i}(\tau_0))_+$, $1 \leq i \leq n$.
3. Let $I = \{i : Z_i > 0\}$ be the index set of positive exceedances and run Algorithm 1 on the data set $(\mathbf{X}_i, Z_i)_{i \in I}$ to estimate the GPD parameters $(\hat{\sigma}(\mathbf{x}), \hat{\gamma}(\mathbf{x}))$.

Output: Estimation of the extreme conditional quantile

$$\hat{Q}_{\mathbf{x}}(\tau) = \hat{Q}_{\mathbf{x}}(\tau_0) + \hat{\sigma}(\mathbf{x}) \frac{\left(\frac{1-\tau}{1-\tau_0}\right)^{-\hat{\gamma}(\mathbf{x})} - 1}{\hat{\gamma}(\mathbf{x})}.$$

the i th observation when predicting at $\mathbf{x} = \mathbf{X}_i$.

3 Parameter tuning and interpretation

3.1 Parameter tuning

Our gradient boosting procedure for GPD modelling includes several parameters that need to be tuned properly for good results. We discuss in this section the interpretation of the different parameters and how to choose them. We introduce data driven choices based on cross validation for the most sensitive parameters and suggest sensible default values for the remaining parameters. This concerns the tuning parameters of Algorithm 1 that takes the sample of exceedances as input and we therefore consider cross-validation within the sample of exceedances.

3.1.1 Tree number B

The number of trees is the most important regularization parameter. The boosting procedure starts from a constant model, that is usually an underfit, and adds recursively trees that adapt the model to the data, leading eventually to an overfit.

We recommend repeated K -fold cross-validation based on the deviance for a data driven choice of B . Given a maximal tree number B_{max} and a division of the data set into K folds $\mathcal{D}_1, \dots, \mathcal{D}_K$, we repeatedly run the algorithm with B_{max} iterations on the data with one fold left-out and then compute the deviance on the left-out fold as a function of B . Adding up the deviances for the different folds, we obtain the cross-validation deviance.

More formally, we define

$$\text{DEV}_{CV}(B) = \sum_{k=1}^K \sum_{i \in \mathcal{D}_k} \ell_{Z_i}(\hat{\theta}_B^{-\mathcal{D}_k}(\mathbf{X}_i)), \quad B = 0, \dots, B_{max}, \quad (14)$$

where $\hat{\theta}_B^{-\mathcal{D}_k}$ denotes the model with B trees trained on the data sample with the k th fold \mathcal{D}_k held out. Due to large values of the deviance on extreme observations, the cross-validation deviance is prone to fluctuations with respect to the partition into folds and we therefore recommend repeated cross-validation. A typical choice is $K = 5$ or 10 with 5 repetitions. The choice of B is then the minimizer of the cross-validation deviance.

3.1.2 Tree depth (D^σ, D^γ)

The gradient boosting algorithm outputs a sum of tree functions. The complexity of the model is therefore determined by the depth parameters D^σ and D^γ , also called interaction depths. A zero depth tree corresponds to a constant tree with no split, so that $D^\sigma = 0$ or $D^\gamma = 0$ yield models with constant scale or shape parameters, respectively. Since the extreme value index γ is notoriously difficult to estimate, it is common in extreme value theory to assume a constant value $\gamma(\mathbf{x}) \equiv \gamma$ so that the case $D^\gamma = 0$ is particularly important. A tree with depth 1, also called a stump, makes only one single split on a single variable. As a result, $D^\sigma = 1$ (resp. $D^\gamma = 1$) corresponds to an additive model in the predictors for $\sigma(\mathbf{x})$ (resp. $\gamma(\mathbf{x})$). Trees with larger depth allow to introduce interaction effects between the predictors of order equal to the depth parameter. We refer to [21, Section 10.11] for a more detailed discussion on tree depth and interaction order in gradient boosting.

In practice, the depth parameter is quite hard to tune and we recommend to consider depth no larger than 3, also because interactions of higher order are difficult to interpret.

Based on our experience, sensible default values are $D^\sigma = 2$ and $D^\gamma = 1$. But more interestingly, cross-validation can be used to select the depth parameters. The left panel of Figure 1 shows a typical cross-validation diagnostic in the context of the simulation study detailed in Section 4. Here $B_{max} = 500$ and depths parameter $(D^\sigma, D^\gamma) = (1, 0), (1, 1), (2, 1)$ and $(2, 2)$ are considered. The plot shows that sensible choices are $B \approx 200$ and $(D^\sigma, D^\gamma) = (1, 0)$ or $(1, 1)$ (more details given in Section 4). The histogram in the right panel shows that, depending on the randomly simulated sample, B typically lies in the range $[100, 250]$, where the deviance is relatively flat ($(D^\sigma, D^\gamma) = (1, 0)$ is fixed here).

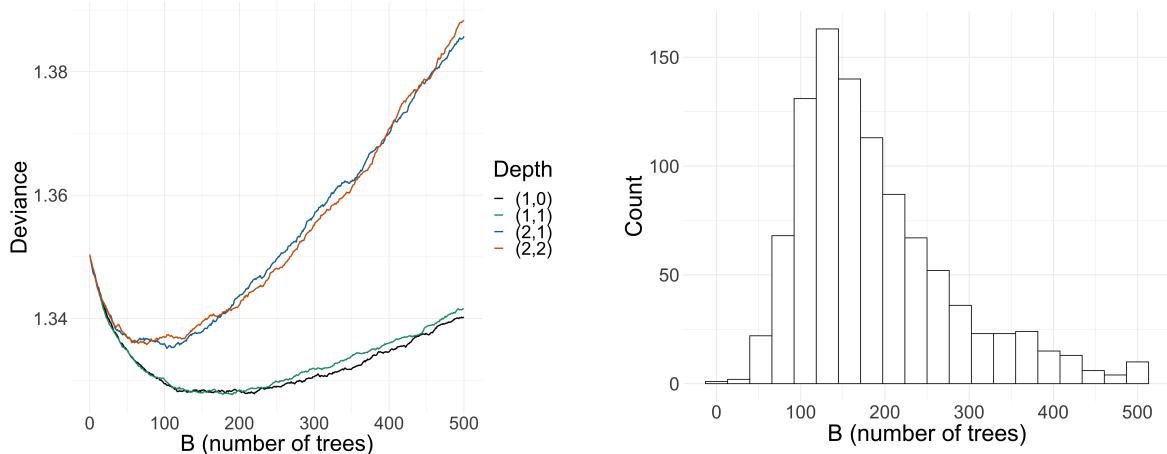


Figure 1: Left panel: cross-validation deviance given by (14) against B for one random sample and depth $(D^\sigma, D^\gamma) = (1, 0), (1, 1), (2, 1)$ and $(2, 2)$. Right panel: selected values of B for 1000 samples when $(D^\sigma, D^\gamma) = (1, 0)$ is fixed. The design of the simulation study is Model 1 described in Section 4.

3.1.3 Learning rates $(\lambda^\sigma, \lambda^\gamma)$

As usual in gradient boosting, there is a balance between the learning rate and the number of trees. As noted in [28], multiplying the learning rate by 0.1 roughly requires 10 times more trees for a similar result. It is common to fix the learning rate to a small value, typically 0.01 or 0.001, and to consider the tree number as the main parameter. Since in our case we have two parallel gradient boosting procedures with different learning rates, we reparameterize them as $(\lambda_{scale}, \lambda_{ratio}) = (\lambda^\sigma, \lambda^\sigma/\lambda^\gamma)$. The balance described above is expressed between B and λ_{scale} and we propose the default $\lambda_{scale} = 0.01$, leaving the number of trees B as the primary parameter. The ratio of the learning rates is important as γ generally requires stronger regularization than σ and ranges on smaller scales. Therefore it is natural to choose $\lambda_{ratio} > 1$.

3.1.4 Remaining tuning parameters

The minimum leaf sizes $L_{\min}^\sigma, L_{\min}^\gamma$ and subsample fraction s play the role of regularization parameters. The minimum leaf size makes sure that the splits do not try to isolate a single high observation of the gradient and that the leaves contain enough observations so that averaging provides a smoother gradient. Subsampling ensures that different trees are fitted on different sub-samples, mitigating the correlation between trees; see [14] and [21, Section 10.12.2] for further discussion on the regularization effect of subsampling. It is common practice that early exploration determines suitable values for these parameters. Depending on the problem and the sample size, we recommend the range $[0.4, 0.8]$ for s and $[10, \frac{n}{50}]$ for L_{\min} .

