

Extreme Conformal Prediction: Reliable Intervals for High-Impact Events

OLIVIER C. PASCHE^{1,2,*}, HENRY LAM², AND SEBASTIAN ENGELKE¹

¹*Research Institute for Statistics and Information Science, University of Geneva, Switzerland*

²*Department of Industrial Engineering and Operations Research, Columbia University, New York, USA*

^{*}*Corresponding author: olivier.pasche@unige.ch*

Abstract

Conformal prediction is a popular method to construct prediction intervals for black-box machine learning models with marginal coverage guarantees. In applications with potentially high-impact events, such as flooding or financial crises, regulators often require very high confidence for such intervals. However, if the desired level of confidence is too large relative to the amount of data used for calibration, then classical conformal methods provide infinitely wide, thus, uninformative prediction intervals. In this paper, we propose a new method to overcome this limitation. We bridge extreme value statistics and conformal prediction to provide reliable and informative prediction intervals with high-confidence coverage, which can be constructed using any black-box extreme quantile regression method. The advantages of this extreme conformal prediction method are illustrated in a simulation study and in an application to flood risk forecasting.

1 Introduction

Conformal prediction is a simple approach to producing prediction sets from any regression or classification model. For a covariate vector \mathbf{X} with values in $\mathcal{X} \subseteq \mathbb{R}^p$ and corresponding response variable Y , the goal of classical conformal prediction is to build a prediction set $C(\mathbf{x})$ satisfying marginal coverage

$$\mathbb{P}(Y_{\text{test}} \in C(\mathbf{X}_{\text{test}})) \geq 1 - \alpha, \quad (1)$$

for a desired confidence level $1 - \alpha \in (0, 1)$, for any new observations $(\mathbf{X}_{\text{test}}, Y_{\text{test}})$. To obtain such a prediction set, we assume that a prediction model \hat{f} was fitted on a training data set from the distribution \mathcal{P} of (\mathbf{X}, Y) , and that a new calibration data set $(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_{n_c}, Y_{n_c})$ of n_c independent samples from the same distribution is available. For a specific score function $s : \mathcal{X} \times \mathbb{R} \rightarrow \mathbb{R}$ that may depend on \hat{f} and acts on the predictors and responses, consider the calibration scores $s_i = s(\mathbf{X}_i, Y_i)$ for $i = 1, \dots, n_c$. For some level $\alpha \in (0, 1)$, denoting by \hat{q}_α the $(\lceil (n_c + 1)(1 - \alpha) \rceil / n_c)$ -quantile of the scores s_1, \dots, s_{n_c} , the prediction set

$$C(\mathbf{x}) = \{y : s(\mathbf{x}, y) \leq \hat{q}_\alpha\}. \quad (2)$$

has the desired $1 - \alpha$ coverage. That is, it satisfies [Eq. \(1\)](#) for a new test sample $(\mathbf{X}_{\text{test}}, Y_{\text{test}})$ from the same distribution \mathcal{P} . This well-established framework is the so-called split conformal approach (e.g. [Lei et al., 2018](#); [Romano et al., 2019](#)). More generally, originally started in [Vovk et al. \(2022\)](#), the

conformalization idea that leverages quantile-based construction of prediction sets elicits a range of variants, with focus on optimal data usage and applying to different problems, including jackknife+ (Alaa and van der Schaar, 2020; Barber et al., 2021), cross-conformal prediction (Vovk, 2015) and ensemble-based approaches (Kim et al., 2020; Gupta et al., 2022).

Conformal prediction intervals are widely used for confidence levels $1 - \alpha$ of moderate value relative to the sample size n_c (e.g., 90% and $n_c = 1000$), where enough of the calibration scores are above this quantile. In many applications, however, test points $Y_{\text{test}} \notin C(X_{\text{test}})$ that fall outside of the confidence set correspond to a high-impact event with serious consequences for the environment, human lives or the economy. Examples for such risk-sensitive applications are the protection of cities and energy infrastructure from flooding (Keef et al., 2009; Asadi et al., 2015) or the financial reserves of banks and insurance companies (van Oordt and Zhou, 2019; Dupuis et al., 2023). In these cases, much larger values of confidence $1 - \alpha$, close to one, will be required, sometimes even by law. Classical methods from conformal prediction fail for those requirements, since the quantile \hat{q}_α as defined above is not a useful estimator when the level $\alpha < 1/(n_c + 1)$. Indeed, in this case, less than one observation then exceeds the $(1 - \alpha)$ -quantile on average, and \hat{q}_α is infinite (or ill-defined). Even for slightly larger values α close to that limit, the variance of \hat{q}_α can be huge.

Extreme value theory provides statistical tools for accurate estimation beyond the data range (de Haan and Ferreira, 2006). The tools have proven successful for improving extrapolation properties of machine learning methods in regression (de Carvalho et al., 2022; Huet et al., 2024; Buriticá and Engelke, 2024), classification (Jalalzai et al., 2018) and generative methods (Boulaguiem et al., 2022).

In this paper, we propose a new methodology that bridges the wide applicability of conformal prediction with extrapolation tools from extreme value statistics to construct reliable prediction sets for high-impact events. In a first step, in order to obtain a good pretrained model \hat{f} beyond the data range, we rely on flexible machine learning methods from extreme quantile regression (Velthoen et al., 2019; Gnecco et al., 2024; Pasche and Engelke, 2024; Richards and Huser, 2024). Second, we rely on the classical and theoretically justified peaks-over-threshold approach, which consists of using the generalised Pareto distribution (GPD) to extrapolate, for example, quantile estimates beyond the range of empirical observations (Balkema and de Haan, 1974; Pickands III, 1975). For a confidence level $1 - \alpha$ close to one, we leverage the GPD fitted to the calibration scores s_1, \dots, s_{n_c} to obtain a reliable estimate of \hat{q}_α beyond the calibration data. The resulting extreme conformal prediction intervals have better properties compared to those from the classical empirical approach for large confidence requirements. In a simulation study, we show that our method improves existing approaches in terms of better coverage, in the sense of Eq. (1), and of informativeness of the prediction interval.

We illustrate the advantages of our approach in an application to flood risk assessment. Using several of the flexible machine learning methods as base predictions, it provides high-confidence one-day-ahead interval forecasts of the conditional range for water flow. We show that using conformal prediction intervals based on extreme value theory improves the coverage of the classical method, which either yields uninformative intervals or exhibits undercoverage and, therefore, seriously underestimates the risk of high-impact events.

2 Background on conformal prediction

2.1 Split conformal prediction

As in the introduction, let X be a covariate vector taking values in $\mathcal{X} \subseteq \mathbb{R}^p$, and Y the response random variable of interest. We will consider, here and in the sequel, the regression case where the response Y is real-valued in \mathbb{R} . We also suppose that we have access to a calibration set of

n_c independent observations $(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_{n_c}, Y_{n_c})$. Classical (split) conformal prediction builds prediction intervals (PIs) as in Eq. (2) that have desired coverage under very weak assumptions. For instance, if the calibration set is an exchangeable sample, and the test point $(\mathbf{X}_{\text{test}}, Y_{\text{test}})$ is independent and of the same distribution, then the unconditional coverage guarantee Eq. (1) holds (Papadopoulos et al., 2002).

Importantly, the probability measure in Eq. (1) is with respect to the randomness in the calibration set jointly with the test point, that is, $\{(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_{n_c}, Y_{n_c}), (\mathbf{X}_{\text{test}}, Y_{\text{test}})\}$. In fact, when conditioning on the calibration set, the distribution of the coverage is a beta distribution, that is,

$$\mathbb{P}\left(Y_{\text{test}} \in C(\mathbf{x}) \mid \{(\mathbf{X}_i, Y_i)\}_{i=1}^{n_c}\right) \sim \text{Beta}(n+1-l, l), \quad l := \lfloor (n+1)\alpha \rfloor. \quad (3)$$

Running the conformal prediction twice on different calibration sets, therefore yields PIs with different coverage probabilities. The guarantee in Eq. (1) says that when averaging out the calibration set, the coverage is at least $1-\alpha$.

Furthermore, the marginal coverage property in Eq. (1) only guarantees “overall” marginal coverage of the prediction set $C(\mathbf{x})$, but does not imply the conditional coverage property

$$\mathbb{P}(Y_{\text{test}} \in C(\mathbf{x}) \mid \mathbf{X}_{\text{test}} = \mathbf{x}) \geq 1-\alpha, \quad \forall \mathbf{x} \in \mathcal{X}. \quad (4)$$

The latter is generally impossible to guarantee in such a general setting. How close $C(\mathbf{x})$ is to satisfying the conditional coverage property depends on the quality of the given pretrained model. For example, for the conformalized quantile regression approach described in Section 2.2, it depends on the accuracy of the initial quantile regression model $\hat{Q}_{\alpha/2}(\mathbf{x}), \hat{Q}_{1-\alpha/2}(\mathbf{x})$.

