Noname manuscript No.
(will be inserted by the editor)

Scoring Predictions at Extreme Quantiles

Axel Gandy · Kaushik Jana · Almut E. D. Veraart

Received: date / Accepted: date

Abstract Prediction of quantiles at extreme tails is of interest in numerous applications. Extreme value modelling provides various competing predictors for this point prediction problem. A common method of assessment of a set of competing predictors is to evaluate their predictive performance in a given situation. However, due to the extreme nature of this inference problem, it can be possible that the predicted quantiles are not seen in the historical records, particularly when the sample size is small. This situation poses a problem to the validation of the prediction with its realisation. In this article, we propose two non-parametric scoring approaches to assess extreme quantile prediction mechanisms. The proposed assessment methods are based on predicting a sequence of equally extreme quantiles on different parts of the data. We then use the quantile scoring function to evaluate the competing predictors. The performance of the scoring methods is compared with the conventional scoring method and the superiority of the former methods are demonstrated in a simulation study. The methods are then applied to reanalyse cyber Netflow data from Los Alamos National Laboratory and daily precipitation data at a station in California available from Global Historical Climatology Network.

Keywords Extreme value · High quantile · Quantile score

Axel Gandy
Department of Mathematics
Imperial College London, London, SW7 2AZ, U.K.
E-mail: a.gandy@imperial.ac.uk

Kaushik Jana (corresponding author)
Department of Mathematics
Imperial College London, London, SW7 2AZ, U.K.
E-mail: k.jana@imperial.ac.uk

Almut E. D. Veraart
Department of Mathematics
Imperial College London, London, SW7 2AZ, U.K.
E-mail: a.veraart@imperial.ac.uk

1 Introduction

Prediction of extreme quantiles or equivalently, so-called return levels, is an important problem in various applications of extreme value analysis including but not limited to meteorology, hydrology, climatology and finance. The main task of inference in many of these prediction problems involves the computation of probabilities of yet unobserved rare events that are not seen in the historical records. Specifically, the problem may involve the estimation of a high quantile, say at level p, based on a sample of size n, where p is very close to 1 so that $n(1-p)$ is small. Examples include the prediction of a quantile of the distribution of precipitation which could be realized once in every 200 years based on 50 to 100 years of rainfall data, or the prediction of high financial loss based on few years of data.

Extreme value theory (EVT) provides tools to predict extreme events, based on the assumption that the underlying distribution of the normalized random variable resides in the domain of attraction of an extreme value distribution (Fisher and Tippett, 1928). Thus extreme quantiles can be predicted by estimating the parameters of the specific extreme value distribution and the normalizing constants. Detailed reviews on different models and methods of estimation in this context can be found in Embrechts et al. (1997), and Coles (2001), among others. Many distinct models could be applicable for predicting an extreme quantile in a given situation, thereby calling for a comparative assessment of their predictive performance.

To our knowledge, relatively little work has been done on the evaluation of quantile predictor (point forecast mechanism) in the context of extreme events. Friederichs and Thorarinsdottir (2012) considered probabilistic forecasts and derived the continuous rank probability score (CRPS) for two common extreme value distributions to assess forecasts for predictive distribution of the peak wind prediction. This method was also used in Bentzien and Friederichs (2014a) and Scheuerer and Moller (2015) for precipitation forecast ensembles through the assumption of a parametric form for the prediction distribution. Lerch et al. (2017) proposed a weighted CRPS for a probabilistic forecast by imposing weights on the tail of the distribution with emphasis on extreme events. In the case of point prediction, specifically for evaluating quantile prediction, the standard tool to use is the quantile score which is a strictly proper and consistent scoring rule for quantile functional (Gneiting and Raftery, 2007). In recent work, Brehmer and Strokorb (2019) demonstrate that even a proper scoring rule can not distinguish tail properties in a very strong sense.

Another challenge of assessment of a very high (or a very low) quantile estimate is that, due to the sparseness of the data at extreme tails, the estimate may fall outside the range of the observed data, therefore, leaving no practical information about the estimate to validate. This instance also implies that the quantile score is optimized trivially at a specific estimate (see equation (4) of Section 2).

To address this problem, we propose to predict a new set of equally extreme quantiles using different subsets of the data set to compare various quantile prediction methods. The new quantile points are chosen such that their predictions are observed in the rest of the data (test sample). Each of the competing predictors is thus used to estimate the series of quantiles on different parts of the data and validated in the rest of the data. This method then uses a combined quantile scoring function and cross-validation to assess different competing extreme quantile predictors.

We propose two approaches. The first one uses a smaller part of the data to predict a new set of quantiles and validate the predictions using the remaining larger part of the data. In the second approach, the quantiles are predicted on a more substantial part of data and validated in the smaller part of the data.

We examine the performance of the two methods under different data generating processes through a simulation study and compare their performance with the quantile scoring method. The proposed two methods are more efficient than the quantile scoring method when the GPD model is reasonably well fitted in different parts of the data. We also observe that the first approach, which uses a larger part of data for validation of the competing predictor performs more favourably than the second approach.

The outline of the paper is as follows. In Section 2, we present two different methods of assessment of the extreme quantile predictors, describe different choices of the tuning parameters, and list a few examples of the extreme quantile predictor. We present a simulation study in Section 3 to demonstrate the performance of the optimum scoring predictors under different data-generating mechanisms. In Section 4, we apply the proposed methodologies to two different data sets: the LANL Netflow data and the precipitation data corresponding to one station of the Global Historical Climatology Network data. We conclude with some remarks in Section 5.

2 Methodology

Consider identically distributed, continuous random variables $Y, Y_1, \ldots, Y_n \in \mathbb{R}$, for $n \in \mathbb{N}$. We write $\mathbf{Y} = (Y_1, \ldots, Y_n)$ for the corresponding random sample of size n and $\mathbf{y} = (y_1, \ldots, y_n)$ for a realisation.

Let $p \in (0, 1)$. Recall that the p-th quantile of Y is defined as

$$q_Y(p) = \inf\{y : P(Y \leq y) \geq p\}.$$

Let Q denote a quantile predictor function, such that $(p, \mathbf{Y}) \mapsto Q(p, \mathbf{Y})$ is a predictor of $q_Y(p)$ and $Q(p, \mathbf{y})$ is its empirical counterpart.

Suppose we wish to estimate the quantile $q_Y(p^0)$ for $p^0 \in (0, 1)$ close to 1 and there are $m \in \mathbb{N}$ competing quantile predictors denoted by Q_i for $i = 1, \ldots, m$. Let

$$\mathbb{Q} = \{Q_1, \ldots, Q_m\}$$

denote the set of these m competing predictors.

Evaluation of point predictors is typically done by means of a scoring function $s(a, b)$ that assigns a numerical score when the point forecast a is issued and the observation b is realized (Gneiting, 2011). For evaluating a point predictor consistently, it is recommended to use a strictly proper and consistent scoring rule for the functional of interest. The quantile score is such a rule for assessing quantile predictors at given probability levels and used in various applications (Bentzien and Friederichs, 2014b).

