Automated threshold selection and associated inference uncertainty for univariate extremes

Conor Murphy*, Jonathan A. Tawn
Department of Mathematics and Statistics, Lancaster University
and
Zak Varty
Department of Mathematics, Imperial College London

October 10, 2024

Abstract

Threshold selection is a fundamental problem in any threshold-based extreme value analysis. While models are asymptotically motivated, selecting an appropriate threshold for finite samples is difficult and highly subjective through standard methods. Inference for high quantiles can also be highly sensitive to the choice of threshold. Too low a threshold choice leads to bias in the fit of the extreme value model, while too high a choice leads to unnecessary additional uncertainty in the estimation of model parameters. We develop a novel methodology for automated threshold selection that directly tackles this bias-variance trade-off. We also develop a method to account for the uncertainty in the threshold estimation and propagate this uncertainty through to high quantile inference. Through a simulation study, we demonstrate the effectiveness of our method for threshold selection and subsequent extreme quantile estimation, relative to the leading existing methods, and show how the method's effectiveness is not sensitive to the tuning parameters. We apply our method to the well-known, troublesome example of the River Nidd dataset.

Keywords: extreme values, generalised Pareto distribution, river flows, return level, threshold selection, uncertainty quantification.

*This paper is based on work completed while Conor Murphy was part of the EPSRC funded STOR-i centre for doctoral training (EP/S022252/1), with part-funding from Shell Research Ltd.

1 Introduction

An inherent challenge in risk modelling is the estimation of high quantiles, known as *return levels*, beyond observed values. Such inference is important for designing policies or protections against future extreme events, e.g., in finance or hydrology (Smith, 2003; Coles et al., 2003). Extreme value methods achieve this extrapolation by using asymptotically exact models to approximate the tail of a distribution above a high, within-sample, threshold u. The choice of this threshold is fundamental in providing meaningful inference. Here, we develop novel methods for automatic selection of the threshold and for propagating the uncertainty in this selection into return level inferences.

Throughout, we assume that all data are realisations of an independent and identically-distributed (iid) univariate continuous random variable X with unknown distribution function F, with upper endpoint $x^F := \sup\{x : F(x) < 1\}$. Under weak conditions, Pickands (1975) shows that for $X > u$, with $u < x^F$, the distribution of the rescaled excess $Y = X - u$, converges to the generalised Pareto distribution (GPD) as $u \to x^F$. To use this limit result in practice, a within-sample threshold u is chosen, above which this limit result is treated as exact. Specifically, whatever the form of F, the excesses Y of u are modelled by the single flexible GPD(σ_u, ξ) family, with distribution function

$$H(y; \sigma_u, \xi) = 1 - (1 + \xi y/\sigma_u)_+^{-1/\xi}, \tag{1}$$

with $y > 0$, $w_+ = \max(w, 0)$, $(\sigma_u, \xi) \in \mathbb{R}_+ \times \mathbb{R}$ being scale and shape parameters. The exponential distribution arises when $\xi = 0$, i.e., as $\xi \to 0$ in distribution (1), whereas for

$\xi > 0$, the distribution tail decay is polynomial. For $\xi < 0$, X has a finite upper end-point at $u - \sigma_u/\xi$ but is unbounded above for $\xi \geq 0$. To estimate the $(1-p)^{\text{th}}$ quantile, x_p, of X, for $p < \lambda_u := \mathbb{P}(X > u)$, we can solve $\hat{F}(x_p) = 1 - p$, where $\hat{F}(x_p) = 1 - \hat{\lambda}_u[1 - H(x_p - u; \hat{\sigma}_u, \hat{\xi})]$, $\hat{\lambda}_u$ is the proportion of the realisations of X exceeding u and $(\hat{\sigma}_u, \hat{\xi})$ are maximum likelihood estimates (MLEs) obtained by using realisations of the threshold excesses. Davison and Smith (1990) overview the properties of the GPD.

Threshold selection involves a bias-variance trade-off: too low a threshold is likely to violate the asymptotic basis of the GPD, leading to bias, whilst too high a threshold results in very few threshold excesses with which to fit the model, leading to large parameter and return level uncertainty. Thus, we must choose as low a threshold as possible subject to the GPD providing a reasonable fit to the data. There are a wide variety of methods aiming to tackle this problem (Scarrott and MacDonald, 2012; Belzile et al., 2023) with the most commonly used methods suffering from subjectivity and sensitivity to tuning parameters.