2.2 Conformalized quantile regression

Conformalized quantile regression was first proposed by Romano et al. (2019) and is also described in Angelopoulos and Bates (2023). Suppose that we have access to a black-box quantile regression model trained to estimate the conditional quantiles $\hat{Q}_{\alpha/2}(\mathbf{x})$ and $\hat{Q}_{1-\alpha/2}(\mathbf{x})$ of Y , given $\mathbf{X} = \mathbf{x}$, at probability levels $\alpha/2$ and $1-\alpha/2$, respectively. Then, conformalized quantile regression uses the score function

$$s(\mathbf{x}, y) := \max\{\hat{Q}_{\alpha/2}(\mathbf{x}) - y, y - \hat{Q}_{1-\alpha/2}(\mathbf{x})\}. \quad (5)$$

Following the general procedure described in the introduction, this leads to the final prediction set in Eq. (2) being the interval

$$C(\mathbf{x}) = [\hat{Q}_{\alpha/2}(\mathbf{x}) - \hat{q}_\alpha, \hat{Q}_{1-\alpha/2}(\mathbf{x}) + \hat{q}_\alpha], \quad (6)$$

where \hat{q}_α is the $(\lceil (n_c+1)(1-\alpha) \rceil / n_c)$ -quantile of the calibration scores s_1, \dots, s_{n_c} , i.e. the order statistic $s_{(\lceil (n_c+1)(1-\alpha) \rceil)}$.

Intuitively, the procedure either widens (with a positive \hat{q}_α) or narrows (with a negative \hat{q}_α) the initial interval $[\hat{Q}_{\alpha/2}(\mathbf{x}), \hat{Q}_{1-\alpha/2}(\mathbf{x})]$ so that it covers $\lceil (n_c+1)(1-\alpha) \rceil$ of the n_c calibration observations. Note that the resulting prediction intervals satisfy the marginal coverage Eq. (1), but there is no guarantee that the conditional coverage in Eq. (4) is satisfied. In fact, the more accurate the initial quantile regression models $\hat{Q}_{\alpha/2}(\mathbf{x})$ and $\hat{Q}_{1-\alpha/2}(\mathbf{x})$ are, the better the conditional coverage will be.

2.3 Limitation for extreme confidence levels

For high-impact events, regulators often require predictions with very high coverage probabilities to ensure that protective infrastructures or measures are sufficient. In particular, in such risk-sensitive applications, the level α in Eq. (1) is typically close to 0 and may satisfy $\alpha < 1/(n_c+1)$. This is

generally referred to as an extreme confidence or probability level since, on average, there is less than one observation above the $(1 - \alpha)$ -quantile in a sample of size n_c . Note that the size n_c of the calibration set is typically fairly small, since these data cannot be used for model fitting, often resulting in extreme scenarios even for relatively moderate levels of α .

The classical construction of the conformal prediction intervals described in the introduction requires the computation of \hat{q}_α , the empirical $(\lceil (n_c + 1)(1 - \alpha) \rceil / n_c)$ -quantile of the calibration set scores. For extreme confidence levels $1 - \alpha$, this quantile level

$$\lceil (n_c + 1)(1 - \alpha) \rceil / n_c > \lceil n_c \rceil / n_c = 1,$$

in which case \hat{q}_α is ill-defined and, by convention, set to infinity (Romano et al., 2019; Angelopoulos and Bates, 2023). This results in degenerate trivial prediction intervals $C(\mathbf{x}) = (-\infty, \infty)$, for all $\mathbf{x} \in \mathcal{X}$. Although this interval satisfies the coverage Eq. (1), it is of no practical utility.

3 Extreme conformal prediction

We propose an approach based on extreme value statistics to construct non-degenerate conformal prediction intervals at extreme confidence levels $1 - \alpha > n_c / (n_c + 1)$. Similarly to classical conformalized quantile regression (Romano et al., 2019), our method requires two steps:

1. fitting a quantile regression model at level $1 - \alpha$ on a training data set of size n ;
2. calibrating the scores s_1, \dots, s_{n_c} on an independent data set.

For extreme confidence levels, both steps typically require extrapolation beyond the data range. Indeed, if α is close to 0, and in particular if $\alpha < 1/(n + 1)$ is also extreme in the training data, then usual quantile regression will not be accurate. Instead, extreme quantile regression methods should be used. There is large literature on such methods based on linear models (Chernozhukov, 2005), additive models (Chavez-Demoulin and Davison, 2005; Youngman, 2019), or more flexible machine learning models such as gradient boosting (Velthoen et al., 2023), random forest (Gnecco et al., 2024) or neural networks (Pasche and Engelke, 2024; Richards and Huser, 2024). Importantly, the model in step 1. can be a black-box, in the sense that we do not require theoretical guarantees. We discuss the extrapolation for step 2. in this section and come back to examples of extreme quantile regression models in Sections 4 and 5.

3.1 Unilateral prediction intervals

Extreme conformal prediction intervals are most relevant in cases where very large values of the response variable Y lead to severe negative impacts. In such cases, reliable prediction intervals for $Y \mid \mathbf{X} = \mathbf{x}$, which contain the realisation of Y with very high marginal probability, are a crucial forecasting tool. They can be used to determine whether a dangerous level of the response could potentially be reached. This also allows the reduction of false negatives (e.g., in the form of missing warnings) that can be critical to the system. In those risk assessment scenarios, it is often a single of the two tail directions which is of risk. This is the case for the risk of flooding discussed in Section 5 but also, for instance, for high temperatures in dry areas at risk of wildfires, and financial asset returns at risk of large losses. Without loss of generality, we, thus, suppose that one is interested in single-sided prediction intervals; two-sided intervals can be constructed analogously.

The classical bilateral procedure can be adapted to unilateral prediction intervals using the score function

$$s(\mathbf{x}, y) := y - \hat{Q}_{1-\alpha}(\mathbf{x}), \tag{7}$$

instead of Eq. (5), where $\hat{Q}_{1-\alpha}(\mathbf{x})$ is the pretrained quantile regression model at level $1 - \alpha$. Let $y_{\min} \in \mathbb{R} \cup \{-\infty\}$ be the lower endpoint of the distribution of Y (or of the conditional distribution $Y \mid \mathbf{X} = \mathbf{x}$, if known). Then, following the usual procedure, the resulting interval is

$$C(\mathbf{x}) = \left(y_{\min}, \hat{Q}_{1-\alpha}(\mathbf{x}) + \hat{q}_\alpha \right], \quad (8)$$

where \hat{q}_α is, in classical conformal prediction, the $(\lceil (n_c + 1)(1 - \alpha) \rceil / n_c)$ -quantile of the calibration scores s_1, \dots, s_{n_c} , i.e., the order statistic $s_{(\lceil (n_c + 1)(1 - \alpha) \rceil)}$.

3.2 Calibrative extrapolation

As discussed in Section 2.3, when an extreme confidence level $1 - \alpha > n_c / (n_c + 1)$ is required, using order statistics to estimate \hat{q}_α would lead to degenerate intervals. Therefore, an alternative approach is needed to estimate a value \hat{q}_α^e from the calibration set such that

$$\mathbb{P}(s_{\text{test}} \leq \hat{q}_\alpha^e) \geq 1 - \alpha. \quad (9)$$

We propose to rely on the classical peaks-over-threshold methodology from extreme value theory to find such a quantile estimate. The tail of the distribution of the calibration score $S := s(\mathbf{X}, Y)$ can be approximated by the generalized Pareto distribution (GPD) above a high threshold u by

$$\mathbb{P}(S > y) = \mathbb{P}(S > u) \mathbb{P}(S > y \mid S > u) \approx \mathbb{P}(S > u) \left(1 + \xi \frac{y - u}{\sigma(u)} \right)_+^{-1/\xi}, \quad \forall y \geq u, \quad (10)$$

where $\xi \in \mathbb{R}$ and $\sigma(u) > 0$ are the shape and scale parameters and u is an intermediate threshold. Under very mild assumptions on the distribution F_S of S , this approximation is theoretically justified as u tends to the upper endpoint of F_S (Balkema and de Haan, 1974; Pickands III, 1975). In practice, u is typically chosen as the empirical $\tau_0 = (1 - k/n_c)$ -quantile \hat{Q}_{τ_0} of S for some $k < n_c$. The tuning parameter k is the number of exceedances used for estimation of the parameters σ and ξ , for instance, by maximum likelihood. Quantiles of S can then be estimated at probability levels beyond the data range using the approximation

$$\hat{Q}_{\tilde{\tau}}^{\text{GPD}} := \hat{Q}_{\tau_0} + \frac{\hat{\sigma}}{\hat{\xi}} \left[\left(\frac{1 - \tau_0}{1 - \tilde{\tau}} \right)^{\hat{\xi}} - 1 \right], \quad \forall \tilde{\tau} > \tau_0. \quad (11)$$

Theoretical guarantees of these estimators typically require that $k \rightarrow \infty$ and $k/n_c \rightarrow 0$ as $n_c \rightarrow \infty$ to ensure the correct trade-off between bias and variance; see de Haan and Ferreira (2006) for more details.