In the context of evaluating quantile forecasts, one typically uses the check loss function, see (Koenker, 1984), defined as

$$\rho_p(a, b) = (a - b)(p - I(a < b)), \qquad \text{for } a, b \in \mathbb{R}.$$

We will now introduce our notation for the corresponding quantile scores we are going to consider. Let $\mathbf{x}^{(1)}$ and $\mathbf{x}^{(2)}$ denote n_1- and, respectively, n_2-dimensional realisations of subsamples of $\mathbf{Y}$. Then we define by

$$S(Q(\cdot, \cdot), p, \mathbf{x}^{(1)}, \mathbf{x}^{(2)}) = \frac{1}{n_2} \sum_{i=1}^{n_2} \rho_p(Q(p, \mathbf{x}^{(1)}), x_i^{(2)}) \tag{1}$$

the average quantile score for the predictor $Q(p, \cdot)$ based on the empirical test (sub-) sample $\mathbf{x}^{(1)}$ and the empirical validation (sub-) sample $\mathbf{x}^{(2)}$.

In the case when $\mathbf{x}^{(1)} = \mathbf{x}^{(2)} = \mathbf{y}$, equation (1) simplifies to

$$S(Q(\cdot, \cdot), p, \mathbf{y}) := S(Q(\cdot, \cdot), p, \mathbf{y}, \mathbf{y}) = \frac{1}{n} \sum_{i=1}^{n} \rho_p(Q(p, \mathbf{y}), y_i). \tag{2}$$

Given the set $\mathbb{Q}$ of competing predictors, the optimum quantile predictor $\widehat{Q}_{qs}$ is defined as the minimizer of the average quantile score via

$$\widehat{Q}_{qs} = \widehat{Q}_{qs}(\cdot, \cdot) = \operatorname*{arg\,min}_{Q \in \mathbb{Q}} S(Q(\cdot, \cdot), p^0, \mathbf{y}). \tag{3}$$

Then $\widehat{q}_{qs}(p^0) = \widehat{Q}_{qs}(p^0, \mathbf{y})$ denotes estimated p^0-quantile using the predictor $\widehat{Q}_{qs}$ and the data $\mathbf{y}$.

Predicting quantiles from very extreme tails is a difficult task due to the sparseness of the data in the tails of the distribution, especially for heavy-tailed distributions, see (Wang et al., 2012). This fact also makes it challenging to evaluate the predictions, as they may not be realized inside the range of the observed data.

In the situation, where all predictors in $\mathbb{Q}$ predict a quantile larger than $\max\{y_1, \ldots, y_n\}$, the optimal quantile prediction is the smallest prediction from $\mathbb{Q}$. Indeed, for every $Q(p, \mathbf{y}) \in \mathbb{Q}$, equation (2) reduces to

$$S(Q(p, \cdot), \mathbf{y}) = \frac{p}{n} \sum_{i=1}^{n} \big(Q(p, \mathbf{y}) - y_i\big). \tag{4}$$

This may not be a reasonable answer when assessing the prediction of very large quantiles.

This shortcoming motivated us to consider an improved scoring method for assessing predictors for high quantiles. The developed method can be easily adapted to the extreme lower quantiles with p close to 0.

More specifically, we propose two methods for evaluating the performance of the predictors of $\mathbb{Q}$ by training and assessing them on different parts of the data set. Both methods employ a k-fold cross-validation, where Method 1 uses a small training and large validation sample, whereas Method 2 uses a large training and a small validation sample.

2.1 Main idea

Suppose we want to predict the p^0th quantile for p^0 close to 1. Let us consider the problem of predicting the p^cth quantile where, $0 < p^c < p^0 < 1$, based on a sub-sample (training sample) of size n^c of the original sample of size n.

We then propose to choose p^c and n^c such that estimating the p^cth quantile based on a sub-sample of size n^c is *equally extreme* as estimating the p^0th quantile based on the original sample of size n. By equally extreme, we mean, the expected number of exceedances above the estimated p^cth and p^0th quantile are the same in training and original sample, respectively. This condition implies that

$$n^c(1-p^c) = n(1-p^0). \tag{5}$$

Now we have one equation and two unknowns (n^c and p^c). In order to be able to determine n^c and p^c uniquely, we impose an additional condition:

Let $\alpha > 0$ denote the *expected number of observations exceeding the p^cth quantile in the remainder of the sample*, i.e. in the validation sample of size $n - n^c$. We assume that α is fixed a priori, hence we obtain the condition that, for a *given* α:

$$(n-n^c)(1-p^c) = \alpha. \tag{6}$$

The system of the two equations (5) and (6) has a unique solution for n^c and p^c given by

$$n^c = \frac{n}{1+\frac{\alpha}{n(1-p^0)}}, \tag{7}$$

$$p^c = p^0 - \frac{\alpha}{n}. \tag{8}$$

In practice, the choice of α will determine the balance between the number of exceedances in the training and the validation sample. A smaller (or larger) value of α leads to a new quantile (p^cth) closer to (or further away from) the target quantile (p^0th) and leaves fewer (more) observations for the evaluation of the predictions.

2.1.1 Using k-fold cross-validation

We will now describe how our idea can be implemented in a k-fold cross-validation framework.

Suppose n is given and p^0 and α have been fixed, and n^c and p^c have been determined using equations (7) and (8), respectively.

We split the original sample $\mathbf{y}$ into k sub-samples of size $\frac{n}{k}$ denoted by $\mathbf{y}^{(i)}$, for $i = 1, \dots, k$. Also, $\mathbf{y}^{\setminus(i)}$ denotes the sub-sample of $\mathbf{y}$ of size $\frac{(k-1)}{k}n$, which excludes all the observations contained in $\mathbf{y}^{(i)}$.

We have to find k for fixed n, p^0, α. In order to do this, we investigate two routes:

- Method 1: Small training sample (one fold) and large validation sample (k-1 folds).
- Method 2: Large training sample (k-1 folds) and small validation sample (one fold).

We note that the choice of α determines how much importance is given to the estimation versus the validation of the extremes. A relatively high value of α leads to more exceedances available for validation, but less for estimation and vice versa.

We note that in the following we will be treating α as our tuning parameter and derive an expression for the number of folds k in terms of α. Equivalently, one could fix the number of folds k first and derive the corresponding expression for α.

2.2 Method 1: Small training sample and large validation sample

We propose an unconventional version of the k-fold cross-validation, where at a time, one sub-sample $\mathbf{y}^{(i)}$ is used as the training sample and the collection of the remaining $k-1$ sub-samples given by $\mathbf{y}^{\setminus(i)}$ is used as the validation sample. Thus each data point is used once for training and $k-1$ times for validation purposes. (In a conventional k-fold cross-validation set-up, one would use a collection of $k-1$ folds for training and one fold for validation.)