The parameter τ_0 stands for the probability level of the intermediate quantile used as threshold. Threshold selection is a long standing problem in extreme value theory

[e.g., 11, 10]. A higher threshold yields a better approximation by the GPD distribution but fewer exceedances, leading to reduced bias and higher variance. Some guidelines for threshold selection in practice are provided in Section 5, where we present an application to precipitation forecast statistical post-processing.

3.2 Tools for model interpretation

Contrary to a single tree, boosting models that aggregate hundreds or thousands of trees are difficult to represent but diagnostic plots are available to ease the interpretation. We briefly discuss variable importance and partial dependence plots, which are straightforward modifications to our framework of the tools detailed in [21, Section 10.13].

3.2.1 Variable importance

Boosting is quite robust to the curse of dimensionality and often provides good results even in the presence of high dimensional predictors and noise variables. Understanding which predictors are the most important is crucial for model interpretation. Variable importance is used for this purpose and we discuss here the permutation score and the relative importance.

The permutation score helps to evaluate the impact of a predictor on the model deviance and is not specific to boosting. The relation between a predictor and the response is disturbed by shuffling the values of this predictor and measuring the difference in the deviance before and after shuffling. More precisely, for predictor X_j , we define

$$I(X_j) = \sum_{i=1}^n \ell_{Z_i} \left(\hat{\theta} \left(\mathbf{X}_i^{(j)} \right) \right) - \sum_{i=1}^n \ell_{Z_i} \left(\hat{\theta} \left(\mathbf{X}_i \right) \right), \quad (15)$$

where $\hat{\theta}$ is the estimator given in (12) and $\mathbf{X}_1^{(j)}, \dots, \mathbf{X}_n^{(j)}$ denote the same input vectors as

$\mathbf{X}_1, \dots, \mathbf{X}_n$ except that the j th components are randomly shuffled. A large permutation score $I(X_j)$ indicates a strong effect of X_j in the boosting model. Since the scores are relative, it is customary to assign to the largest the value of 100 and scale the others accordingly.

The relative importance is specific to tree based methods such as boosting or random forests and uses the structure of the trees in the model. It is discussed for instance in [21, Section 10.13.1]. Recall that during the construction of the trees, the splits are performed so as to minimize the residual sum of squares (RSS) of the gradient and each split causes a decrease in the RSS. The more informative splits are those causing a large decrease in the RSS. The relative importance of a given variable X_j is obtained by considering all the splits due to this variable in the sequence of trees, and by summing up the decrease in RSS due to those splits. Because we have two sequences of trees, we compute relative importance of variable X_j in the estimation of σ and γ separately by considering the sequence of trees (T_b^σ) and (T_b^γ) respectively.

3.2.2 Partial dependence plot

Once the most relevant variables have been identified, the next attempt is to understand the dependence between the predictors and the response. Partial dependence plots offer a nice graphical diagnostic of the partial influence of a predictor X_j on the outputs $\hat{\sigma}(\mathbf{x})$, $\hat{\gamma}(\mathbf{x})$ or $\hat{Q}_{\mathbf{x}}(\tau)$; see [21, Section 10.13.2]. The partial dependence plot for $\hat{\sigma}$ with respect to X_j is the graph of the function $x \mapsto \frac{1}{n} \sum_{i=1}^n \hat{\sigma}(\mathbf{X}_i^{-j,x})$, where the vector $\mathbf{X}_i^{-j,x}$ is equal to \mathbf{X}_i except that the j th component has been replaced by x . Notice that dependence between the predictors is not taken into account so that this is not an estimate of $\mathbb{E}[\hat{\sigma}(\mathbf{X}) \mid X_j = x]$, except if X_j is independent of the other predictors. In the particular case when an additive

model is built, i.e., $D^\sigma = 1$, the partial dependence plot with respect to X_j is equal to the effect of the variable X_j up to an additive constant. Partial dependence plots with respect to several covariates can be defined and plotted similarly, at least in dimension 2 or 3.

4 Simulation studies

To demonstrate the performance of our method, we conduct two numerical experiments. We generate n independent samples with d covariates $\mathbf{X} = (X_1, \dots, X_d)$ distributed from an independent uniform distribution on $[-1, 1]^d$, with $(n, d) = (2000, 40)$ or $(5000, 10)$, depending on the complexity of the model. We aim to estimate the conditional quantile function $Q_{\mathbf{x}}(\tau)$ corresponding to extreme probability levels $\tau \in \{0.99, 0.995, 0.9995\}$. We choose the level $\tau_0 = 0.8$ for the intermediate quantile and it is worthwhile to note that the effective sample size $n(1 - \tau_0)$ for the gradient boosting step is then only 400 for $n = 2000$.

The local smoothing based methods mentioned in the introduction [16, 8] become cumbersome in our simulation setting because of the sparsity of data in high dimension. We compare our `gbex` method to two quantile regression approaches, the quantile regression forest (`qrf`) from [24] and the generalized random forest (`grf`) from [1]. Moreover, we consider two existing methods from extreme value theory that use GPD modeling of the exceedances. One is the classical estimator of extreme quantile without using covariates, thus $\gamma(\mathbf{x}) \equiv \gamma$ and $\sigma(\mathbf{x}) \equiv \sigma$, which we call the `constant` method. The other one is the `evgam` method of [34] that assumes generalized additive models for $\gamma(\mathbf{x})$ and $\sigma(\mathbf{x})$.

To evaluate the performance over the full predictor domain $[-1, 1]^d$ we consider the integrated squared error (ISE) defined for a fixed quantile level τ and the i th replication

of the data set by

$$\text{ISE}_i = \int_{[-1,1]^d} \left(\hat{Q}_{\mathbf{x}}^{(i)}(\tau) - Q_{\mathbf{x}}(\tau) \right)^2 d\mathbf{x}, \quad (16)$$

where $\hat{Q}_{\mathbf{x}}^{(i)}(\tau)$ is the quantile estimated from the model. We use a Halton sequence, a low discrepancy quasi-random sequence [e.g., 25, p. 29], in order to efficiently evaluate the high dimensional integral in the ISE computation. Averaging over the $R = 1000$ replications, we obtain the mean integrated squared error (MISE).

Our first model is designed to check robustness of the methods against noise variables. This model is constructed in a similar way as the example studied in [1, Section 5] and it has a predictor dimension of $d = 40$, of which one covariate is signal and the remaining are noise variables.

- **Model 1:** Given $\mathbf{X} = \mathbf{x} \in \mathbb{R}^{40}$, Y follows a Student's t -distribution with 4 degrees of freedom and scale

$$\text{scale}(\mathbf{x}) = 1 + \mathbf{1}(x_1 > 0).$$

This is a heavy-tailed model where the GPD approximation has a constant shape parameter $\gamma(\mathbf{x}) \equiv 1/4$ and the scale parameter is a step function in X_1 . More precisely, $\sigma(\mathbf{x}) = \sigma(\tau_0)(1 + \mathbf{1}(x_1 > 0))$ where $\sigma(\tau_0)$ is a multiplicative constant depending on the threshold parameter τ_0 .

In our second model, we consider a more complex response surface where both the scale and shape parameters depend on the covariates and interactions of order 2 are introduced.

- **Model 2:** Given $\mathbf{X} = \mathbf{x} \in \mathbb{R}^{10}$, Y follows a Student's t -distribution with degree of freedom $\text{df}(\mathbf{x})$ depending on x_1 through

$$\text{df}(\mathbf{x}) = 7(1 + \exp(4x_1 + 1.2))^{-1} + 3,$$

and scale parameter $\text{scale}(\mathbf{x})$ depending on (x_1, x_2) through

$$\text{scale}(\mathbf{x}) = 1 + 6\varphi(x_1, x_2),$$

where φ denotes the density function of a bivariate normal distribution with standard normal margins and correlation 0.9. The numerical constants are chosen so that the GPD approximation of Y given $\mathbf{X} = \mathbf{x}$ has parameters $\gamma(\mathbf{x}) = 1/\text{df}(\mathbf{x})$ in the range $[0.10, 0.33]$ for $\mathbf{x} \in [-1, 1]^d$, $d = 10$.

4.1 Tuning parameters and cross validation

We generate samples of size $n = 2000$ and 5000 , respectively from Model 1 and Model 2. We set the following tuning parameters for `gbex`: the learning rate $\lambda_{\text{scale}} = 0.01$ and the sample fraction $s = 75\%$ for both models; $\lambda_{\text{ratio}} = 15$ for Model 1 and $\lambda_{\text{ratio}} = 7$ for Model 2.