Asymptotically as $n_c \rightarrow \infty$ and under additional second-order conditions, using $\hat{q}_\alpha^e := \hat{Q}_{1-\alpha}^{\text{GPD}}$ would satisfy Eq. (9) with equality (in a suitable limiting sense). However, on a finite calibration sample and with a fixed threshold u , $\hat{Q}_{1-\alpha}^{\text{GPD}}$ can underestimate the true quantile due to estimation and approximation biases, respectively (Roodman, 2018; Bücher and Zhou, 2021; Zeder et al., 2023). We, therefore, follow here a more conservative approach. Alternatively to choosing \hat{q}_α^e as $\hat{Q}_{1-\alpha}^{\text{GPD}}$, we may use the upper endpoint of a $(1 - \alpha_2)$ -confidence interval for $F_S^{-1}(1 - \alpha_1)$, the $(1 - \alpha_1)$ -quantile of the calibration scores S , for two suitable levels $\alpha_1, \alpha_2 \in (0, 1)$. The following proposition shows that, if the upper endpoint of the confidence interval has correct coverage, then the resulting extreme conformal prediction interval satisfies Eq. (1).

Proposition 3.1. *Let $\alpha_1, \alpha_2 \in (0, 1)$, and $[Q_L, Q_U]$ be a $(1 - \alpha_2)$ -confidence interval for $F_S^{-1}(1 - \alpha_1)$, the $(1 - \alpha_1)$ -quantile of the calibration scores S . If the confidence interval has correct coverage, i.e. $\mathbb{P}(U_L \leq F_S^{-1}(1 - \alpha_1) \leq U_Q) \geq 1 - \alpha_2$, and if*

$$1 - \alpha \leq (1 - \alpha_1)(1 - \alpha_2), \quad (12)$$

then $\mathbb{P}(s_{\text{test}} \leq Q_U) \geq 1 - \alpha$.

Proof. Let $[Q_L, Q_U]$ be the $(1 - \alpha_2)$ -confidence interval for $q := F_S^{-1}(1 - \alpha_1)$. Then, by assumption,

$$\mathbb{P}(q \leq U_Q) \geq \mathbb{P}(U_L \leq q \leq U_Q) \geq 1 - \alpha_2, \text{ and } \mathbb{P}(s_{\text{test}} \leq q) \geq 1 - \alpha_1.$$

As s_{test} and Q_U are independent,

$$\begin{aligned} \mathbb{P}(s_{\text{test}} \leq Q_U) &= \mathbb{P}(s_{\text{test}} - q \leq Q_U - q) \geq \mathbb{P}(\{s_{\text{test}} \leq q\} \cap \{Q_U \geq q\}) = \\ &= \mathbb{P}(s_{\text{test}} \leq q) \mathbb{P}(Q_U \geq q) \geq (1 - \alpha_1)(1 - \alpha_2) \geq 1 - \alpha. \end{aligned}$$

□

Two natural choices for α_1 and α_2 that satisfy Eq. (12) are $\alpha_1 = \alpha_2 = \alpha/2$, which is analogous to a Bonferroni correction (Bonferroni, 1936), and $\alpha_1 = \alpha_2 = 1 - (1 - \alpha)^{1/2}$, which is analogous to the Šidák correction (Šidák, 1967). Although the Šidák correction is slightly less conservative, the difference is negligible for small values of α .

There are several well-studied approaches for obtaining extreme quantile confidence intervals (CI) using the GPD approximation (Coles, 2001; Davison and Hinkley, 1997; Davison et al., 2003; de Haan and Zhou, 2022), including the profile likelihood CI, bootstrap CI and delta method CI with normal approximation. The profile likelihood CI typically represents uncertainty best and yields the most conservative CI upper endpoint. However, with small sample sizes when α is very small, estimating the upper endpoint of the CI can be challenging since the derivative of the right side of the profile log-likelihood is close to zero. This sometimes makes finding the crosspoint between the profile curve and the chi-square confidence line numerically difficult. This may lead to numerically undefined or infinite CI upper endpoints. The bootstrap approach comes in different variations, both for the sampling step (nonparametric, parametric) and for the aggregation step (basic, percentile, normal, etc...). We here focus on the nonparametric bootstrap with percentile aggregation, as there are the most commonly used. The bootstrap can give reliable confidence intervals but might be less conservative than the profile-likelihood approach. Finally, the delta method CI is computationally less expensive than the other two alternatives, but it provides intervals that are asymmetric around the quantile estimate, which is not realistic for large quantiles. Moreover, it can also suffer from numerical instability issues, due to its matrix inversion step, and fail to yield meaningful CIs.

4 Simulation study

4.1 Experimental setup

To assess the different conformalization methods, we perform a simulation study with several scenarios. The data is generated from

$$\begin{cases} X \sim \mathcal{U}([-1, 1]^{10}), \\ Y | X = \mathbf{x} \sim \sigma(\mathbf{x}) \cdot \varepsilon_Y, \end{cases} \quad (13)$$

with $\sigma(\mathbf{x}) := 1 + 6\phi(x_1, x_2)$, where ϕ is the bivariate Gaussian density with correlation 0.9. We consider two scenarios for the noise variable ε_Y : a heavy-tailed Student t distribution $t_{\alpha(\mathbf{x})}$, with covariate-dependent tail index $\alpha(\mathbf{x}) = 1/\xi(\mathbf{x}) := 7 \cdot \{1 + \exp(4x_1 + 1.2)\}^{-1} + 3$, and a light-tailed Gaussian $\mathcal{N}(0, 1)$ distribution. The former choice corresponds exactly to the generating process used in two extreme quantile regression benchmark studies (Velthoen et al., 2023; Pasche and Engelke, 2024).

We consider several sizes for the calibration sets, with $n_c \in \{10^3, 10^{3.5}, 10^4\}$ observations. For each calibration size, we repeat the experiments 100 times. We consider extreme PI confidence levels $1 - \alpha$, with $\alpha \in \{10^{-3}, 10^{-3.5}, 10^{-4}, 10^{-4.5}, 10^{-5}\}$. We consider two choices for the initial

“pretrained” quantile predictions: the conditional-quantile ground truth and a pretrained extreme quantile regression model. The former aims at assessing the methods with ideal initial predictions. As all conformalization methods are translation invariant, adding first-order bias to the pretrained model would always lead to the same final PIs. For the latter choice, we use the extreme quantile regression neural networks (EQRN) model, as it performed best on this benchmark dataset (Pasche and Engelke, 2024). It is pretrained on 5,000 observations generated from Eq. (13). Its hyperparameters and architecture were selected based on validation GPD deviance with a grid search. For each calibration size, repetition, confidence level, and initial predictions, we perform the following conformalization procedures and assess their average population coverage on a separate test set of 10^6 observations.

- GPD profile: \hat{q}_α^e is the endpoint of the profile-likelihood CI for the calibration-score quantile.
- GPD bootstrap: \hat{q}_α^e is the endpoint of the nonparametric bootstrap percentile CI for the calibration-score quantile.
- GPD delta: \hat{q}_α^e is the endpoint of the delta method CI for the calibration-score quantile.
- GPD simple: \hat{q}_α^e is the GPD $(1 - \alpha)$ -quantile estimate of the calibration scores.
- Classical: Classical (split, single-sided) conformalized quantile regression based on the calibration-score order statistics.

Each of the GPD-based procedures use the empirical 0.95-quantile as the threshold u .

4.2 Coverage results

Figure 1 shows the distribution of the computed test coverage for each considered conformalization method, confidence level, and calibration size, for the Student t noise and ground truth original predictions. The chosen confidence levels are particularly large relative to the size of the calibration sets. Hence, for most scenarios, $\alpha < 1/(n_c + 1)$. In those cases, the classical conformalization method always yields trivial infinite intervals. They have a coverage of 1, but are uninformative and of no practical use. On the other hand, the other methods, relying on peaks over threshold extrapolation instead of empirical quantiles, are able to yield finite PIs, even when $\alpha \ll 1/(n_c + 1)$.

The simple GPD estimates of the calibration-score $(1 - \alpha)$ -quantile do not seem to provide sufficient coverage with small calibration samples and for the larger confidence levels, likely due to the GPD estimation error or approximation biases. The other three methods, relying on the confidence intervals for the score quantiles (and Proposition 3.1), achieve much better coverage and, in most cases, satisfy Eq. (1) as their coverage is larger than $1 - \alpha$ on average. However, those GPD CI-based methods seem consistently overconservative for lower confidence levels.

In general, the profile likelihood method seems the most conservative, compared to the nonparametric bootstrap and delta method alternatives, as anticipated. It also satisfies the coverage guarantee in all scenarios. Its downside is the numerical difficulty, described in Section 3.2. With the implementation at hand, this issue arose, in the worst case, in 85% of the repetitions for $n_c = 1000$ and the lowest α value, but quickly decreased to, at most, 2% for $n_c = 3163$ and 0% for $n_c = 10000$. This instability is understandable in the former truly extreme case, as the confidence level is more than two orders of magnitude larger than the level for which PIs are obtainable with the classical conformal method and as the likelihood only relies on 50 observations to estimate a $(1 - 5 \cdot 10^{-6})$ -confidence interval for a $(1 - 5 \cdot 10^{-6})$ -quantile.