I.e. we set the size of the training sample as $n^c = n/k$, which results in

$$k = \frac{n}{n^c} = 1 + \frac{\alpha}{n(1-p^0)} \iff \alpha = n(1-p^0)(k-1).$$

Here we see that we can either fix α and solve for k or vice versa. In the former case, we would typically set $k = \lfloor 1 + \frac{\alpha}{n(1-p^0)} \rfloor$, to ensure that k is integer-valued. Also, recall that $p^c = p^0 - \frac{\alpha}{n}$.

We can then define the score based on the k-fold cross-validation as

$$SCV^{(1)}(Q, p^0, \alpha, \mathbf{y}) = \frac{1}{k} \sum_{j=1}^{k} S(Q(\cdot,\cdot), p^c, \mathbf{y}^{(j)}, \mathbf{y}^{\setminus(j)}). \tag{9}$$

Since the tuning parameter α clearly affects the score, we propose to define a more general score in the next subsection, which takes a variety of choices of α into account.

2.2.1 Using a variety of tuning parameters α

Consider $l \in \mathbb{N}$ distinct values of the tuning parameter α as $\alpha_1, \alpha_2, \ldots, \alpha_l$ and set $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_l)$. For each choice α_i for $i = 1, \ldots, l$, we derive the corresponding n_i^c and p_i^c via the equations (7) and (8). I.e.

$$n_i^c = \frac{n}{1 + \frac{\alpha_i}{n(1-p^0)}}, \qquad\qquad p_i^c = p^0 - \frac{\alpha_i}{n}.$$

Then we set $k_i = \lfloor n/n_i^c \rfloor = \lfloor 1 + \frac{\alpha_i}{n(1-p^0)} \rfloor$.

For a predictor $Q \in \mathbb{Q}$, we define a combined score for prediction of the p^0th quantile by

$$SCV^{(1)}(Q, p^0, \boldsymbol{\alpha}, \mathbf{y}) = \frac{1}{l} \sum_{i=1}^{l} \frac{1}{k_i} \sum_{j=1}^{k_i} S(Q(\cdot,\cdot), p_i^c, \mathbf{y}_i^{(j)}, \mathbf{y}_i^{\setminus(j)}), \tag{10}$$

where $\mathbf{y}_i^{(j)}$ is the j-th sample fold of size $n_i^c = n/k_i$ of the original sample $\mathbf{y}$, and $\mathbf{y}_i^{\setminus(j)}$ denotes the collection of the $k_i - 1$ sub-samples of $\mathbf{y}$ excluding $\mathbf{y}_i^{(j)}$.

This combined score function pools information from assessing a given predictor at multiple and similar extreme trial quantile levels.

Remark 1 We note that we have assigned equal weights to the cross-validation scores corresponding to different choices of α. It will be interesting to explore in more detail whether alternative weights should be considered when combining the individual scores.

We define the optimal predictor as the minimizer of the combined quantile score

$$\widehat{Q}_1 = \widehat{Q}_1(\cdot,\cdot) = \underset{Q \in \mathbb{Q}}{\arg\min}\, SCV^{(1)}(Q, p^0, \boldsymbol{\alpha}, \mathbf{y}). \tag{11}$$

Note that the dependence of $\widehat{Q}_1$ on the choice of the tuning parameter $\boldsymbol{\alpha}$ is not reflected in the notation to simplify the exposition.

Then $\widehat{q}_1(p^0) = \widehat{Q}_1(p^0, \mathbf{y})$ denotes estimated p^0-quantile using the predictor $\widehat{Q}_1$ and the data $\mathbf{y}$.

2.3 Method 2: Large training sample and small validation sample

Alternatively, one could use a traditional k-fold cross-validation set up to compare the competing predictors, where $(k-1)$ sub-samples are used for training and just one sub-sample for validation. Hence we have

$$n^c = (k-1)\frac{n}{k}. \tag{12}$$

The number of folds k needs to be determined depending on the fixed parameters n, p^0, α. Using equations (7) and (12), we solve for k and get

$$k = \frac{n(1-p^0)}{\alpha} + 1 \iff \alpha = n(1-p^0)(k-1)^{-1}.$$

Hence, we set

$$k = \left\lfloor \frac{n(1-p^0)}{\alpha} + 1 \right\rfloor. \tag{13}$$

As before, we also have that

$$p^c = p^0 - \frac{\alpha}{n}. \tag{14}$$

As in Method 1, we can then define the score based on the k-fold cross-validation as

$$SCV^{(2)}(Q, p^0, \alpha, \mathbf{y}) = \frac{1}{k}\sum_{j=1}^{k} S(Q(\cdot,\cdot), p^c, \mathbf{y}^{\setminus(j)}, \mathbf{y}^{(j)}). \tag{15}$$

2.3.1 Using a variety of tuning parameters α

As for Method 1, we can also consider a variety of choices of the tuning parameter α. As before, consider $l \in \mathbb{N}$ distinct values of the tuning parameter α as $\alpha_1, \alpha_2, \ldots, \alpha_l$ and set $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_l)$. For each choice α_i for $i = 1, \ldots, l$, we derive the corresponding n_i^c and p_i^c via the equations (12) and (14). I.e.

$$n_i^c = n^c = (k-1)\frac{n}{k}, \qquad\qquad p_i^c = p^0 - \frac{\alpha_i}{n}.$$

For a predictor $Q \in \mathbb{Q}$, we define a combined score for prediction of the p^0th quantile by

$$SCV^{(2)}(Q, p^0, \boldsymbol{\alpha}, \mathbf{y}) = \frac{1}{l}\sum_{i=1}^{l}\frac{1}{k_i}\sum_{j=1}^{k_i} S(Q(\cdot,\cdot), p_i^c, \mathbf{y}_i^{\setminus(j)}, \mathbf{y}_i^{(j)}), \tag{16}$$

where $\mathbf{y}_i^{(j)}$ is the j-th sample fold of size n/k_i of the original sample $\mathbf{y}$, and $\mathbf{y}_i^{\setminus(j)}$ denotes the collection of the $k_i - 1$ sub-samples of $\mathbf{y}$ excluding $\mathbf{y}_i^{(j)}$.

We define the optimal predictor as the minimizer of the combined quantile score

$$\widehat{Q}_2 = \widehat{Q}_2(\cdot,\cdot) = \operatorname*{arg\,min}_{Q\in\mathbb{Q}} SCV^{(2)}(Q, p^0, \boldsymbol{\alpha}, \mathbf{y}). \tag{17}$$

Note again that the dependence of $\widehat{Q}_2$ on the choice of the tuning parameter $\boldsymbol{\alpha}$ is not reflected in the notation.

Then $\widehat{q}_2(p^0) = \widehat{Q}_2(p^0, \mathbf{y})$ denotes estimated p^0-quantile using the predictor $\widehat{Q}_2$ and the data $\mathbf{y}$.