As discussed in Section 3.1, the number of trees B is the most important regularization parameter and the depth parameters (D^σ, D^γ) determine the complexity of the fitted model. Therefore, we investigate how these tuning parameters influence the performance of our estimator in terms of MISE. Figure 2 shows the results for Model 1 (left panel) and for Model 2 (right panel). The curves represent the MISE of `gbex` as a function of B for various depth parameters (D^σ, D^γ) . The right panel clearly shows that for Model 2 the choice $(D^\sigma, D^\gamma) = (1, 1)$ does not account for the model complexity adequately, which leads to a high MISE. Indeed, boosting with depth one tries to fit an additive model but the scale parameter of Model 2 depends on (X_1, X_2) in a non-additive way. On the other hand, for Model 1, which is an additive model with the optimal depth $(D^\sigma, D^\gamma) = (1, 0)$, the curves suggest that assuming unnecessary complexity of the model might lead to suboptimal

behavior of the estimator: the choice $(2, 1)$ yields higher MISE than the other two choices and the MISE stays low for a shorter range of B . In general, higher depths help the model to adapt the data faster but then overfitting is prone to occur more rapidly when B increases. The horizontal dashed lines in Figure 2 represent the resulting MISE of our estimator when B is chosen via cross-validation with deviance loss given in (14), with $K = 5$ folds and 10 replications. The plots confirm that the data driven choice of B results in near optimal MISE for fixed depth parameters (with dashed horizontal lines close to the minimum of the curve with the same color). We additionally apply cross-validation to select both B and (D^σ, D^γ) simultaneously. The resulting MISE is represented by the black dashed line, which is very close to the minimum of all the dashed lines. Overall, the results confirm the good performance of the proposed cross-validation procedure.

For the rest of the simulation study, we set $(D^\sigma, D^\gamma) = (1, 1)$ for Model 1 and $(D^\sigma, D^\gamma) = (3, 1)$ for Model 2 and choose B with cross-validation.

4.2 Comparison with different methods

The comparison of our **gbex** method to the other three approaches **qrf**, **grf** and **constant**, is presented in Figure 3. The results for Model 1 and Model 2 are given in the first and second row, respectively. For the probability level $\tau = 0.99, 0.99$ and 0.9995 in the left, middle and right column, the figure shows the boxplots of ISE defined in (16) and the MISE represented by the vertical black line. The MISE grows as the probability level increases for all methods, however **gbex** clearly outperforms the other three approaches with a much smaller MISE and a much lower variation of ISE. When the probability level τ is close to or larger than $(1 - 1/n)$ (right column), both **grf** and **qrf** lead to extremely large ISE outliers so that the ISE mean is larger than the third quartile (black line outside the box).

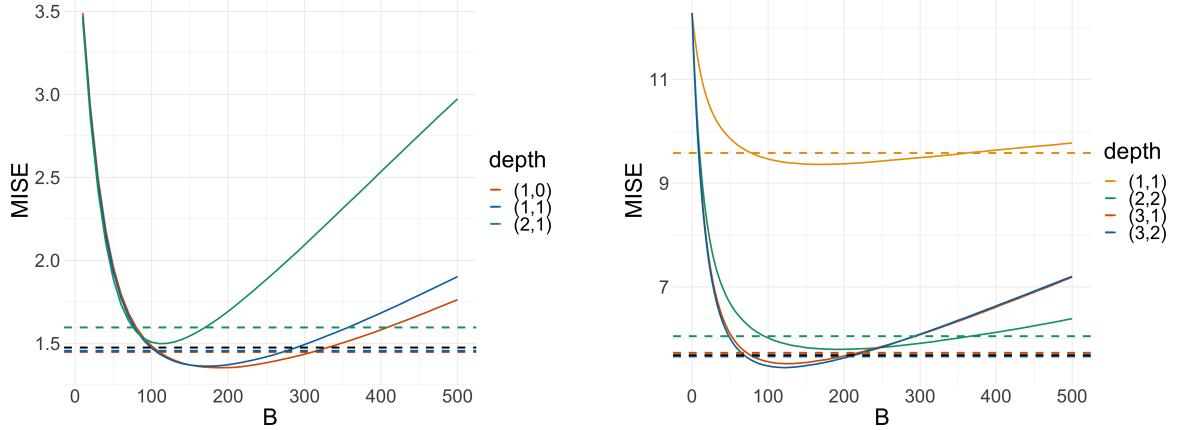


Figure 2: The MISE for Model 1 (left panel) and Model 2 (right panel) of the `gbex` extreme quantile estimator with probability level $\tau = 0.995$ as a function of B for various depth parameters (curves); the MISE of the `gbex` estimator with adaptive choice of B for various depth parameters (horizontal dashed lines); the MISE of the `gbex` estimator with both tree number and depth parameters selected by cross-validation (black dashed line).

Some extreme outliers of ISE are left out of the boxplots to have a clear comparison. We have also investigated other models (Burr, GPD) and the comparison results are reported in Appendix B.

In Figure 3, we have not included `evgam`, the main competitor from extreme value theory. The reason is that for high-dimensional predictor spaces with many noise variables as in our simulations, the additive model for the GPD parameters suffers severely from the curse of dimensionality. Indeed, in an additional simulation from Model 1 with a varying dimension d of the predictor space (not shown here), the MISE of `evgam` grows quickly as a function of d . The MISE of `gbex`, on the other hand, remains fairly constant with growing number of noise variables. The simulation result reveals that the MISE of `gbex` with $d = 40$

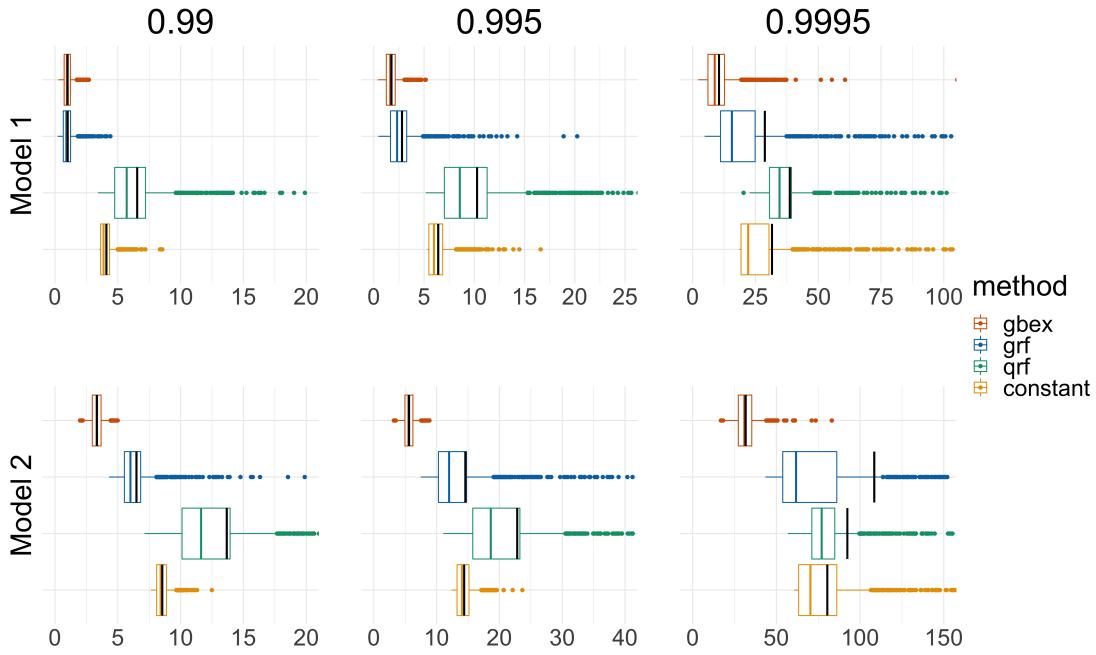


Figure 3: Boxplot of ISE based on 1000 replications for the four quantile estimators (`gbex`, `grf`, `qrf` and `constant`) at different probability levels $\tau = 0.99$ (left), 0.995 (middle) and 0.9995 (right) for Model 1 (top) and Model 2 (bottom). Some outliers of `grf` and `qrf` are left out for a clearer comparison. The black vertical lines indicate the MISE.

is similar to the MISE of `evgam` with $d = 4$. This underlines the robustness of `gbex` against the curse of dimensionality and noise variables, which is a prominent advantage of tree based methods.

4.3 Diagnostic plots

We finally look at the model interpretation diagnostics. Figure 4 shows the permutation importance scores defined in (15) for both models, based on 1000 replications. The boxplots clearly show that this score is able to identify the signal variable(s). Note that there are 39 noise variables for Model 1 and 8 for Model 2. The scores of the noise variables behave all similarly and only a limited number are displayed. For Model 2, the permutation score is higher for X_1 than for X_2 , due to the fact that X_1 contributes to both shape and scale functions while X_2 only contributes to the scale function.