The bootstrap and delta-method approaches seem less overconservative for the more moderate α values, but slightly undercover in the scenarios with the lowest α values. Nevertheless, they still significantly outperform the simple GPD approach and the infinite classical PIs. Contrary to the profile likelihood approach, the bootstrap method never fails to provide finite estimates. On the

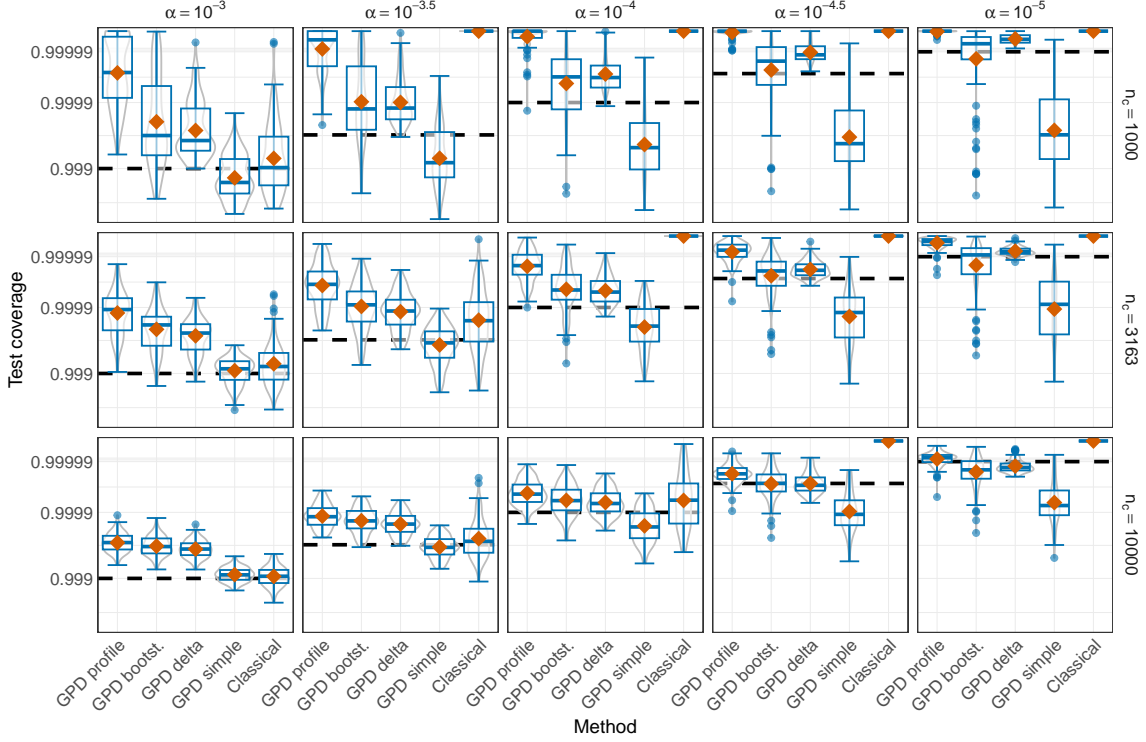


Figure 1: Boxplots of the test coverage probability of the quantile PIs (analytically computed for a grid of test observation and averaged over \mathcal{X}) for different conformalization methods, conformal confidence values $1 - \alpha$ (columns, labelled with α), and calibration sample sizes n_c (rows), for the Student t distributed noise and quantile ground-truth predictions.

other hand, the delta-method approach also suffers from stability issues with small calibration sizes, regardless of the confidence level.

Figure A.1 in Appendix A.1 shows the same coverage distribution when the data-generating process has light-tailed Gaussian noise, instead of heavy-tailed Student t_4 noise. In comparison, all methods tend to result in significantly more conservative intervals in terms of the coverage. In particular, all three GPD CI-based approaches always result in more coverage than necessary.

Figure 2 shows the coverage distributions for the EQRN predictions and Student t_4 noise. We observe that, although being accurate predictions of the conditional quantile, in terms of integrated squared error (Pasche and Engelke, 2024), the EQRN predictions, in this case, undercover when considered as a PI endpoint. The conformalization results closely match those for the ground-truth quantile predictions, although the coverage seems smaller for the largest confidence levels, for all methods. The scenario with $\alpha = 10^{-5}$ and the largest sample size is the only one for which the GPD profile approach seems to slightly undercover. All the other non-infinite alternatives also undercover for this largest confidence level. The extreme conformal prediction methods all outperform the original EQRN prediction in all scenarios, except for the GPD simple on small calibration sizes.

As a takeaway, our practical recommendation for conservative and informative high-confidence PIs is to use the profile-likelihood method if its estimate is finite and the bootstrap-based estimate otherwise. This combination results in the profile-likelihood conservativeness, in most scenarios, and avoids the potential infinite intervals with the bootstrap estimate, which is still conservative enough in the majority of scenarios, in cases of numerical difficulties. We call this method the “GPD safeprofile” PI. Alternatively, considering the maximum of the bootstrap and delta-method PI endpoints as a replacement for unstable profile likelihood situations could be more conservative but might be less stable.

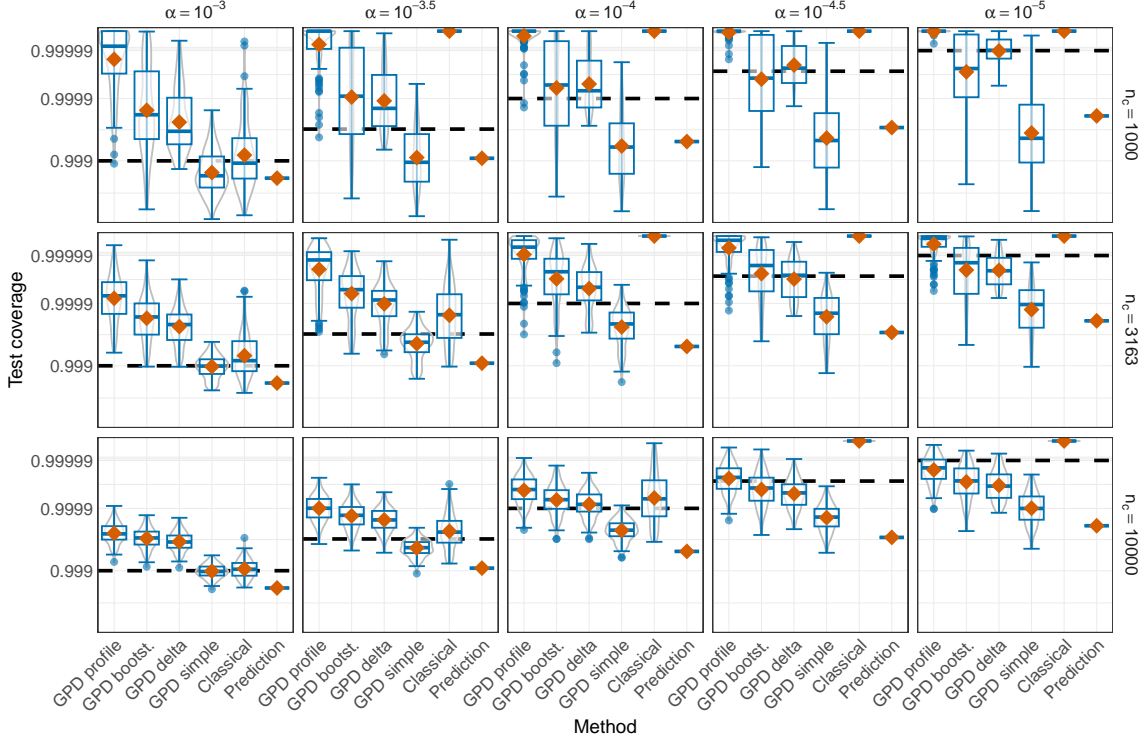


Figure 2: Boxplots of the test coverage probability of the quantile PIs (analytically computed for a grid of test observation and averaged over \mathcal{X}) for different conformalization methods, conformal confidence values $1 - \alpha$ (columns, labelled with α), and calibration sample sizes n_c (rows), for the Student t distributed noise and EQRN predictions.

5 Application to flood risk forecasting

5.1 Description and aim

Flooding is one of the most impactful natural hazards in terms of infrastructure and economic damage, and of the endangerment of human lives. Methods from extreme value theory have proven successful for assessing flood risk and providing reliable worst-case scenarios (e.g., [Katz et al., 2002](#); [Keef et al., 2009](#); [Asadi et al., 2018](#); [Engelke and Hitz, 2020](#); [Engelke and Ivanovs, 2021](#)).

In Switzerland, the Federal Office for the Environment (FOEN) monitors the river flow with numerous gauging stations throughout the river network. In its capital city, Bern, extreme water flow events of the Aare river led to several major floods, causing some of the most severe infrastructural and economic flooding damages recorded in the country. The main driver of strong water-flow events is the cumulative amount of upstream precipitation. In this study, we rely on the average daily discharge measures of the Aare river (in m^3s^{-1}) provided by the FOEN¹, and on recordings of daily precipitation (in mm) at various meteorological stations, obtained from MeteoSwiss². This version of the dataset was preprocessed and analysed in previous studies ([Pasche et al., 2023](#); [Pasche and Engelke, 2024](#)).