Description	Requirement	Method 1	Method 2
Training sample size n^c	$\frac{n}{1+\frac{\alpha}{n(1-p^0)}}$	$\frac{n}{k}$	$(k-1)\frac{n}{k}$
Validation sample size $n-n^c$	$\frac{\alpha n}{n(1-p^0)+\alpha}$	$(k-1)\frac{n}{k}$	$\frac{n}{k}$
Number of folds k		$\lfloor 1+\frac{\alpha}{n(1-p^0)} \rfloor$	$\lfloor 1+\frac{n(1-p^0)}{\alpha} \rfloor$
p^c	$p^0-\frac{\alpha}{n}$		
$(n-n^c)(1-p^c)$	α	$n(1-p^0)(k-1)$	$n(1-p^0)(k-1)^{-1}$
$n(1-p^c)$	$n(1-p^0)+\alpha$	$n(1-p^0)k$	$n(1-p^0)\frac{k}{k-1}$

Table 1 Overview of the two methods. Recall that α denotes the expected number of exceedances above the p^cth quantile in the validation sample. The quantity $n(1-p^c)$ represents the expected number of exceedances above the p^c quantile in the entire sample.

2.4 Summary of the two methods and choice of α

We summarise the set-up of Methods 1 and 2 in Table 1. Recall that the tuning parameter α specifies the average number of observations in the validation data that exceeds the p^c- quantile estimated in the training sample. The performance of the predictors are evaluated in the validation sample and the best predictor is the minimizer of the average quantile score. Therefore, the best predictor is the optimum quantile estimate (according to the quantile score) in the validation sample among all quantile predictors under consideration. But for a given validation sample of fixed size, it is only possible to get a good classical optimizer when there are at least few observations larger than the desired quantile. Hence, the tuning parameter α has to be chosen accordingly.

We note that suitable ranges of α will differ depending on which of the two methods is used. While Method 1 can allow for larger values of α (in the sense of $\alpha > 1$), we would expect that in Method 2, we should choose $\alpha < 1$ to reflect the fact that, for extreme quantiles at levels $p^c < p^0$, we would not expect to see exceedances in a relatively small validation sample based on only one fold in the cross-validation.

In our simulation study and empirical work, we consider the choice that $p^0 = 1 - \frac{1}{2n}$, in which case the target quantile is so extreme that we might not even see a single observation above that threshold. We summarise the two methods for this particular case in Table 2.

For Method 1, motivated by work by Wang et al. (2012) and Velthoen et al. (2019), we choose $\alpha \in \mathbb{N} \cap [1, n^{1/4}]$. A restriction to integer-valued αs is not necessary but aids our interpretation. We then determine the corresponding values of k based on the choice of α.

In order to make the comparison between Methods 1 and 2 easier, we then take the values of k derived in Method 1 and convert them to the corresponding values of α in Method 2. For $n = 7500$, this procedure is summarised in Table 3.

Related to our choice of multiple αs, we note that the use of multiple quantile levels to estimate regression coefficients appeared in the quantile regression

Description	Requirement	Method 1	Method 2
Training sample size n^c	$\frac{n}{1+2\alpha}$	$\frac{n}{k}$	$(k-1)\frac{n}{k}$
Validation sample size $n-n^c$	$\frac{2\alpha n}{1+2\alpha}$	$(k-1)\frac{n}{k}$	$\frac{n}{k}$
Number of folds k		$\lfloor 1+2\alpha \rfloor$	$\lfloor 1+\frac{1}{2\alpha} \rfloor$
p^c	$1-\frac{(1+2\alpha)}{2n}$		
$(n-n^c)(1-p^c)$	α	$\frac{1}{2}(k-1)$	$\frac{1}{2}(k-1)^{-1}$
$n(1-p^c)$	$\frac{1}{2}+\alpha$	$\frac{1}{2}k$	$\frac{1}{2}\frac{k}{k-1}$

Table 2 Special case of Table 1 when when $p^0 = 1-\frac{1}{2n}$.

	k	3	5	9	17
Method 1	α	1	2	4	8
	$p^c = 1-\frac{(1+2\alpha)}{n}$	$1-\frac{3}{n}$	$1-\frac{5}{n}$	$1-\frac{9}{n}$	$1-\frac{17}{n}$
Method 2	α	$\frac{1}{2}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$
	$p^c = 1-\frac{(1+2\alpha)}{2n}$	$1-\frac{2}{2n}$	$1-\frac{5}{8n}$	$1-\frac{9}{16n}$	$1-\frac{17}{32n}$

Table 3 Parameter choices for the simulation study: Consider the example when $p^0 = 1-\frac{1}{2n}$. From the formulas in the table we can read of that a choice of $\alpha \in \{1,2,4,8\}$ in Method 1 corresponds to $k \in \{3,5,9,17\}$. Also, choosing $\alpha \in \{1/4, 1/8, 1/16, 1/32\}$ in Method 2 also corresponds to $k \in \{3,5,9,17\}$.

context (Koenker, 1984, 2004; Zou and Yuan, 2008) and in extreme quantile estimation (Wang et al., 2012; Velthoen et al., 2019).

2.5 Examples of extreme quantile predictor

The two commonly used models for predicting extreme quantiles are the Generalized Extreme Value Distribution (GEV) for block maxima and the Generalized Pareto Distribution (GPD) for peaks-over-thresholds of the variable of interest.

GEV model: The GEV(μ,σ,ξ) model is parametrized by the location parameter μ, the scale parameter σ and the shape parameter ξ. Its cumulative distribution function is given by

$$F_Y(y|\mu,\sigma,\xi) = \begin{cases} \exp\left(-\left(1+\xi\frac{y-\mu}{\sigma}\right)^{-1/\xi}\right), & \xi \geq 0, \\ \exp\left(-\exp\left(-\frac{y-\mu}{\sigma}\right)\right), & \xi = 0, \end{cases} \tag{18}$$

where $1+\xi[(y-\mu)/\sigma] > 0$ for $\xi \neq 0$.

For a given level p and a realisation $\mathbf{y}$, the p-th quantile of Y can be estimated as

$$Q(p,\mathbf{y}) = \begin{cases} \hat{\mu} + \frac{\hat{\sigma}}{\hat{\xi}}\left[1-(-\log p)^{-\hat{\xi}}\right], & \text{if } \hat{\xi} \neq 0, \\ \hat{\mu} - \hat{\sigma}\log(-\log p), & \text{if } \hat{\xi} = 0, \end{cases} \tag{19}$$

where $\hat{\mu}$, $\hat{\sigma}$ and $\hat{\xi}$ are estimates of μ, σ and ξ.

GPD model: Let us assume that there exists a non-degenerate limiting distribution for appropriately linearly rescaled excesses of a sequence of independently and identically distributed observations $Y_1, \ldots, Y_n$ above a threshold u. Then under general regularity conditions, the limiting distribution will be a GPD as $u \to \infty$ (Pickands, 1975). The GPD is parametrized by a shape parameter ξ and a threshold-dependent scale parameter σ_u, with cumulative distribution function F,

$$F_Y(y|u, \sigma_u, \xi) = \begin{cases} 1 - \left[1 + \xi \left(\frac{y-u}{\sigma_u}\right)\right]^{-1/\xi}, & y \geq u, \xi \geq 0, \\ 1 - \exp\left[\frac{y-u}{\sigma_u}\right], & u < y < u - \sigma_u/\xi, \xi < 0, \end{cases} \tag{20}$$

and $u \in \mathbb{R}$ and $\sigma_u > 0$.