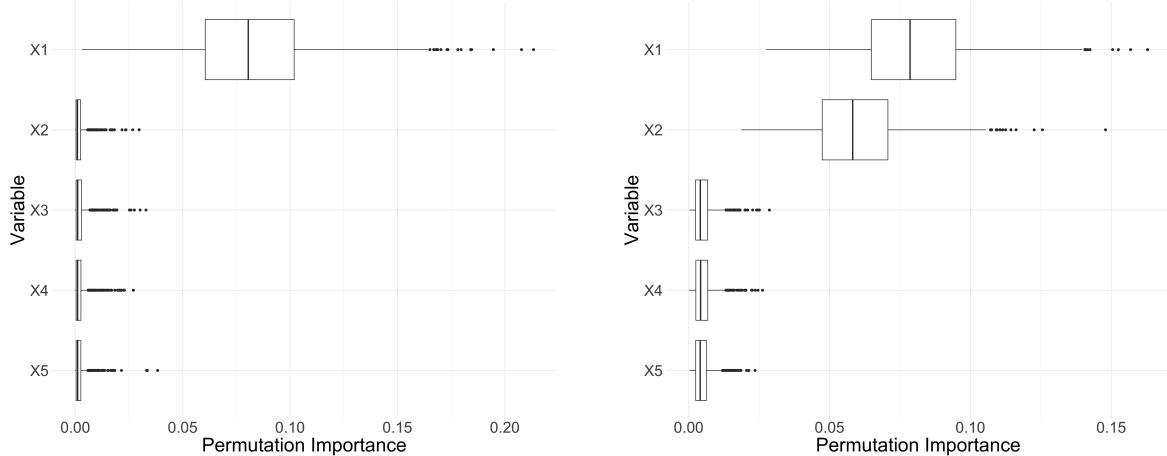


Figure 4: Boxplots of permutation scores defined in (15) for X_j , $j = 1, \dots, 5$, based on 1000 samples. Left panel: Model 1, where only X_1 contains signal. Right panel: Model 2, where only X_1 and X_2 contain signal.

The left panel of Figure 5 presents a typical partial dependence plot (Section 3.2) for $\hat{\sigma}$ based on one random sample from Model 1. This plot clearly suggests that $\hat{\sigma}$ is a step function of X_1 and does not depend on the noise variables. The partial dependence plot for

$\hat{\gamma}$ indicates that the shape does not change with respect to any of the covariates. For this model, the partial dependence plots are in perfect agreement with the simulation design. For Model 2, the left panel of Figure 6 shows the partial dependence plot of the scale parameter with respect to X_1 and X_2 . We see that the model detects the right pattern of larger values on the diagonal and in the center. The right panel shows that the model identifies the impact of X_1 on the shape parameter while the partial dependence plot of the other variables is fairly constant, again in agreement with the simulation design.

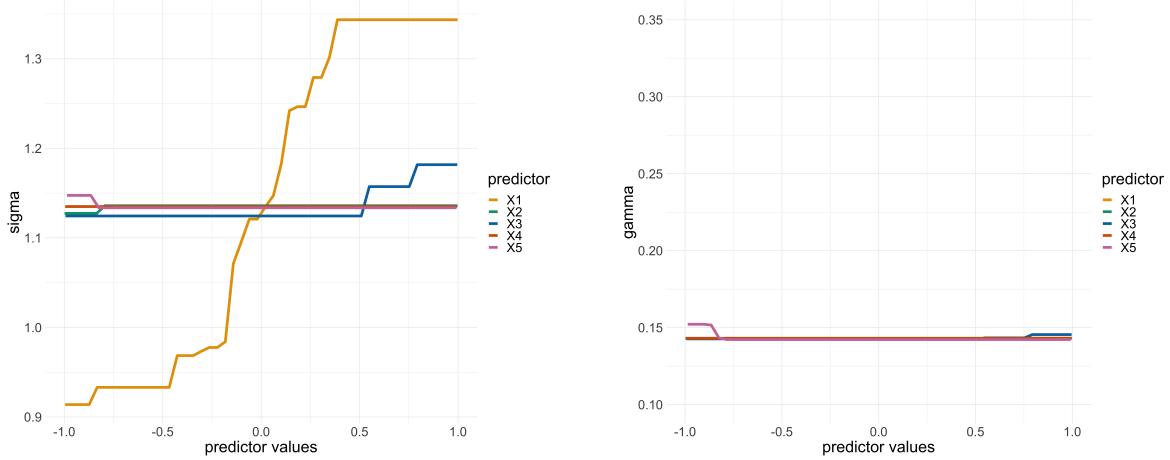


Figure 5: Partial dependence plots of $\hat{\sigma}$ (left panel) and of $\hat{\gamma}$ (right panel) with respect to X_j , $j = 1, \dots, 5$, based on one random sample of Model 1.

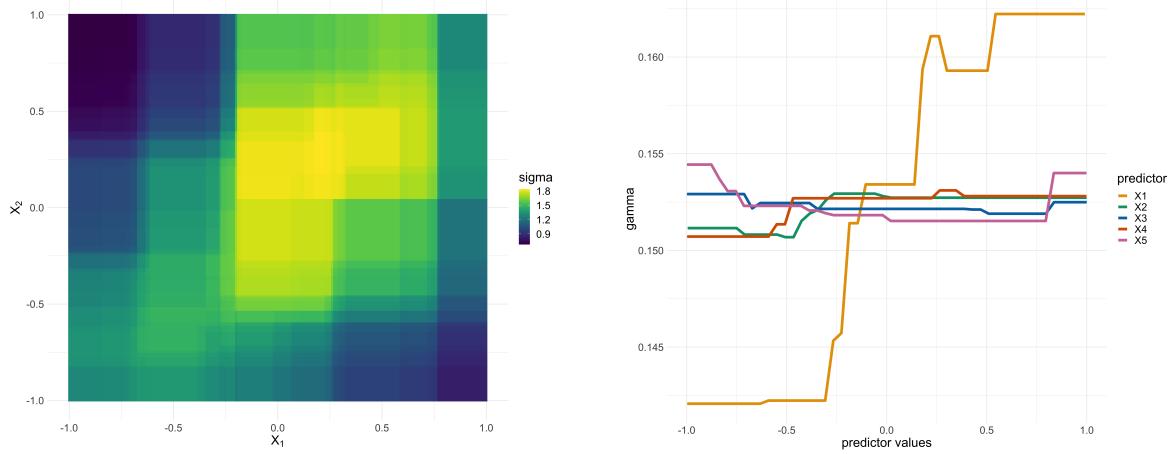


Figure 6: Left panel: partial dependence plots of $\hat{\sigma}$ with respect to (X_1, X_2) . Right panel: partial dependence plot of $\hat{\gamma}$ with respect to X_j , $j = 1, \dots, 5$. Both experiments corresponds to one random sample of Model 2.

5 Application to precipitation forecast

Extreme precipitation events can have disruptive consequences on our society. Accurate predictions are vital for taking preventive measures such as pumping water out of the system to prevent flooding. We apply our `gbex` method to predict extreme quantiles of daily precipitation using the output of numerical weather prediction (NWP) models.

Weather forecasts rely on NWP models that are based on non-linear differential equations from physics describing the atmospheric flow. The solutions to these equations with respect to initial conditions and parametrizations of unresolved processes form a forecast that is deterministic in nature. Introducing uncertainty in these initializations yields an ensemble forecast that consists of multiple members. In this application, we use the ensem-

ble forecast from the European Centre for Medium-Range Weather Forecasts (ECMWF) as covariates in `gbex` to predict the daily precipitation. Using NWP output for further statistical inference to improve forecasts is known as statistical post-processing.

5.1 Precipitation Data

Our data set consists of ECMWF ensemble forecasts of daily accumulated precipitation and the corresponding observations at seven meteorological stations spread across the Netherlands (De Bilt, De Kooy, Eelde, Schiphol, Maastricht, Twente and Vlissingen)¹. We use about 9 years of data, from January 1st, 2011, until November 30th, 2019, with sample size $n = 3256$. We fit separate models for each station with response variable Y equal to the observed precipitation at the station between 00 UTC and 24 UTC.