With our proposed extreme conformal approach, we aim to provide high-confidence one-day-ahead interval forecasts of the conditional range for water flow. We rely on several extreme quantile regression models pretrained to forecast the one-day-ahead extreme quantiles of the Aare water flow in Bern, given observations of upstream precipitation at six locations in Bern’s water catchment and of the average daily water flow at an upstream gauging station, during the previous 10 days.

¹<https://www.hydrodaten.admin.ch/>.

²<https://gate.meteoswiss.ch/idaweb>.

Figure A.2 in Appendix A.1 shows the location of those meteorological and gauging stations. We then use the GPD safe-profile extreme conformalization approach recommended in Section 4 to obtain the desired one-day-ahead PIs. The extreme quantile regression models we consider are EQRN, GBEX (Velthoen et al., 2023), EGAM (Youngman, 2019), and EXQAR (Li and Wang, 2019). We also consider the simple unconditional GPD quantile estimates as a comparative baseline.

The models were pretrained on data from 1939 to 1951. They were all fine-tuned with a grid search for hyperparameter selection. We use data from 1958 to 1999 for calibration and testing (48 years). The observations after 1999 are not considered, due to a major distribution shift³. We choose 10 years as the default calibration size in the first part of the analysis, but vary this size from 3 to 15 years in the second part. We use multiples of complete years to keep a seasonal balance in the calibration and test sets.

5.2 Results

Figure 3 compares the number of observations exceeding the PIs from each method during the test period, using predictions of the different pretrained models, for a range of moderate to extreme confidence levels. The number of observations expected to exceed the PIs during the 38-year test period varies from 2,776, for $1 - \alpha = 0.8$, to only 1.4, for the largest level $1 - \alpha = 0.9999$. Using the original model predictions as PIs leads, in most cases, to undercoverage. Although the best-performing predictions seem to vary around the target coverage, they fail to provide satisfactory coverage consistently. The classical conformalization seems effective for the lowest two confidence levels but worsens the coverage for the following moderately extreme levels, compared to the initial predictions. At the largest level, going beyond the calibration order statistics, the classical method yields infinite-endpoint PIs. The extreme conformalization method yields finite PIs with significantly better coverage for confidence levels above 0.95, for which the approaches differ. The PI coverage consistently strictly satisfies the target confidence levels for each initial prediction model. The profile likelihood procedure is stable for almost all models and confidence levels. It was replaced by the bootstrap PIs only once, with the EGAM predictions.

Figure 4 shows the initial EQRN predictions, which seem to fit the data variation best, and their conformalized PI endpoints for two of the considered confidence levels, including the largest. The infinite classical PIs are non-informative for the largest level. On the other hand, the extreme PIs, providing correct test coverage, are not overly large compared to the data variation, the original predictions, and the unconditional quantile estimates. All other extreme conformal corrections for the EQRN predictions are even smaller in magnitude. During the considered year, the original predictions at the largest confidence level are exceeded once, on 25th July 1973. This exceedance is covered by the conformalized extreme PI.

Figure 5 shows the evolution of the test coverage with the calibration size, for all predictions, methods, and confidence levels. We observe that the extreme PIs significantly outperform the classical conformalization in terms of empirical test coverage, for all relevant levels. It always provides informative finite PIs, contrary to the classical approach that yields infinite PIs when $\alpha < 1/(n_c + 1)$, that is, for calibration sizes up to 8 years with $1 - \alpha = 0.9997$, and for all sizes with $1 - \alpha = 0.9999$. The extreme PI has valid coverage in almost all combinations, including when the classical approach and/or original predictions significantly undercover. In the few undercovered situations, it still outperforms both the original predictions and the classical approach.

³See the flood report of the FOEN at <https://www.hydrodaten.admin.ch/en/2135.html>.

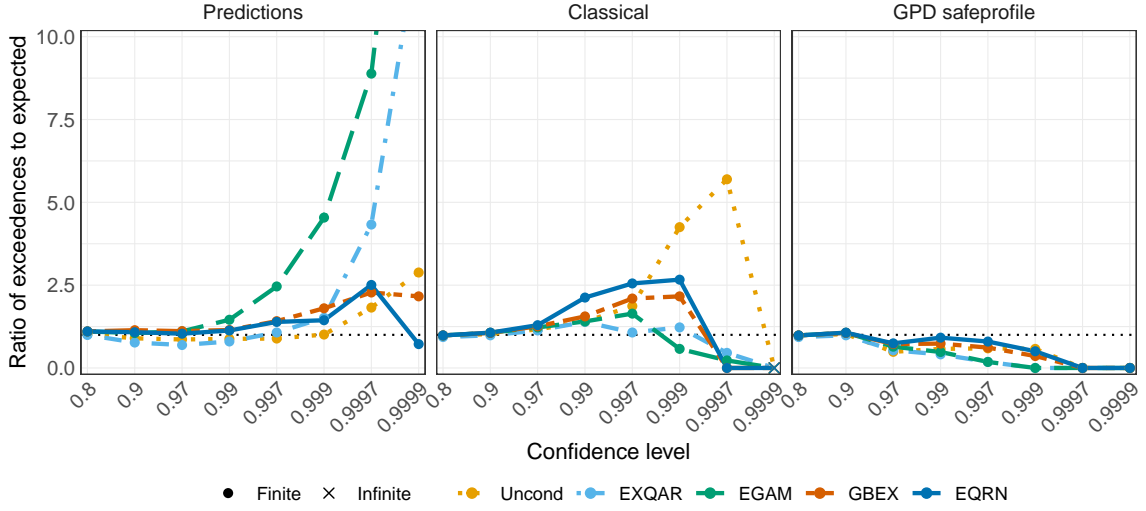


Figure 3: Number of observations exceeding the PIs during the test period as a ratio to the expected number of exceedances for different confidence levels and each pretrained prediction model, for the original predictions (left), the classical conformalization (center), and the GPD safeprofile method (right).

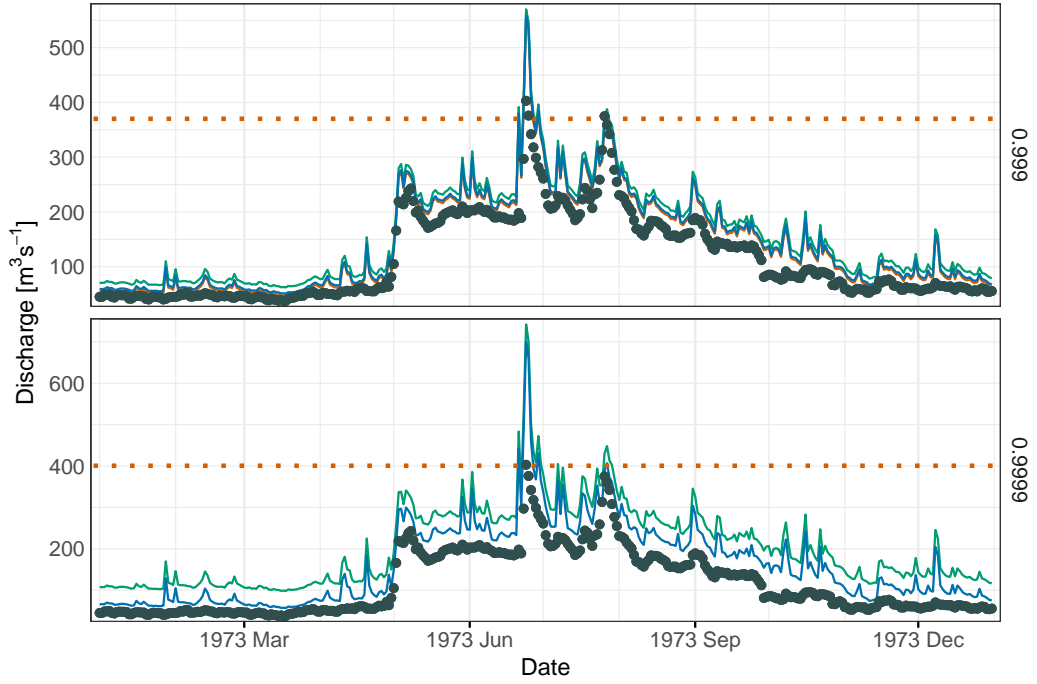


Figure 4: Original EQRN prediction (blue), classical conformal PI (red), and extreme conformal PI (green), at confidence levels 0.999 (top panel) and 0.9999 (bottom panel), during one of the test years. The classical conformal PI is infinite at level 0.9999. The observations (points) and the unconditional GPD $(1 - \alpha)$ -quantile estimates (dotted lines) are also shown.