The GPD model depends on a threshold u, which needs to be chosen up front. The choice of the threshold is a crucial issue, as too low of a threshold leads to a bias from model misspecification, and too high of a threshold increases the variance of the estimators: a bias-variance trade-off. Many threshold selection procedures have been proposed in the literature, see (Davison and Smith, 1990; Drees et al., 2000; Coles, 2001), among many others.

Once an appropriate threshold has been selected, the parameters can be estimated by the maximum likelihood method and, subsequently, the target high quantile estimate can be extracted by plugging in the estimates of u, ξ and σ_u in

$$Q(p, \mathbf{y}) = \begin{cases} u + \frac{\hat{\sigma}}{\hat{\xi}} \left[\left(\frac{\hat{\zeta}_u}{1-p}\right)^{\hat{\xi}} - 1\right], & \text{if } \xi \neq 0, \\ u + \hat{\sigma} \log\left(\frac{\hat{\zeta}_u}{1-p}\right), & \text{if } \xi = 0, \end{cases} \tag{21}$$

where $\hat{\sigma}$, and $\hat{\xi}$ are the maximum likelihood estimates based on the GPD assumption for the exceedances ((Coles, 2001), Section 4.3.3). Also, $\zeta_u = P(Y > u)$ can be approximated by the ratio $\frac{n_u}{n}$ where n_u is the number of exceedances above the threshold u in the given sample of size n.

3 Simulation Study

In this section, we carry out a simulation study to examine the performance of the optimum score predictors $\widehat{Q}_1$ and $\widehat{Q}_2$ obtained through the proposed scoring methods and compare them with that of the conventional optimal quantile score predictor $\widehat{Q}_{qs}$ at large quantiles.

We simulate samples from the random variable Y following the four sets of distribution with densities:

(i) $f(y) = f_{GPD(10,\,1,\,\xi)}(y)$ with three choices of the shape parameter ξ: (a) ξ= -0.5 , (b) ξ=0, and (c) ξ=0.5,

(ii) $f(y) = \lambda f_{U(0,\,10)}(y) + (1-\lambda) f_{GPD(10,\,1,\,0.5)}(y)$, with two different choices of the weight parameter λ: (a) λ= 0.5 and (b) λ=0.99,

(iii) $f(y) = 0.5 f_{GPD(10,\ 1,\ 0.1)}(y) + 0.5 f_{GPD(10,\ 1,\ 0.5)}(y)$,
(iv) $f(y) = f_{Gamma(1,1,0.1)}(y)$,

where $f_{U(0,10)}$ denotes the probability density function (pdf) of the Uniform distribution (0, 10) and $f_{GPD(\mu,\ \sigma,\ \xi)}$ denotes the pdf of the GPD with location μ, scale σ, and shape ξ. The three different values of the shape parameter in model-(i) correspond to the three types of the extreme value families: $\xi = -0.5$ (Weibull), $\xi = 0$ (Gumbel) and $\xi = 0.5$ (Frechet). We choose the standard scale $\sigma = 1$ and location $\mu = 10$ in each of the cases: (i) to (iii). The model-(ii), is a mixture with mixing probability λ such that it produces lower 100λ percent samples from $U(0, 10)$ and upper 100(1-λ) percent samples from the GPD. We choose $\lambda = 0.5$ and $\lambda = 0.99$ in model-(ii). The model-(iii) is obtained by equally mixing two GPDs with different shape parameters. Model-(iv) is the Gamma distribution with rate, scale and shape parameter taking values as 1, 1 and 0.1, respectively.

For each of the simulated datasets, we apply the conventional method $\widehat{Q}_{qs}$ and the proposed scoring methods $\widehat{Q}_1$ and $\widehat{Q}_2$ to evaluate and select the optimum score predictors of the quantile at the level $p^0 = (1 - \frac{1}{2n})$ based on a sample of size n. We take sample size $n = 7500$, this value of n can be thought of as 50 years of daily observations (150 days each season) in the precipitation rate data analyzed in Section 4.2.

As a competing set of predictors for the target quantile level p^0, we consider the GPD-model-based predictors given in (21) estimated by the maximum likelihood method. For selecting the threshold in the GPD model, we consider the "rules-of-thumb" approach (Ferreira et al., 2003), which uses a fixed upper fraction of data to fit the GPD. We consider two sets of thresholds. The first set is based on ten upper order statistics: 150, 125, 100, 75, 50, 40, 30, 20, 10 and 3, and refer to this set as A. Here the largest order statistics corresponds to the lowest GPD threshold. We label the predictors Q_i where, for the index i, we use numbers starting from 1 to 10, such that, 1 and 10 correspond to the predictors with the highest and lowest value of the upper order statistics, respectively.

Note that, each of the members of the set A uses an equal number of observations fitted to the GPD in both the training and original prediction. We then consider another set of thresholds, determined by sample percentiles with probability levels: 0.98, 0.9833, 0.9867, 0.99, 0.993, 0.995, 0.996, 0.9973, 0.9987, and 0.9996, and refer this set as B. These probability levels are chosen such that the number of exceedances is approximately the same for A and B for the original sample of size $n = 7500$. Note that for each of the members of B, the number of samples fitted to the GPD in the original prediction is larger than those in the training stage (as the values of the thresholds are different for different sample sizes). For notational convenience, as in set A, we index the members of B by the numbers starting from 11 to 20. We also added the sample quantile at level p^0 as a predictor in the competing set and denote it as 0-th estimate. Here, out of sample quantiles are estimated by the maximum of the data set.

For choosing the values of the tuning parameter α for $\widehat{Q}_1$ and $\widehat{Q}_2$, we use the strategy summarised in Table 3, resulting in $\boldsymbol{\alpha} = (1, 2, 4, 8)$ (where $[n^{1/4}] \approx 9$). We found that the use of additional values of α does not change the performance of the proposed methods noticeably.

We generate $L = 1000$ replicates each of size $n = 7500$ from the models (i)–(iv). For each of the simulated dataset, we use three methods, $\widehat{Q}_{qs}$, $\widehat{Q}_1$ and $\widehat{Q}_2$ to find the optimum quantile score predictors chosen from the competing predictor-set of p^0th quantile, and for each of the three methods, we compute the root mean squared error (RMSE) across L iterations as

$$RMSE = \sqrt{\frac{1}{L}\sum_{i=1}^{L}(\widehat{q}_{\cdot}^{(i)}(p^0) - q_Y(p^0))^2}, \tag{22}$$

where $\widehat{q}_{\cdot}^{(i)}(p^0)$ is the optimal p^0-quantile prediction in the ith iteration, which is based on any of the three predictors: $\widehat{Q}_{qs}$, $\widehat{Q}_1$ and $\widehat{Q}_2$. For instance, if $\widehat{Q}_1$ is the optimal predictor in the ith iteration, then

$$\widehat{q}_{\cdot}^{(i)}(p^0) = \widehat{q}_1^{(i)}(p^0) = \widehat{Q}_1(p^0, \mathbf{y}).$$