As for the covariates, we use ECMWF ensemble forecasts of daily accumulated precipitation that is computed the day before at 12UTC. The ensemble forecast contains 51 members. For efficiency, we use two summary statistics, namely the standard deviation of the ensemble members and the upper order statistics (the maximum of the ensemble members). Because most part of the Netherlands is flat and the distance between stations is not large, we include the ensemble summary statistics of all stations as covariates for the model of each station. To account for seasonality, we additionally consider the sine and cosine with a period of 365 for the day of the year. The total covariate dimension is $d = 7 \times 2 + 2 = 16$, for each model. We denote our data as $(Y_i^{(l)}, \mathbf{X}_i)$, where $\mathbf{X}_i \in \mathbb{R}^{16}$, $i = 1, \dots, n = 3256$ and $l = 1, \dots, 7$. For station l , we apply the `gbex` Algorithm 2 to $\{(Y_i^{(l)}, \mathbf{X}_i), i = 1, \dots, n\}$ to obtain estimates of $Q_{\mathbf{X}}^{(l)}(\tau)$.

¹Observed daily precipitation can be obtained from <http://projects.knmi.nl/klimatologie/daggegevens/selectie.cgi>

5.2 Model fitting

For model fitting, we have observed in a preliminary analysis that the output is sensitive to the initial value of (γ, σ) and we propose a specific strategy that provides better results than the default initialization. We consider a common initial value for the shape γ for all the stations and different initial values of σ for the different stations, which leads to $\theta_0 = (\gamma, \sigma_1, \dots, \sigma_7)$. More precisely, we obtain the initial values by optimizing the log-likelihood function

$$L(\theta_0) = \sum_{l=1}^7 \sum_{i=1}^n \left[(1 + 1/\gamma) \log \left(1 + \gamma \frac{Y_i^{(l)} - c}{\sigma_l} \right) + \log \sigma_l \right] \mathbb{1}_{\{Y_i^{(l)} - c > 0\}},$$

where c is a large threshold chosen such that the estimate of γ becomes stable.

We apply `gbex` as detailed in Algorithm 2 with $\tau_0 = 0.8$ for each model. We choose all tuning parameters except for B to be the same for the seven models, in such a way to achieve the overall best combined deviance score for all stations. This prevents overfitting for a specific station and it results in the following choices: $(D^\sigma, D^\gamma) = (2, 1)$, $(\lambda_{scale}, \lambda_{ratio}) = (0.01, 12)$, $s = 50\%$, and $(L_{\min}^\sigma, L_{\min}^\gamma) = (15, 45)$. Figure 7 shows the cross-validated deviance as a function of the number of trees B for different depth levels at two stations. The deviance behaves quite similar for the two stations and we choose $(D^\sigma, D^\gamma) = (2, 1)$ for all stations. The optimal B for each station is then chosen as the minimizer of the cross-validated deviance.

5.3 Results

We first look into the variable importance scores for the fitted models and focus on the relative importance to understand which variables affect the scale and shape parameters, respectively. Figure 8 shows the relative importance for γ and σ , where the scores for

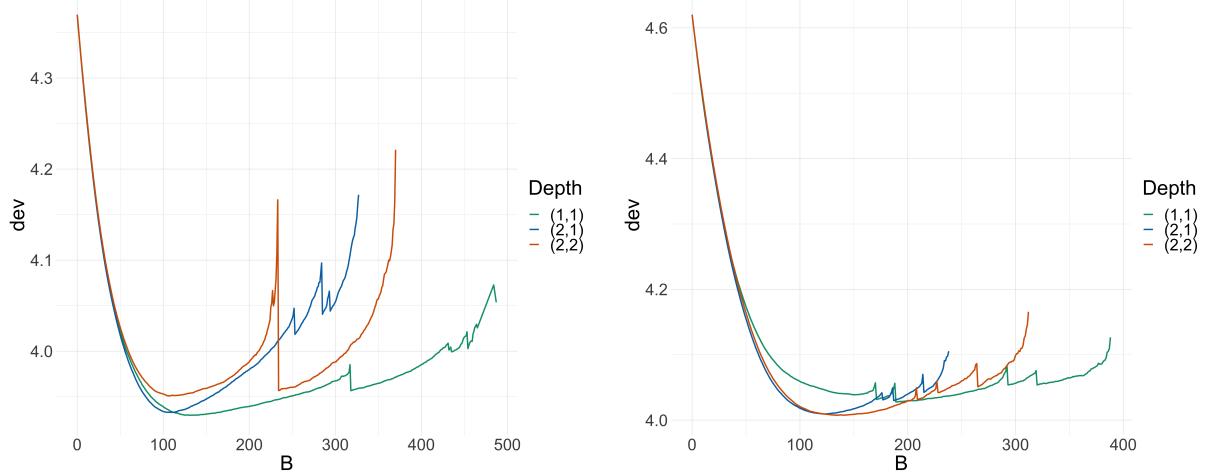


Figure 7: Cross-validation deviance given by (14) against B for the data at stations Eelde (left) and Schiphol (right) in the application in Section 5.

the variable `ens_sd` and `ens_up` correspond to the aggregation of 7 scores (one for each station). It is interesting to note that for the shape γ , the `day of year` is the most important variable in six out of seven models. This motivates to investigate the seasonality pattern in the extreme precipitation. The partial dependence plots of $\hat{\gamma}^{(l)}$ (left panel) and $\hat{Q}_\mathbf{X}^{(l)}(0.995)$ (right panel) with respect to the `day of year` are presented in Figure 9 for all stations. They indicate that the tail of the precipitation is heavier in summer and autumn than in winter and spring. The curves in the left panel resemble step functions and higher values of $\hat{\gamma}$ correspond to June, July and August for five stations. For the other two stations Twente and Vlissingen, it is shifted towards autumn.

Another relevant question concerns the contribution of ensemble statistics of other stations in forecasting the extreme precipitation of a specific location. To this end, we add the permutation scores of ensemble standard deviation and ensemble upper order statistics

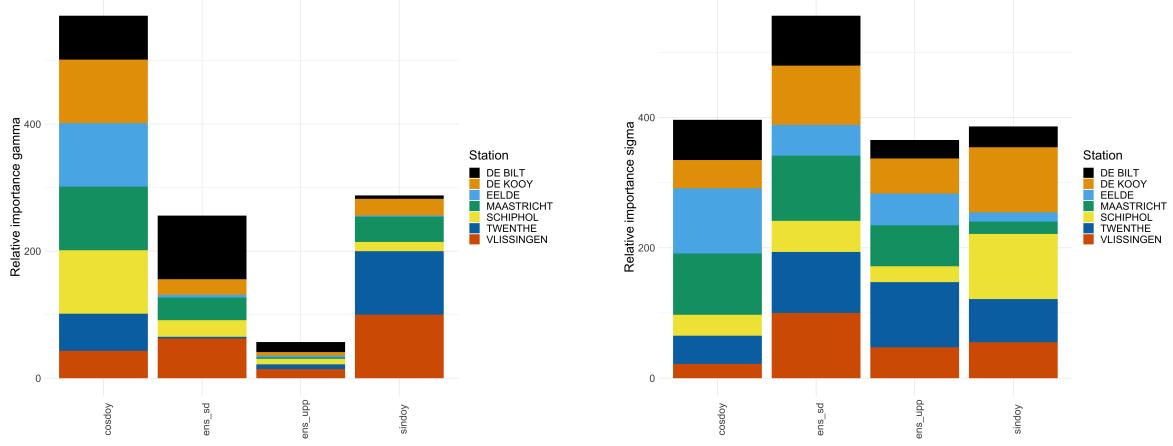


Figure 8: Relative variable importance score for γ (left) and σ (right). For each model, the scores are normalized such that the maximum score is 100. The scores for the variables `ens_sd` (resp. `ens_up`) at the 7 stations are aggregated into a single score.

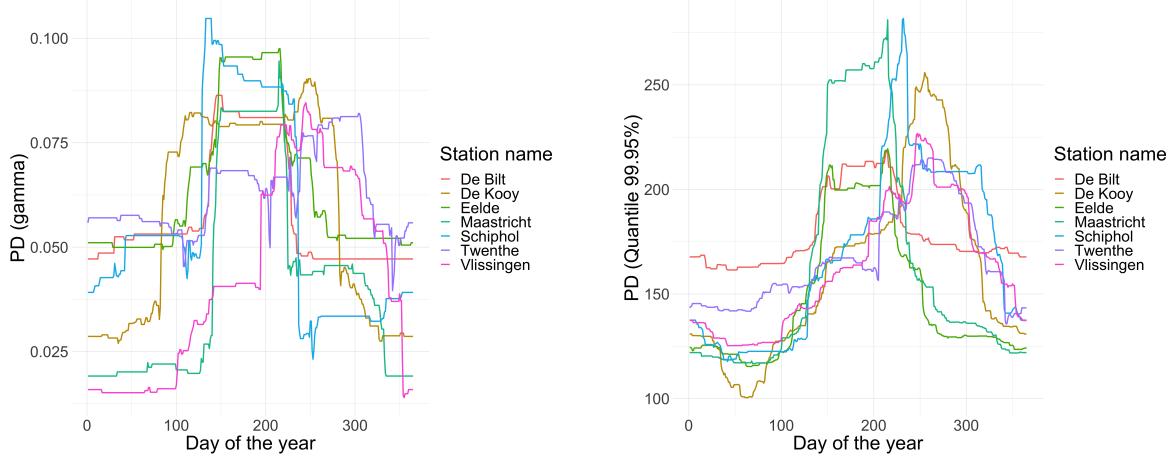


Figure 9: Partial dependence plots of $\hat{\gamma}^{(l)}$ (left panel) and $\hat{Q}_{\mathbf{X}}^{(l)}(0.9995)$ (right panel, in 0.1mm) with respect to `day of year`.