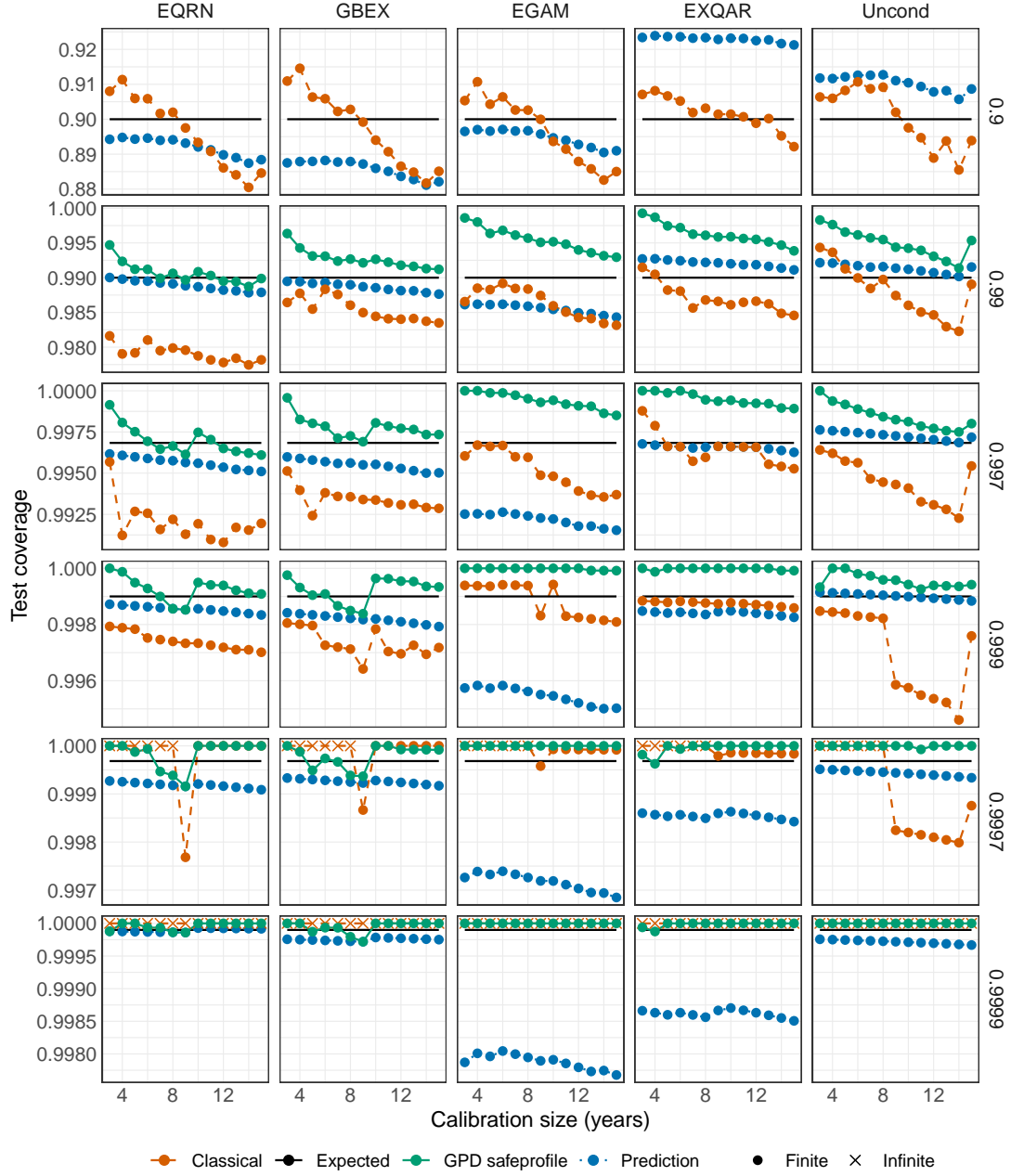


Figure 5: Empirical test coverage of the prediction intervals for a range of calibration-set sizes, for each conformalization method, pretrained model (columns), and confidence level $1 - \alpha$ (rows). For $1 - \alpha < 0.95$, corresponding to the quantile level of the GPD threshold u , the GPD approach coincides with the classical.

6 Conclusion

We propose a conformalization method which relies on extreme value statistics to provide conservative and non-degenerate prediction intervals for reliable risk assessment of high-impact events. The novel method uses the well-studied peaks-over-threshold approach, which leverages the generalized Pareto distribution to extrapolate the necessary conformal correction on the calibration data to the required extreme confidence levels. It uses a conservative confidence interval solution for robustness against estimation and approximation biases.

In the simulation study and the application to river-flow forecasts, our extreme conformal prediction method consistently provides better coverage than both the original predictions and the classical conformalization method at large confidence levels. For the largest levels, the classical approach results in infinite (or undefined) PIs, which are of no practical use. On the other hand, our recommended method yields informative finite intervals that consistently achieve the desired coverage for confidence levels up to several orders of magnitude larger than what is feasible with classical methods. Importantly, our extreme conformal prediction approach can be used in combination with any extreme quantile regression model, including black-box machine learning methods without known asymptotic guarantees.

One downside of our approach is its potential overconservativeness in certain scenarios, e.g. for moderately extreme confidence levels and some lighter-tailed data distributions. This is, at least in part, a consequence of our conservative CI-based solution, used to circumvent possible estimation and/or approximation biases in the peaks-over-threshold GPD estimation. Contrary to the finite sample properties of classical conformal prediction, the GPD relies on asymptotic results and lacks those finite sample guarantees. This is an unavoidable trade-off for the ability to extrapolate beyond the moderate levels for which the classical empirical-quantile-based method is feasible. In the general framing of this work, strict finite-sample guarantees for quantiles extrapolated beyond the data range are not theoretically possible without additional assumptions, due to the broad spectrum of potential sub-asymptotic tail behaviours (Thomas, 2015; Lhaut et al., 2022).

This work establishes a first method for extreme-confidence PIs in the context of the well-established split-conformalized regression. Further work is possible to include more specialised or alternative conformal approaches. One direction is to enhance the efficiency of data usage, using, for example, full conformal prediction or potential variants of jackknife (Barber et al., 2021; Steinberger and Leeb, 2016), infinitesimal jackknife (Alaa and van der Schaar, 2020), or cross-validation (Vovk, 2015). However, such approaches typically have a much greater computational cost, as they require refitting the predictive model multiple times. Other directions include extensions to non-stationary data (see e.g. Barber et al., 2023). Finally, approaches other than conformalization could also warrant investigations, such as building PIs based on the so-called high-quality criterion and using deep learning (Pearce et al., 2018; Khosravi et al., 2011). How to incorporate extreme value statistics to extrapolate prediction intervals to cover high-impact events in these methodologies appears to be largely open.

Declarations

Acknowledgements

This research project was conducted while the first author, O. C. Pasche, was a visiting scholar at the Department of Industrial Engineering and Operations Research, at Columbia University. He thanks the department and the university for their hospitality during this period.

Funding

O. C. Pasche and S. Engelke were supported by the Swiss National Science Foundation Eccellenza Grant 186858. H. Lam was supported by the InnoHK initiative of the Innovation and Technology Commission of the Hong Kong Special Administrative Region Government, Laboratory for AI-Powered Financial Technologies, and the Columbia Innovation Hub Award.

Availability of supporting data

In the Application to flood risk forecasting, we use river discharge and precipitation data recorded in Switzerland between 1930 and 1999, in the Rhine and Aare basins. The precipitation records can be ordered online for free from MeteoSwiss, on <https://gate.meteoswiss.ch/idaweb>, and the discharge records from the Swiss Federal Office for the Environment (FOEN), on <https://www.bafu.admin.ch/bafu/en/home/topics/water/state/data/obtaining-monitoring-data-on-the-topic-of-water/hydrological-data-service-for-watercourses-and-lakes.html>.

APPENDIX

A.1 Additional figures

A.1.1 Simulation study

Figure A.1 shows the distribution of the computed test coverage for each considered conformalization method, confidence level, and calibration size, for the light-tailed Gaussian-noise data and ground truth original predictions.