Table 4 shows the RMSE of the three optimum predictors $\widehat{Q}_{qs}$, $\widehat{Q}_1$ and $\widehat{Q}_2$ chosen from the sets of the predictors A, B and union of the two sets, $A \cup B$. The first proposed approach $\widehat{Q}_1$ outperforms the conventional method $\widehat{Q}_{qs}$ in most of the cases. The only exception is model-(ii) : (b), where the optimum predictor $\widehat{Q}_1$ produces a higher RMSE than $\widehat{Q}_{qs}$. The above data generating model consists of fewer GPD components, on average only 75 GPD observations in the whole sample. This returns roughly four GPD samples in the training data of size $n_i^c = 441$ for a given $\alpha_4 = 8$. An inadequate sample from the GPD implies that most of the competing GPD based predictors in $A \cup B$ are poorly fitted, thus resulting in a high bias and a high RMSE. An increase of the GPD component in (ii) : (a) (50 per cent data from GPD) returns a lower RMSE for $\widehat{Q}_1$ compared to $\widehat{Q}_{qs}$. It is also found that the presence of only 5 per cent of the GPD component in model-(ii) (for $n = 7500$ and $\alpha = 8$ it provides at least 20 GPD random sample in the training sample) enables the competing predictors to fit the GPD reasonably well. In this scenario, $\widehat{Q}_1$ outperforms the other two methods (results are not reported here).

Note that, the conventional scoring method $\widehat{Q}_{qs}$ uses the whole sample of size 7500 to train the predictor set. Therefore, in the case of model-$(ii), b$, all the competing predictors are trained on at least 75 GPD sample points (1 per cent of the total sample). Thus they are reasonably well fitted to the GPD, and $\widehat{Q}_{qs}$ outperforms $\widehat{Q}_1$ (row 4 of table 4). It can be concluded that if the competing GPD based predictors are reasonably well fitted in the training sample, then the proposed method $\widehat{Q}_1$ performs better than $\widehat{Q}_{qs}$. The latter estimator is more efficient than the former when the data contains only a few extreme components.

Table 4 RMSE of the optimum score predictors $\widehat{Q}_{qs}$, $\widehat{Q}_1$ and $\widehat{Q}_2$ for the quantile at probability level $p^0 = (1 - \frac{1}{2n})$, based on samples of size $n = 7500$ and using $\alpha = \{1, 2, 4, 8\}$. All the results are based on 1000 iterations.

	AB			A			B		
Model	$\widehat{Q}_{qs}$	$\widehat{Q}_1$	$\widehat{Q}_2$	$\widehat{Q}_{qs}$	$\widehat{Q}_1$	$\widehat{Q}_2$	$\widehat{Q}_{qs}$	$\widehat{Q}_1$	$\widehat{Q}_2$
$(i) : (a)$	0.012	0.011	0.012	0.012	0.013	0.013	0.012	0.010	0.012
$(i) : (b)$	1.171	0.558	0.934	1.172	1.120	1.141	1.170	0.473	0.943
$(i) : (c)$	585.95	68.70	112.95	585.77	69.21	110.87	586.10	44.47	109.15
$(ii) : (a)$	230.92	87.265	92.63	228.92	192.03	92.81	228.92	90.72	92.62
$(ii) : (b)$	25.175	411.714	62.334	25.15	28.86	23.15	25.15	12.67	22.68
(iii)	231.256	97.315	98.983	229.29	192.66	94.58	229.28	93.12	98.36
(iv)	1.114	1.054	1.021	1.107	1.031	1.027	1.108	1.031	1.027

In contrast to $\widehat{Q}_1$, the second approach $\widehat{Q}_2$ uses most of the data to train the competing predictors. For example, in model $(ii) : (b)$ the method uses a training sample of size 5000 (with $k = 3$ in (12)) consisting of 50 GPD sample. The presence of a sufficient number of GPD samples enables all the predictors to be well fitted to the GPD, and the optimum predictor $\widehat{Q}_2$ produces a smaller error than $\widehat{Q}_1$. Table 4 also shows that $\widehat{Q}_2$ outperforms the classical optimum score predictor $\widehat{Q}_{qs}$ in all the cases. However, it is evident that, in general, $\widehat{Q}_1$ is more efficient than $\widehat{Q}_2$. The superiority of $\widehat{Q}_2$ over $\widehat{Q}_1$ in model-$(ii) : (b)$, could be because of a better estimation of the competing quantile predictors (based on a larger training sample of size 6750) and at the same time having a reasonable amount of exceedances in the cross-validation (for example $\alpha = 8$).

Figure 1 shows the frequency distributions of the competing predictors set $A \cup B$ selected as optimum through $\widehat{Q}_{qs}$, $\widehat{Q}_1$ and $\widehat{Q}_2$ when the data is generated from the mixture model-$(ii) : (a)$. It is found that $\widehat{Q}_{qs}$ uniformly selects the predictors with slightly more preference to higher threshold based predictors. Whereas $\widehat{Q}_1$ frequently selects lower thresholds, indicating a preference to the predictors based on a large number of observations to be fitted to the GPD. It is also found that $\widehat{Q}_2$ selects the lower thresholds of both A and B more frequently. It can be said that $\widehat{Q}_2$ has less discriminate power compared to $\widehat{Q}_1$ for distinguishing sets A and B. Note that, for the original sample of size $n = 7500$, the training sample size for $\widehat{Q}_2$ turns out to more than 5000 giving a reasonably large sample to estimate all the predictors of $A \cup B$. When the training sample size is reduced to 2500, the method tends to select the predictor from the set A (the results are omitted for brevity).

We then compare the above three methods with the median of the competing predictors and a randomly chosen predictor from the competing set. As before, we use the predictor set $\{0\} \cup A \cup B$ for estimating the quantile at probability level $p^0 = 1 - \frac{1}{2n}$ and use the tuning parameter $\alpha \in \{1, 2, 4, 8\}$. Table 5 shows the RMSE of the optimum score predictors $\widehat{Q}_{qs}$, $\widehat{Q}_1$, $\widehat{Q}_2$, along with those of the random and the median prediction. As expected, the performance of the random predictor is worst compared to the other predictors.