A novel automated approach to threshold selection is introduced by Varty et al. (2021) specifically for modelling large, human-induced earthquakes. These data are complex due to improvements in measurement equipment over time. The major implication of such change is that data are missing-not-at-random, with the dataset appearing to be realisations of a non-identically distributed variable, requiring a threshold $u(t)$ which varies with time t, even though the underlying process is believed to be identically distributed over t. Since excesses of $u(t)$ do not have the same GPD parameters over time, Varty et al. (2021) transform these to a common standard exponential distribution via the probability integral transform, using estimates of $(u(t), \sigma_{u(t)}, \xi)$. They then quantify the model fit using a metric based on a QQ-

plot and select a time-varying threshold that optimizes this metric. The key novel aspect of their assessment is the use of bootstrapping methods in the metric evaluation which fully accounts for the uncertainty in the GPD fit, which varies across threshold choices.

Due to the lack of existing threshold selection methods designed for the context of Varty et al. (2021), that paper focuses on the data analysis rather than investigating the performance of the threshold selection method. We explore how their ideas can be best adapted to threshold selection in a univariate, iid data context. We find that a variant of the Varty et al. (2021) metric improves the performance and leads to substantially better results than existing automated methods, including greater stability with respect to tuning parameters.

We differ from Varty et al. (2021) as we study both threshold selection and return level estimation when the truth is known. We also address an entirely different problem of how to incorporate the uncertainty resulting from threshold selection into return level estimation. Existing methods typically treat the threshold, once it has been selected, as known, for subsequent return level inference. The available data above candidate threshold choices are often few and so inference can be highly sensitive to the chosen threshold. Reliance on a single threshold leads to poor calibration of estimation uncertainty and as a result, can mislead inference. In particular, we show that the resulting confidence intervals for such an approach considerably under-estimate the intended coverage. We propose a novel and simple method, based on a double-bootstrap procedure, that incorporates the uncertainty in the selected threshold during inference. We show that the coverage probabilities of confidence intervals from our approach are close to the required nominal levels, thus ensuring our inferences provide meaningful information for design policies.

Ultimately, our aim is to provide a threshold selection method that does not require any user decisions to achieve adequate results. For example, the method should not be sensitive to the choice of candidate threshold grid, it should not require the estimation of a mode to select this grid, it should not have a limit on the number of candidate thresholds for a given sample size, nor should it exclude the possibility that the available data have been pre-processed, such as containing only the exceedances of some arbitrary level.

In Section 2, we illustrate problems with threshold selection and outline existing strategies. Section 3 describes the core existing automated methods while Section 4 introduces our procedure for the selection of a threshold, contrasting it with that of Varty et al. (2021). Section 5 presents our proposed method for incorporating threshold uncertainty into return level inference. In Section 6, the proposed methods are compared against existing methods on simulated data. In Section 7, we apply our methodology to the widely studied troublesome dataset of the River Nidd, first analysed by Davison and Smith (1990).

2 Background

The *threshold stability property* of the GPD is key in many threshold selection approaches: if excesses of a threshold u follow a GPD then excesses of a higher threshold v ($u < v < x^F$) will also follow a GPD, with adjusted parameter values, i.e., if $X - u|(X > u) \sim \text{GPD}(\sigma_u, \xi)$, then $X - v|(X > v) \sim \text{GPD}(\sigma_u + \xi(v - u), \xi)$, see supplementary material S:2.1. By this property, the GPD shape parameter ξ should have the same value for all valid choices of threshold. A modelling threshold can be selected as the lowest value for which this property holds, accounting for the sampling variability in the estimates of ξ. The conventional method

for this assessment is known as a *parameter stability plot* (Coles, 2001). This plot displays the estimates of ξ and their associated confidence intervals (CIs) for a set of candidate thresholds. The threshold is selected as the lowest value for which the estimate of ξ for that level is consistent with estimates of ξ at all higher thresholds. Throughout the paper, we use maximum likelihood estimation and parametric bootstrap-based CIs.