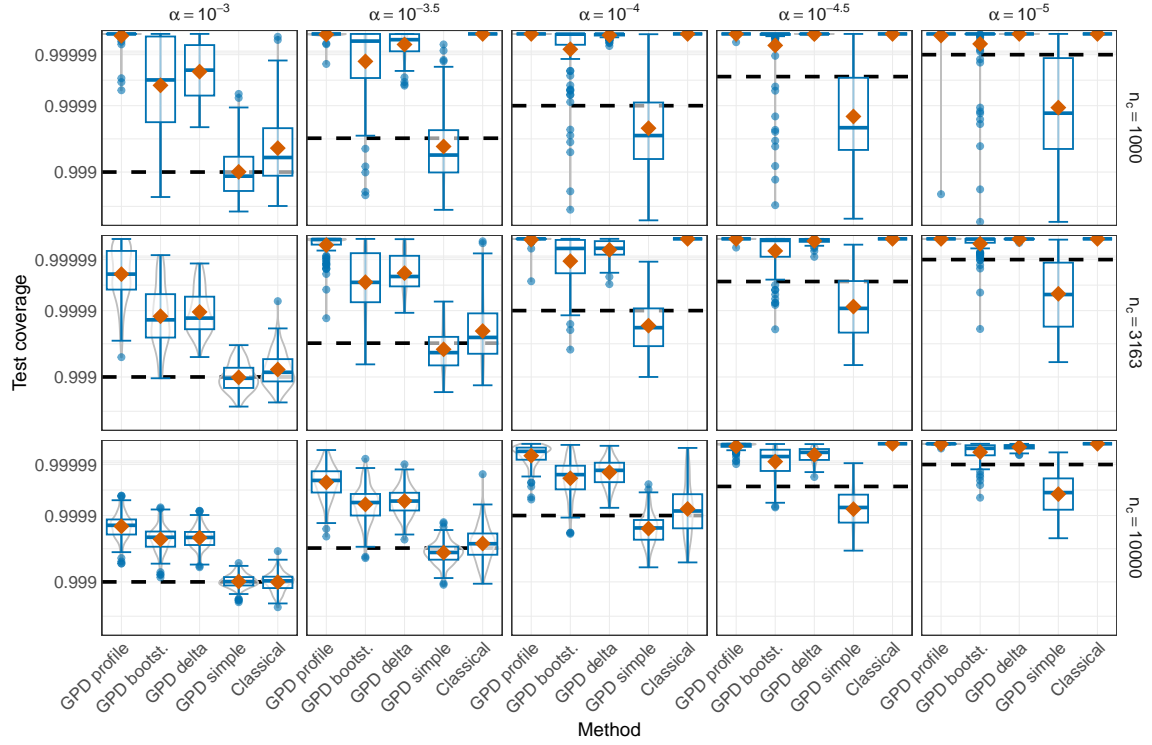


Figure A.1: Boxplots of the test coverage probability of the quantile PIs (analytically computed for a grid of test observation and averaged over \mathcal{X}) for different conformalization methods, conformal confidence values $1 - \alpha$ (columns, labelled with α), and calibration sample sizes n_c (rows), for the Gaussian distributed noise and quantile ground-truth predictions.

A.1.2 Application to river-flow forecasts

Figure A.2 shows the locations of the meteorological and gauging stations corresponding to the variables used in the model forecasts.

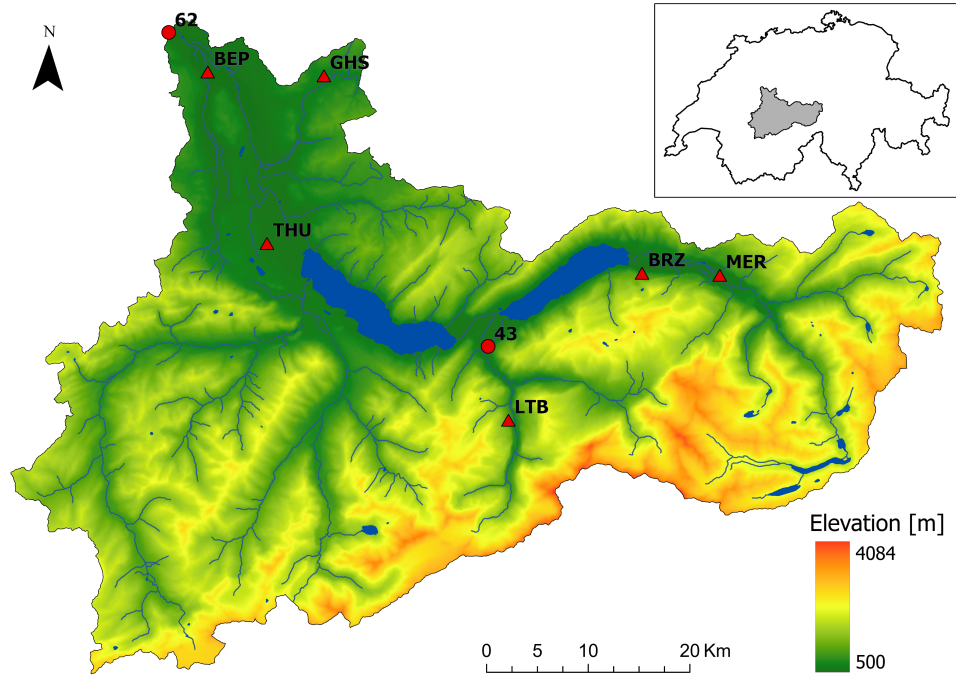


Figure A.2: Topographic map of water catchment of the gauging station in Bern–Schönau (62) on the Aare in Switzerland. Another upstream gauging station in Gsteig (43), on the Lütschine river, and six meteorological stations with precipitation measurements (triangles) are also shown (source: [Pasche and Engelke, 2024](#)).

References

- Alaa, A. and van der Schaar, M. Discriminative jackknife: Quantifying uncertainty in deep learning via higher-order influence functions. In *Proc. 37th Int. Conf. Mach. Learn.*, volume 119, pages 165–174. PMLR, 2020.
- Angelopoulos, A. N. and Bates, S. Conformal Prediction: A Gentle Introduction. *Found. Trends in Mach. Learn.*, 16(4): 494–591, 2023. doi:[10.1561/2200000101](https://doi.org/10.1561/2200000101).
- Asadi, P., Davison, A. C., and Engelke, S. Extremes on river networks. *Ann. Appl. Stat.*, 9(4):2023–2050, 2015. doi:[10.1214/15-aos863](https://doi.org/10.1214/15-aos863).
- Asadi, P., Engelke, S., and Davison, A. C. Optimal regionalization of extreme value distributions for flood estimation. *J. Hydrol.*, 556:182–193, 2018. doi:[10.1016/j.jhydrol.2017.10.051](https://doi.org/10.1016/j.jhydrol.2017.10.051).
- Balkema, A. A. and de Haan, L. Residual Life Time at Great Age. *Ann. Probab.*, 2(5):792 – 804, 1974. doi:[10.1214/aop/1176996548](https://doi.org/10.1214/aop/1176996548).
- Barber, R. F., Candès, E. J., Ramdas, A., and Tibshirani, R. J. Predictive inference with the jackknife+. *Ann. Stat.*, 49(1): 486–507, 2021. doi:[10.1214/20-AOS1965](https://doi.org/10.1214/20-AOS1965).
- Barber, R. F., Candès, E. J., Ramdas, A., and Tibshirani, R. J. Conformal prediction beyond exchangeability. *Ann. Stat.*, 51(2):816–845, 2023. doi:[10.1214/23-AOS2276](https://doi.org/10.1214/23-AOS2276).
- Bonferroni, C. Teoria statistica delle classi e calcolo delle probabilita. *Pubblicazioni del R istituto superiore di scienze economiche e commerciali di firenze*, 8:3–62, 1936.
- Boulaoui, Y., Zscheischler, J., Vignotto, E., van der Wiel, K., and Engelke, S. Modeling and simulating spatial extremes by combining extreme value theory with generative adversarial networks. *Environ. Data Sci.*, 1:e5, 2022. doi:[10.1017/eds.2022.4](https://doi.org/10.1017/eds.2022.4).
- Bücher, A. and Zhou, C. A Horse Race between the Block Maxima Method and the Peak-over-Threshold Approach. *Stat. Sci.*, 36(3):360–378, 2021. doi:[10.1214/20-STS795](https://doi.org/10.1214/20-STS795).
- Buriticá, G. and Engelke, S. Progression: an extrapolation principle for regression. *ArXiv preprint*, 2024. doi:[10.48550/arXiv.2410.23246](https://doi.org/10.48550/arXiv.2410.23246).
- Chavez-Demoulin, V. and Davison, A. C. Generalized additive modelling of sample extremes. *J. R. Stat. Soc. C*, 54(1): 207–222, 2005. doi:[10.1111/j.1467-9876.2005.00479.x](https://doi.org/10.1111/j.1467-9876.2005.00479.x).
- Chernozhukov, V. Extremal quantile regression. *Ann. Stat.*, 33(2):806 – 839, 2005. doi:[10.1214/009053604000001165](https://doi.org/10.1214/009053604000001165).
- Coles, S. *An Introduction to Statistical Modeling of Extreme Values*. Springer, 2001. doi:[10.1007/978-1-4471-3675-0](https://doi.org/10.1007/978-1-4471-3675-0).
- Davison, A. C. and Hinkley, D. V. *Bootstrap Methods and their Application*. Cambridge University Press, New York, 1997. doi:[10.1017/CBO9780511802843](https://doi.org/10.1017/CBO9780511802843).
- Davison, A. C., Hinkley, D. V., and Young, G. A. Recent Developments in Bootstrap Methodology. *Stat. Sci.*, 18(2): 141–157, 2003. doi:[10.1214/ss/1063994969](https://doi.org/10.1214/ss/1063994969).
- de Carvalho, M., Kumukova, A., and Dos Reis, G. Regression-type analysis for multivariate extreme values. *Extremes*, 25(4):595–622, 2022. doi:[10.1007/s10687-022-00446-6](https://doi.org/10.1007/s10687-022-00446-6).
- de Haan, L. and Ferreira, A. *Extreme Value Theory*. Springer, 2006. doi:[10.1007/0-387-34471-3](https://doi.org/10.1007/0-387-34471-3).
- de Haan, L. and Zhou, C. Bootstrapping extreme value estimators. *J. Am. Stat. Assoc.*, 119(545):382–393, 2022. doi:[10.1080/01621459.2022.2120400](https://doi.org/10.1080/01621459.2022.2120400).
- Dupuis, D. J., Engelke, S., and Trapin, L. Modeling panels of extremes. *Ann. Appl. Stat.*, 17(1):498–517, 2023. doi:[10.1214/22-AOAS1639](https://doi.org/10.1214/22-AOAS1639).
- Engelke, S. and Hitz, A. Graphical models for extremes (with discussion). *J. R. Stat. Soc. B*, 82(4):871–932, 2020. doi:[10.1111/rssb.12355](https://doi.org/10.1111/rssb.12355).
- Engelke, S. and Ivanovs, J. Sparse structures for multivariate extremes. *Annu. Rev. Stat. Appl.*, 8:241–270, 2021. doi:[10.1146/annurev-statistics-040620-041554](https://doi.org/10.1146/annurev-statistics-040620-041554).
- Gnecco, N., Terefe, E. M., and Engelke, S. Extremal random forests. *J. Am. Stat. Assoc.*, 119(548):3059–3072, 2024. doi:[10.1080/01621459.2023.2300522](https://doi.org/10.1080/01621459.2023.2300522).
- Gupta, C., Kuchibhotla, A. K., and Ramdas, A. Nested conformal prediction and quantile out-of-bag ensemble methods. *Pattern Recognit.*, 127:108496, 2022. doi:[10.1016/j.patcog.2021.108496](https://doi.org/10.1016/j.patcog.2021.108496).
- Huet, N., Cléménçon, S., and Sabourin, A. On regression in extreme regions. *ArXiv preprint*, 2024. doi:[10.48550/arXiv.2303.03084](https://doi.org/10.48550/arXiv.2303.03084).
- Jalalzai, H., Cléménçon, S., and Sabourin, A. On binary classification in extreme regions. In *Adv. Neural Inf. Process. Syst.*, volume 31, pages 3092–3100, 2018.