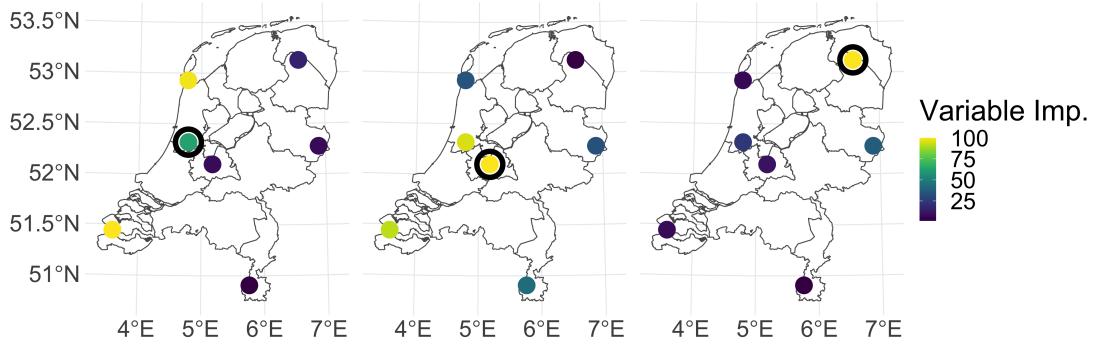


Figure 10: Normalized permutation scores of ensemble statistics per location for three models: Schiphol (left), De Bilt (middle), Eelde (right). The black circle indicates the station for which the model is fitted. From North to South, the stations are: Eelde, De Kooy, Twente, Schiphol, De Bilt, Vlissingen, Maastricht.

per station, resulting in seven scores for each model. We then normalize these scores such that the maximum score is 100. The results for three stations are visualized in Figure 10. First, quite surprisingly, when forecasting the extreme precipitation at Schiphol (left plot), the ensemble forecast relies on the information from Vlissingen and De Kooy even more than the information at Schiphol, which might be explained by a coastal effect. Similarly, the model at De Bilt (middle plot) uses the information from Schiphol and Vlissingen. For other stations like Eelde (right plot), the own information of the station is the most important. The maps of the other four stations (De Kooy, Maastricht, Vlissingen and Twente) are very similar to that of Eelde.

Our method can be used to provide relevant information for weather warning systems. The Dutch meteorology institute (KNMI) issues three levels of weather warnings for disruptive weather conditions, namely code yellow, code orange and code red. Code red, the

most severe one, is issued depending on the social impact and safety risk of extreme weather conditions. Code yellow and code orange are issued if some weather quantity such as snow-fall, slipperiness, temperature, or wind speed, reaches a specific level. For precipitation, the threshold is 50 mm (resp. 75 mm) within 24 hours for code yellow (resp. orange). As an illustration on how our method can be informative for the weather warning system, we look into the predicted 99.95% quantile by **gbex** for the month when the maximum observed precipitation (over the time span of our data set) occurs, and compare it to the thresholds of code yellow and code orange. Figure 11 presents the results for three stations: De Kooy, Schiphol and Vlissingen. The maximum observed precipitation were 52.3 mm on July 14, 2011, 67.2 mm on September 8, 2017 and 49.9 mm on October 13, 2013, respectively, for these three stations. For these three days, our prediction of the 99.95% quantile (using only information from the past) indicates a high level of precipitation, comparable to the code orange level. It could therefore be used for effective early warning. Overall, the blue curve (predicted 99.95% quantile) is above the black points (observations) and it captures well the days with heavy precipitation.

We finally assess the goodness of fit of our GPD model and produce QQ-plots comparing the empirical and theoretical quantiles of exceedances above threshold. We use a transformation to the exponential distribution to compare observations stemming from different stations with different covariate values. More precisely, denoting by $Z_i^{(l)}$ the i th exceedance above threshold at station l , then if our model is well-specified $Z_i^{(l)} \sim \text{GPD}(\hat{\sigma}^{(l)}(\mathbf{X}_i), \hat{\gamma}^{(l)}(\mathbf{X}_i))$, and therefore

$$\frac{1}{\hat{\gamma}^{(l)}(\mathbf{X}_i)} \log \left(1 + \frac{\hat{\gamma}^{(l)}(\mathbf{X}_i) Z_i^{(l)}}{\hat{\sigma}^{(l)}(\mathbf{X}_i)} \right) \sim \text{Exp}(1). \quad (17)$$

The corresponding QQ-plots graphically assess the goodness of fit and we can see in Fig-

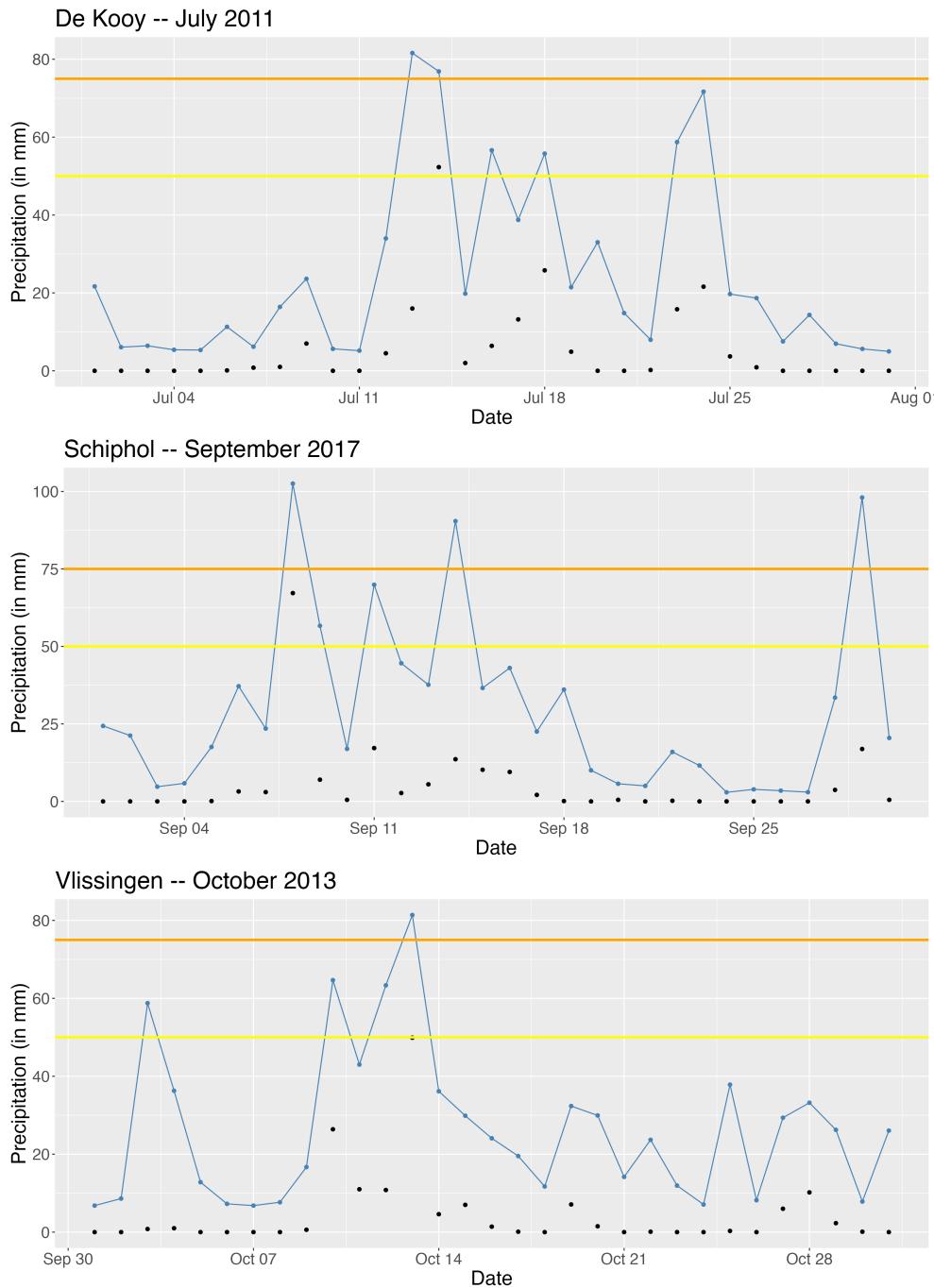


Figure 11: Black points: observed precipitation; blue points: predicted 99.95% quantile; yellow (orange) line: precipitation threshold³⁹ of 50 mm (75 mm) for code yellow (orange) weather warnings.

ure 12 that the **gbex** model (left panel) fits the data well at all stations, outperforming the **constant** model (right panel). Such a plot can be used to compare different choices of the intermediate threshold τ_0 . Points close to the diagonal indicate that not only the regression model is good, but also that the approximation of the exceedances by the generalized Pareto distribution is appropriate at this this threshold level.