Figure 1 shows two parameter stability plots, with the left plot for a simulated dataset of 1000 values generated from the Case 4 distribution, described in Section 6, where excesses of the threshold $u = 1.0$ follow a GPD(0.6, 0.1); and the right plot for 154 measurements from the River Nidd. Each plot has 95% CIs of two types; the delta method and the bootstrap. Profile log-likelihood based CIs were also evaluated but resulted in very similar intervals to the bootstrap method, so they were omitted. The delta method gives narrower CIs, though close to the bootstrap intervals for the larger dataset. Selecting an appropriate threshold using this method is challenging and subjective as the parameter estimates are dependent across threshold choices, there is a high level of uncertainty due to the small sample sizes that characterise extreme value analyses, and the uncertainty increases with threshold choice.

For the Case 4 data, the plot shows that candidate thresholds above (below) 0.3 are possibly appropriate (not appropriate) as CIs for higher candidate thresholds include (exclude) the corresponding shape parameter estimates, and above 0.8 the point estimates appear more stable. Here $(u, \xi) = (1, 0.1)$, so we can see that candidates below 0.3 are not suitable as ξ is outside their CIs, but the true threshold is higher than may be selected using this plot. For the River Nidd, lower candidate threshold values imply a very heavy-tailed distribution ($\hat{\xi} \approx 0.5$), whilst high candidate thresholds imply a very short tail, with estimates ($\hat{\xi} \approx -0.5$).

As a result of this unusual behaviour, the Nidd data has become the primary example for non-trivial threshold selection (Davison and Smith, 1990; Northrop and Coleman, 2014). We apply our new method to this dataset in Section 7. Further examples of the problems encountered when using parameter stability plots are given in supplementary material S:2.2.

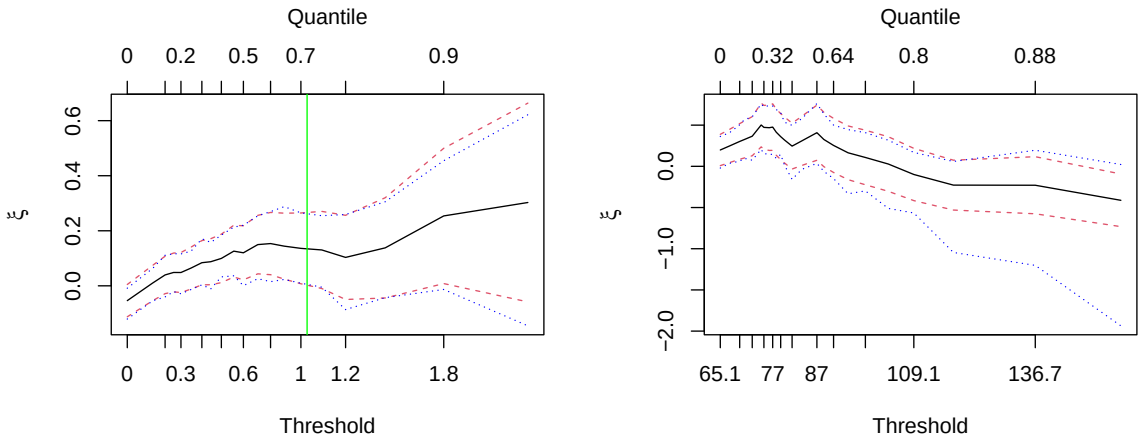

Figure 1: Examples of parameter stability plots with pointwise CIs using the delta-method [dashed] and bootstrapping [dotted] for [left] a simulated dataset with true threshold $u = 1.0$ following Case 4 distribution [green-vertical] and [right] the River Nidd dataset.

Scarrott and MacDonald (2012) and Belzile et al. (2023) review the extensive literature that aims to improve upon parameter stability plots. The latter characterises these methods, with a core reference, as follows: penultimate models (Northrop and Coleman, 2014), goodness-of fit diagnostics (Bader et al., 2018), sequential-changepoint approaches (Wadsworth, 2016), predictive performance (Northrop et al., 2017), and mixture models (Naveau et al., 2016). It also discusses semi-parametric inferences (Danielsson et al., 2001), but it excludes the development by Danielsson et al. (2019), with similarities to the goodness-of-fit approaches.