- Katz, R. W., Parlange, M. B., and Naveau, P. Statistics of extremes in hydrology. *Adv. Water Resour.*, 25:1287–1304, 2002. doi:[10.1016/S0309-1708\(02\)00056-8](https://doi.org/10.1016/S0309-1708(02)00056-8).
- Keef, C., Tawn, J., and Svensson, C. Spatial risk assessment for extreme river flows. *J. R. Stat. Soc. C*, 58(5):601–618, 2009. doi:[10.1111/j.1467-9876.2009.00672.x](https://doi.org/10.1111/j.1467-9876.2009.00672.x).
- Khosravi, A., Nahavandi, S., Creighton, D., and Atiya, A. F. Comprehensive review of neural network-based prediction intervals and new advances. *IEEE Trans. Neural Netw.*, 22(9):1341–1356, 2011. doi:[10.1109/TNN.2011.2162110](https://doi.org/10.1109/TNN.2011.2162110).
- Kim, B., Xu, C., and Barber, R. Predictive inference is free with the jackknife+–after-bootstrap. In *Adv. Neural Inf. Process. Syst.*, volume 33, pages 4138–4149, 2020.
- Lei, J., G’Sell, M., Rinaldo, A., Tibshirani, R. J., and Wasserman, L. Distribution-free predictive inference for regression. *J. Am. Stat. Assoc.*, 113(523):1094–1111, 2018. doi:[10.1080/01621459.2017.1307116](https://doi.org/10.1080/01621459.2017.1307116).
- Lhaut, S., Sabourin, A., and Segers, J. Uniform concentration bounds for frequencies of rare events. *Statist. Probab. Lett.*, 189:109610, 2022. doi:[10.1016/j.spl.2022.109610](https://doi.org/10.1016/j.spl.2022.109610).
- Li, D. and Wang, H. J. Extreme Quantile Estimation for Autoregressive Models. *J. Bus. Econ. Stat.*, 37(4):661–670, 2019. doi:[10.1080/07350015.2017.1408469](https://doi.org/10.1080/07350015.2017.1408469).
- Papadopoulos, H., Proedrou, K., Vovk, V., and Gammerman, A. Inductive confidence machines for regression. In Elomaa, T., Mannila, H., and Toivonen, H., editors, *Mach. Learn.: ECML 2002*, pages 345–356. Springer, 2002. doi:[10.1007/3-540-36755-1_29](https://doi.org/10.1007/3-540-36755-1_29).
- Pasche, O. C. and Engelke, S. Neural networks for extreme quantile regression with an application to forecasting of flood risk. *Ann. Appl. Stat.*, 18(4):2818–2839, 2024. doi:[10.1214/24-AOAS1907](https://doi.org/10.1214/24-AOAS1907).
- Pasche, O. C., Chavez-Demoulin, V., and Davison, A. C. Causal modelling of heavy-tailed variables and confounders with application to river flow. *Extremes*, 26:573–594, 2023. doi:[10.1007/s10687-022-00456-4](https://doi.org/10.1007/s10687-022-00456-4).
- Pearce, T., Brintrup, A., Zaki, M., and Neely, A. High-quality prediction intervals for deep learning: A distribution-free, ensembled approach. In *Proc. 35th Int. Conf. Mach. Learn.*, volume 80, pages 4075–4084. PMLR, 2018.
- Pickands III, J. Statistical Inference Using Extreme Order Statistics. *Ann. Stat.*, 3(1):119 – 131, 1975. doi:[10.1214/aos/1176343003](https://doi.org/10.1214/aos/1176343003).
- Richards, J. and Huser, R. Regression modelling of spatiotemporal extreme U.S. wildfires via partially-interpretable neural networks. *ArXiv preprint*, 2024. doi:[10.48550/arXiv.2208.07581](https://doi.org/10.48550/arXiv.2208.07581).
- Romano, Y., Patterson, E., and Candes, E. Conformalized quantile regression. In *Adv. Neural Inf. Process. Syst.*, volume 32, 2019.
- Roodman, D. Bias and size corrections in extreme value modeling. *Comm. Statist. Theory Methods*, 47(14):3377–3391, 2018. doi:[10.1080/03610926.2017.1353630](https://doi.org/10.1080/03610926.2017.1353630).
- Steinberger, L. and Leeb, H. Leave-one-out prediction intervals in linear regression models with many variables. *ArXiv preprint*, 2016. doi:[10.48550/arXiv.1602.05801](https://doi.org/10.48550/arXiv.1602.05801).
- Thomas, M. *Concentration results on extreme value theory*. PhD thesis, Univeristé Paris Diderot Paris 7, 2015. URL <https://theses.hal.science/tel-01177197>.
- van Oordt, M. and Zhou, C. Systemic risk and bank business models. *J. Appl. Econometrics*, 34(3):365–384, 2019. doi:[10.1002/jae.2666](https://doi.org/10.1002/jae.2666).
- Velthoen, J., Cai, J.-J., Jongbloed, G., and Schmeits, M. Improving precipitation forecasts using extreme quantile regression. *Extremes*, 22(4):599–622, 2019. doi:[10.1007/s10687-019-00355-1](https://doi.org/10.1007/s10687-019-00355-1).
- Velthoen, J., Dombry, C., Cai, J.-J., and Engelke, S. Gradient boosting for extreme quantile regression. *Extremes*, 26: 639–667, 2023. doi:[10.1007/s10687-023-00473-x](https://doi.org/10.1007/s10687-023-00473-x).
- Vovk, V. Cross-conformal predictors. *Ann. Math. Artif. Intell.*, 74:9–28, 2015. doi:[10.1007/s10472-013-9368-4](https://doi.org/10.1007/s10472-013-9368-4).
- Vovk, V., Gammerman, A., and Shafer, G. *Algorithmic Learning in a Random World*, volume 29. Springer, 2nd edition, 2022. doi:[10.1007/978-3-031-06649-8](https://doi.org/10.1007/978-3-031-06649-8).
- Youngman, B. D. Generalized Additive Models for Exceedances of High Thresholds With an Application to Return Level Estimation for U.S. Wind Gusts. *J. Am. Stat. Assoc.*, 114(528):1865–1879, 2019. doi:[10.1080/01621459.2018.1529596](https://doi.org/10.1080/01621459.2018.1529596).
- Zeder, J., Sippel, S., Pasche, O. C., Engelke, S., and Fischer, E. M. The Effect of a Short Observational Record on the Statistics of Temperature Extremes. *Geophys. Res. Lett.*, 50(16), 2023. doi:[10.1029/2023GL104090](https://doi.org/10.1029/2023GL104090).
- Šidák, Z. Rectangular confidence regions for the means of multivariate normal distributions. *J. Am. Stat. Assoc.*, 62(318): 626–633, 1967. doi:[10.1080/01621459.1967.10482935](https://doi.org/10.1080/01621459.1967.10482935).