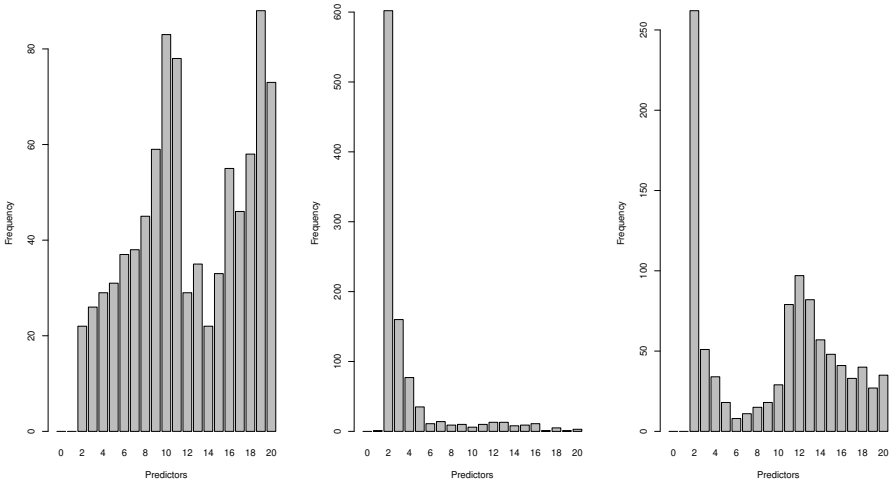

Fig. 1 Frequency distributions of optimum predictors through the scoring methods $\widehat{Q}_{qs}$, $\widehat{Q}_1$ and $\widehat{Q}_2$ plotted in the left, middle and right panel, respectively. The data are simulated from $(ii) : (a)$. The empirical quantile estimate is denoted by '0', the members of the sets A and B are denoted by the numbers 1 to 10 (lower index for lower threshold) and 11 to 20 (lower index for lower threshold), respectively.

Table 5 RMSE of the optimum score predictors $\widehat{Q}_{qs}$, $\widehat{Q}_1$, $\widehat{Q}_2$, median predictor $\widehat{Q}_{med}$ and random predictors $\widehat{Q}_{rand}$ for the quantile at probability level $p^0 = (1 - \frac{1}{2n})$, based on samples of size $n = 7500$. All the results are based on 1000 iterations.

Model	$\widehat{Q}_{qs}$	$\widehat{Q}_1$	$\widehat{Q}_2$	$\widehat{Q}_{med}$	$\widehat{Q}_{rand}$
$(i) : (a)$	0.012	0.011	0.012	0.013	0.013
$(i) : (b)$	1.171	0.558	0.934	1.067	1.137
$(i) : (c)$	585.95	68.70	112.95	133.94	169.29
$(ii) : (a)$	30.92	87.265	92.63	87.17	118.60
$(ii) : (b)$	25.175	411.714	62.334	16.58	322.22
(iii)	231.256	97.315	98.983	107.93	112.74
(iv)	1.114	1.054	1.021	0.989	0.982

The predictor $\widehat{Q}_1$ dominates the median prediction except model $(ii) : (b)$ and (iv).

4 Assessment of quantile predictions at high tails: real data examples

In this section, we illustrate the proposed scoring methods with two different real data examples: cyber Netflow bytes transfer and daily precipitation.

4.1 LANL cyber netflow data

The unified Netflow and host event dataset, available from Los Alamos National Laboratory (LANL, https://csr.lanl.gov/data/2017.html), thoroughly described in Turcotte et al. (2018), is one of the most commonly used datasets in cybersecurity research for anomaly detection with enormous importance in industry and society (Adams and Heard, 2016). The data comprise records describing communication events between various devices connected to the LANL enterprise. The daily Netflow data consisting of hundreds of thousands of records is available for 90 days (day 2 to day 91, starting from a specific epoch time). Each of the flow records (Netflow V9) is an aggregate summary of a bi-directional network communication with the following components: *StartTime, EndTime, SrcIP, DstIP, Protocol, SrcPort, DstPort, SrcPackets, DstPackets, SrcBytes* and *DstBytes* (Turcotte et al., 2017). In this work, we analyze one important component of this data set.

As a variable of interest, we consider the total volume of bytes transfer (through an edge) which started in an epoch time. The data is available for every second of the day, so in a day we have n=86400 observations. Figure 2 shows the histogram of the byte transfer in a typical day (day 3, chosen randomly from the available 90 days), indicating the heavy-tailed nature of the underlying distribution. We consider the target quantile as $p^0 = (1 - \frac{1}{2n})$ and use the competing set of 21 predictors $\{0\} \cup A \cup B$. As in the simulation study, we consider the same set of values of the tuning parameter k (or equivalently α) as $\{3, 5, 719\}$.

As the variable of interest, we consider the total volume of bytes transfer. The optimum scoring predictors through methods $\widehat{Q}_{qs}$, $\widehat{Q}_1$ and $\widehat{Q}_2$ turn out to be based on 20th upper order statistics and percentiles with probability levels 0.99467 and 0.996. The optimal predictions of p^0th quantile of total bytes transfer using the above three methods are 689.24 gigabytes, 2070.581 gigabytes and 1007.86 gigabytes, respectively.

4.2 GHCN daily precipitation data

Prediction of extreme high return values of daily precipitation is a common task in hydrological engineering where these values correspond to return periods of 100 or 1000 years based on relatively few years of data (Bader et al., 2018). Daily precipitation data is available from Global Historical Climatology Network (GHCN) and can be downloaded freely from ftp://ftp.ncdc.noaa.gov/pub/data/ghcn/daily/ for tens of thousands of surface weather sites around the world (Menne et al., 2012). To illustrate the proposed methodology, we consider a sample station chosen randomly from the set of stations in one of the US coastal state, California.

We consider precipitation data for the months, November to March for each year for the period 1940 to 2015 (following Bader et al. (2018)) with the total number of daily data points as n =3303 after excluding zeros and some

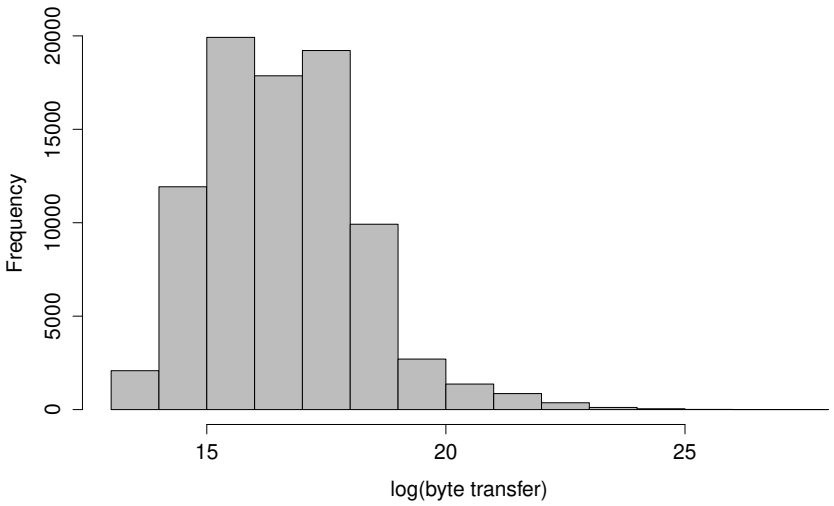

Fig. 2 Histogram of the total volume of bytes transfer in a second on a particular day.

missing observations. Figure 3 shows the histogram of the daily precipitation data, indicating the heavy-tailed nature of the response variable. Here we set our target as to predict the $2n$-observation (150-year) return level, implying $p^0 = (1 - \frac{1}{2n})$. As a set of predictors, we consider 21 unconditional predictors $\{0\} \cup A \cup B$ considered in Section 3. As the values of k (or equivalent values of α), we take the set of values (3, 5, 7, 17). The optimum predictors using the three methods $\widehat{Q}_{qs}$, $\widehat{Q}_1$ and $\widehat{Q}_2$ are the 20th, fourth and the first member of the competing set and the optimum predictions are 1847.21, 2227.97 and 2727.22 millimetres, respectively.