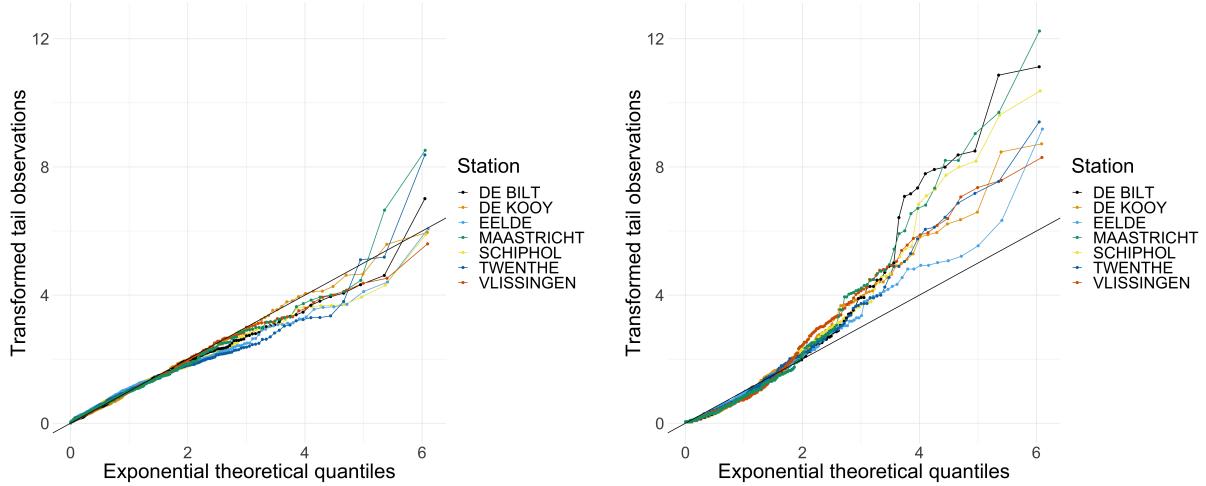


Figure 12: QQ-plots based on (17) for the estimated models at seven stations via **gbex** (left panel) and via the **constant** method (right panel).

6 Conclusion

The existing literature on extreme quantile regression is so-far limited to low-dimensional predictor spaces [8, 16, 30] and simple response surfaces [9, 32, 6, 34]. Our methodological contribution fills a gap in this area. We have developed **gbex**, a gradient boosting procedure for extreme quantile regression that combines the flexibility of machine learning methods

and the rigorous extrapolation from extreme value theory. Our method can handle non-linear complex problems and high-dimensional feature spaces.

We model the tail of the distribution of the response Y by a generalized Pareto distribution (GPD) whose parameters depend on the covariate \mathbf{X} . Based on exceedances over a high threshold, gradient boosting produces a tree ensemble estimating these parameters using the deviance as the objective function. Tuning parameters can effectively be chosen through our proposed cross-validation, or be fixed to sensible default values. In several numerical experiments we highlight the robustness of `gbex` against the curse of dimensionality and noise variables. Diagnostic tools are available to quantify the impact of the signal variables on the response. Our method outperforms quantile regression methods from machine learning and classical methods based on extreme value theory. The method can be applied to complex real-world data sets and we show its merits for post-processing of extreme precipitation forecasts in the Netherlands.

A very natural yet challenging direction for future research is the theoretical analysis of our gradient boosting procedure. A consistency result for large samples is desirable but all existing results in the literature on gradient boosting assume the convexity of the objective function [e.g., 3]. The GPD deviance used as objective function in our setting is not convex in the shape parameter γ . A proper theoretical analysis of `gbex` therefore seems to be very hard and is outside the scope of the present paper.

Acknowledgements

This work was supported by the Netherlands Organisation for Scientific Research (NWO) under grant number 14612, by the French Agence Nationale de la Recherche under grant

number ANR-20-CE40-0025-01, and by the Swiss National Science Foundation under grant number 186858.

Conflict of interest

The authors declare that they have no conflict of interest.

Data availability statement

The datasets generated during and/or analysed during the current study are available in the Github repository <https://github.com/JWelthoen/gbex/>. The daily precipitation data used in Section 5 is also publicly available on <http://projects.knmi.nl/klimatologie/daggegevens/selectie.cgi>

References

- [1] S. Athey, J. Tibshirani, S. Wager. Generalized random forests. *The Annals of Statistics*, 47(2):1148–1178, 2019.
- [2] A. A. Balkema and L. de Haan. Residual life time at great age. *The Annals of Probability*, 2(5):792–804, 1974.
- [3] G. Biau and B. Cadre. Optimization by gradient boosting. In A. Daouia and A. Ruiz-Gazen, editors, *Advances in Contemporary Statistics and Econometrics: Festschrift in Honor of Christine Thomas-Agnan*, pages 23–44. Springer International Publishing, 2021.

- [4] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.
- [5] L. Breiman, J. Friedman, C.J. Stone, and R.A. Olsen. Classification and Regression Trees. Taylor & Francis, 1984.
- [6] V. Chavez-Demoulin and A. Davison. Generalized additive modelling of sample extremes. *Journal of the Royal Statistical Society, series C*, 54, 2005.
- [7] V. Chernozhukov. Extremal quantile regression. *Ann. Statist.*, 33(2):806–839, 2005.
- [8] A. Daouia, L. Gardes, and S. Girard. On kernel smoothing for extremal quantile regression. *Bernoulli*, 19(5B):2557–2589, 2013.
- [9] A. Davison and R. Smith. Models for exceedances over high threshold. *Journal of the Royal Statistical Society, series B*, 52, 1990.
- [10] H. Drees, A. Janßen, S. Resnick, and T. Wang. On a minimum distance procedure for threshold selection in tail analysis. *SIAM Journal on Mathematics of Data Science*, 2(1):75–102, 2020.
- [11] D. J. Dupuis. Exceedances over high thresholds: A guide to threshold selection. *Extremes*, 1(3):251–261, 1999.
- [12] S. Farkas, O. Lopez, and M. Thomas. Cyber claim analysis using generalized pareto regression trees with applications to insurance. *Insurance: Mathematics and Economics*, 98:92–105, 2021.
- [13] J. H. Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

- [14] J. H. Friedman. Stochastic gradient boosting. *Computational Statistics & Data Analysis*, 38(4):367–378, 2002.
- [15] L. Gardes. Tail dimension reduction for extreme quantile estimation. *Extremes*, 21(1):57–95, 2018.
- [16] L. Gardes and G. Stupler. An integrated functional Weissman estimator for conditional extreme quantiles. *REVSTAT*, 17(1):109–144, 2019.
- [17] N. Gnecco, E. M. Terefe, and S. Engelke. Extremal random forests, 2022. URL <https://arxiv.org/abs/2201.12865>.
- [18] O. C. Pasche and S. Engelke. Neural networks for extreme quantile regression with an application to forecasting of flood risk, 2022. URL <https://arxiv.org/abs/2208.07590>.
- [19] J. Richard and R. Huser. A unifying partially-interpretable framework for neural network-based extreme quantile regression, 2022. URL <https://arxiv.org/abs/2208.07581>.
- [20] L. de Haan and A. Ferreira. Extreme Value Theory: An Introduction. Springer, New York, 2006.
- [21] T. Hastie, R. Tibshirani, and J. H. Friedman. *The elements of statistical learning*. Springer Series in Statistics. Springer, New York, second edition, 2009. Data mining, inference, and prediction.
- [22] R. Koenker and G. Bassett Jr. Regression quantiles. *Econometrica: journal of the Econometric Society*, pages 33–50, 1978.