5 Concluding remarks

We have proposed two distribution-free scoring methods for extreme high quantile predictors to assess their predictive performances. These methods rely on cross-validation by partitioning the sample of n into k folds. The first proposed method evaluates the competing predictors by employing them to predict a series of equally extreme quantiles on a fold of size $\frac{n}{k}$ chosen randomly from the given data set. The predictors are then evaluated by their predictive performance in the other parts of the sample of sizes $(k-1)n/k$. The choice of the number of folds k (obtained through choosing values of α) and equally extreme quantile points are obtained by solving two equations for a given tuning parameter set $\boldsymbol{\alpha}$. The second method uses $k-1$ folds, all at a time, to train the predictors and test them on one fold. A numerical

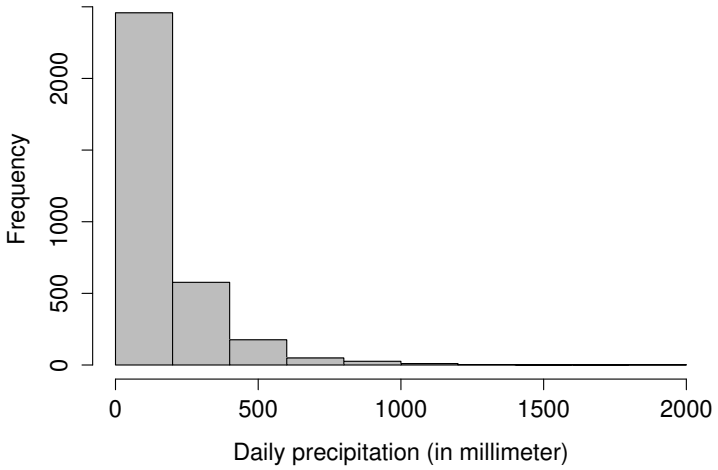

Fig. 3 Histogram of the daily precipitation during the time period 1945-2010 at a met-station in California.

study shows the increased efficiency of the proposed methods compared to the conventional quantile scoring method.

The choice of the tuning parameter $\boldsymbol{\alpha}$ is a crucial issue. For a sample of size n and the given target probability, $p_0 = 1 - \frac{1}{2n}$, a set of choices of the tuning parameter α for the two proposed method is given here. We also compare the choice of αs in the two proposed methods by fixing the number of folds (k). Our simulation study provides a prescription for the choice of the tuning parameter through theoretical justification from quantile regression. This work could be extended by incorporating covariate information in the quantile evaluation and by further investigating the choice of the tuning parameter α.

Acknowledgements The research of Kaushik Jana is supported by the Alan Turing Institute- Lloyd's Register Foundation Programme on Data-Centric Engineering. The authors also like to thank Dr Tobias Fissler of Vienna University of Economics and Business for many helpful discussions.

Statement on Conflict of Interest: The authors declare that there is no conflict of interest.

References

Adams, N. and Heard, N. (2016). *Dynamic Networks and Cyber-Security*. World Scientific (Europe).

Bader, B., Yan, J., and Zhang, X. (2018). Automated threshold selection for extreme value analysis via ordered goodness-of-fit tests with adjustment for false discovery rate. *Ann. Appl. Stat.*, 12(1):310–329.

Bentzien, S. and Friederichs, P. (2014a). Decomposition and graphical portrayal of the quantile score. *Quarterly Journal of the Royal Meteorological Society*, 140(683):1924–1934.

Bentzien, S. and Friederichs, P. (2014b). Decomposition and graphical portrayal of the quantile score. *Quarterly Journal of the Royal Meteorological Society*, 140(683):1924–1934.

Brehmer, J. R. and Strokorb, K. (2019). Why scoring functions cannot assess tail properties. *Electron. J. Statist.*, 13(2):4015–4034.

Coles, S. G. (2001). *An Introduction to Statistical Modeling of Extreme Values*. Springer, London.

Davison, A. C. and Smith, R. L. (1990). Models for exceedances over high thresholds. *Journal of the Royal Statistical Society. Series B (Methodological)*, 52(3):393–442.

Drees, H., de Haan, L., and Resnick, S. (2000). How to make a hill plot. *Ann. Statist.*, 28(1):254–274.

Embrechts, P., Klüppelberg, C., and Mikosch, T. (1997). *Modelling Extremal Events*. Berlin: Springer.

Ferreira, A., de Haan, L., and Peng, L. (2003). On optimising the estimation of high quantiles of a probability distribution. *Statistics*, 37(5):401–434.

Fisher, R. A. and Tippett, L. H. C. (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample. *Mathematical Proceedings of the Cambridge Philosophical Society*, 24(2):180–190.

Friederichs, P. and Thorarinsdottir, T. L. (2012). Forecast verification for extreme value distributions with an application to probabilistic peak wind prediction. *Environmetrics*, 23(7):579–594.

Gneiting, T. (2011). Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106(494):746–762.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378.

Koenker, R. (1984). A note on l-estimates for linear models. *Statistics & Probability Letters*, 2(6):323 – 325.

Koenker, R. (2004). Quantile regression for longitudinal data. *Journal of Multivariate Analysis*, 91(1):74 – 89.

Lerch, S., Thorarinsdottir, T. L., Ravazzolo, F., and Gneiting, T. (2017). Forecaster's dilemma: Extreme events and forecast evaluation. *Statist. Sci.*, 32(1):106–127.

Menne, M. J., Durre, I., Vose, R. S., Gleason, B. E., and Houston, T. G. (2012). An overview of the global historical climatology network-daily database. *Journal of Atmospheric and Oceanic Technology*, 29(7):897–910.

Pickands, J. (1975). Statistical inference using extreme order statistics. *Ann. Statist.*, 3(1):119–131.

Scheuerer, M. and Moller, D. (2015). Probabilistic wind speed forecasting on a grid based on ensemble model output statistics. *Ann. Appl. Stat.*, 9(3):1328–1349.

Turcotte, M. J. M., Kent, A. D., and Hash, C. (2017). Unified host and network data set. *ArXiv*, 1708.07518v1.

Turcotte, M. J. M., Kent, A. D., and Hash, C. (2018). *Unified Host and Network Data Set*, chapter Chapter 1, pages 1–22. World Scientific.

Velthoen, J., Cai, J.-J., Jongbloed, G., and Schmeits, M. (2019). Improving precipitation forecasts using extreme quantile regression. *Extremes*.

Wang, H. J., Li, D., and He, X. (2012). Estimation of high conditional quantiles for heavy-tailed distributions. *Journal of the American Statistical Association*, 107(500):1453–1464.

Zou, H. and Yuan, M. (2008). Composite quantile regression and the oracle model selection theory. *Ann. Statist.*, 36(3):1108–1126.