- [23] V. Mai, and M. Johansson. Stability and Convergence of Stochastic Gradient Clipping: Beyond Lipschitz Continuity and Smoothness. *Proceedings of the 38th International Conference on Machine Learning*, 139(Jul):7325–7335, 2021.
- [24] N. Meinshausen. Quantile regression forests. *Journal of Machine Learning Research*, 7(Jun):983–999, 2006.
- [25] H. Niederreiter. *Random number generation and quasi-Monte Carlo methods*, volume 63 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1992.
- [26] J. Pickands III. Statistical inference using extreme order statistics. *Ann. Statist.*, 3(1):119–131, 01 1975.
- [27] J. Qian, Y. Wu, B. Zhuang, S. Wang, and J. Xiao. Understanding Gradient Clipping In Incremental Gradient Methods. *Proceedings of the 38th International Conference on Machine Learning*, 130(Apr):1504–1512, 2021.
- [28] G. Ridgeway. Generalized boosting models: a guide to the gbm package, 2007. URL <https://cran.r-project.org/web/packages/gbm/vignettes/gbm.pdf>.
- [29] R. L. Smith. Estimating tails of probability distributions. *Ann. Stat.*, 15:1174–1207, 1987.
- [30] J. Velthoen, J.-J. Cai, G. Jongbloed, and M. Schmeits. Improving precipitation forecasts using extreme quantile regression. *Extremes*, 22(4):599–622, 2019.
- [31] S. Wager and S. Athey. Estimation and inference of heterogeneous treatment effects

using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

- [32] H. Wang and C.-L. Tsai. Tail index regression. *Journal of the American Statistical Association*, 104(487):1233–1240, 2009.
- [33] H. J. Wang, D. Li, and X. He. Estimation of high conditional quantiles for heavy-tailed distributions. *Journal of the American Statistical Association*, 107(500):1453–1464, 2012.
- [34] B. D. Youngman. Generalized additive models for exceedances of high thresholds with an application to return level estimation for U.S. wind gusts. *Journal of the American Statistical Association*, 114(528):1865–1879, 2019.

A Likelihood derivatives

The gradient boosting algorithm for GPD modeling makes use of the first and second order derivatives of the negative log likelihood $\ell_z(\theta)$, $\theta = (\sigma, \gamma)$ and $z > 0$. They are respectively given by

$$\begin{aligned}\frac{\partial \ell_z}{\partial \sigma}(\theta) &= \frac{1}{\sigma} \left(1 - \frac{(1 + \gamma)z}{\sigma + \gamma z} \right), \\ \frac{\partial \ell_z}{\partial \gamma}(\theta) &= -\frac{1}{\gamma^2} \log \left(1 + \gamma \frac{z}{\sigma} \right) + \frac{(1 + 1/\gamma)z}{\sigma + \gamma z},\end{aligned}$$

and

$$\begin{aligned}\frac{\partial^2 \ell_z}{\partial \sigma^2}(\theta) &= \frac{1}{\sigma(\sigma + \gamma z)} \left(\frac{z}{\sigma} + \frac{z - \sigma}{\sigma + \gamma z} \right), \\ \frac{\partial^2 \ell_z}{\partial \gamma^2}(\theta) &= \frac{2}{\gamma^3} \log \left(\gamma \frac{z}{\sigma} + 1 \right) - \frac{2z}{\gamma^2(\sigma + \gamma z)} - \frac{(1 + 1/\gamma)z^2}{(\sigma + \gamma z)^2}.\end{aligned}$$

B Additional simulation study

The data generating process is similar to Model 1 in Section 4. The covariate vector $\mathbf{X} \in \mathbb{R}^{40}$ is distributed uniformly on the cube $[-1, 1]^{40}$. We consider three heavy-tailed distributions, namely Burr, GPD and Student's t , as the conditional distribution of Y given \mathbf{X} . For all models, the scale of Y depends on \mathbf{X} through a step function

$$\text{scale}(\mathbf{X}) = 1 + \mathbf{1}(X_1 > 0). \quad (18)$$

The conditional distributions are respectively:

- **Model 3:** a Student's t -distribution with 2 degrees of freedom and the scale given in (18).
- **Model 4:** a GPD in (2) with $\gamma = 0.25$ and $\sigma(\mathbf{x})$ given in (18).
- **Models 5-6:** a Burr distribution with a CDF given by

$$F(y) = 1 - \left(1 + \left(\frac{y}{\text{scale}(\mathbf{x})} \right)^\alpha \right)^\beta.$$

We choose $\alpha = \beta = 2$ for Model 5 and $\alpha = 2, \beta = 1$ for Model 6, which lead to $\gamma = 0.25$ and 0.5, respectively.

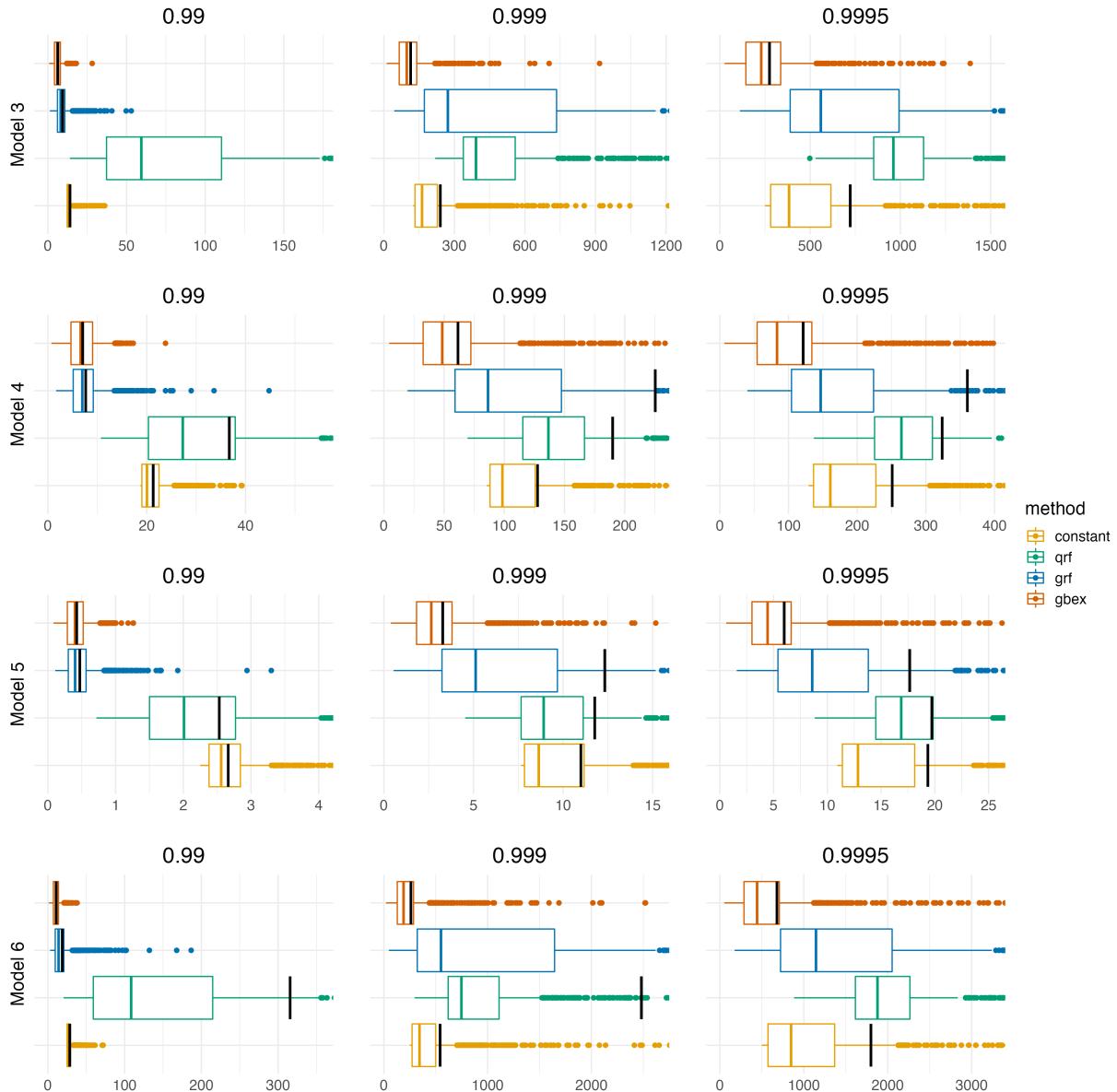


Figure 13: Boxplot of ISE based on 1000 replications for the four quantile estimators (gbex, grf, qrf and constant) at different probability levels $\tau = 0.99, 0.999, 0.9995$ for Models 3-6. Some outliers of grf and qrf are left out for a clearer comparison. The black vertical lines indicate the MISE.