In Section 3, we outline the key aspects of the core automated approaches with which we compare our proposed method. Supplementary material S:2.3 and S:2.4 describe Northrop and Coleman (2014) and Danielsson et al. (2001, 2019) respectively, finding that the former

suffers from subjectivity of interpretation similar to the parameter stability plots. We do not describe any mixture methods in this paper as although they benefit from accounting for threshold uncertainty, their inferences are strongly dependent on the choice of model for below the threshold, which we feel is inconsistent to the strategy of extreme value modelling and is likely to induce bias in the threshold selection and subsequent quantile estimation.

3 Existing automated threshold selection methods

Automated threshold selection methods aim to remove subjectivity from the choice of threshold by selecting an optimal threshold from a set of user-defined candidate thresholds based on optimising some criterion. We outline and compare the approaches of Wadsworth (2016) and Northrop et al. (2017), which we find to perform best of the considered existing methods. Further details of these methods are given in supplementary material S:3.

Wadsworth (2016) addresses the dependence between MLEs of ξ, denoted by $\hat{\boldsymbol{\xi}}$, over candidate thresholds. Using asymptotic theory for the joint distribution of MLEs from overlapping samples of data, $\hat{\boldsymbol{\xi}}$ are transformed to the vector $\hat{\boldsymbol{\xi}}^*$ of normalised increments between successive $\hat{\boldsymbol{\xi}}$ values. For GPD data, asymptotically, $\hat{\boldsymbol{\xi}}^*$ would be iid realisations from a standard normal distribution, whereas if the data above any candidate threshold were not from a GPD, the associated elements of $\hat{\boldsymbol{\xi}}^*$ would be better approximated by a non-standard normal. This changepoint behaviour is used to select the threshold. The underlying asymptotic arguments can cause considerable threshold sensitivity and the failure of the method to converge. Both issues are exacerbated by small samples and we identify systematic failures of the associated open source software when $\xi < 0$. To reduce such problems, Wadsworth (2016, Table 1)

provides guidance on the number of candidate thresholds for a given sample size.

Northrop et al. (2017) model data using the binomial-GPD (BGPD) model, which is GPD above u, with $\lambda_u = \mathbb{P}(X > u)$ a model parameter, and an improper uniform density, of value $1 - \lambda_u$, below u. They use Bayesian inference and, for each candidate threshold, assess the predictive density of GPD fits above a fixed validation threshold v, where v is the largest candidate threshold. The selected threshold maximises the predictive ability of this model, above v, using leave-one-out cross-validation. The method is sensitive to the validation and candidate threshold set and to the prior joint density of the BGPD parameters.

4 Novel metric-based constant threshold selection

4.1 Metric choice

We propose an adaptation of the Varty et al. (2021) approach to identify the threshold u for which the sample excesses, arising from iid and non-missing realisations of a continuous random variable, are most consistent with a GPD model. Both methods use a QQ-plot-based metric to approximate the integrated absolute error (IAE) between the quantiles of the model and the data-generating process. Our method, the *expected quantile discrepancy* (EQD), uses the data on the original scale. In contrast, the method of Varty et al. (2021) transforms the data to an Exponential(1) marginal scale and will be termed the *Varty method*.

The following makes the difference between the two methods precise. Let $\boldsymbol{x}_u = (x_1, \ldots, x_{n_u})$ be the sample of n_u excesses of candidate threshold u and $\boldsymbol{q} = \{q_i = (i-1)/(n_u - 1) : i = 1, \ldots, n_u\}$ be the vector of probability plotting points corresponding to the sample size of $\boldsymbol{x}_u$.

The sample quantile function $Q(\cdot; \boldsymbol{x}_u, \boldsymbol{q}) : [0,1] \to \mathbb{R}^+$ is defined as the linear interpolations of the points $\left\{(q_i, x_u^{(i)}) : i = 1, \ldots, n_u\right\}$, with $x_u^{(i)}$ denoting the i^{th} order statistic of $\boldsymbol{x}_u$ (increasing with i), where any ties are handled similarly through linear interpolation. The transformation to Exponential(1) margins is defined by $T(x; \sigma, \xi) = F_{\text{Exp}}^{-1}\{H(x; \sigma, \xi)\}$ where F_{Exp}^{-1} is the inverse distribution function of an Exponential(1) variable, H is the GPD function (1), and let $T(\boldsymbol{x}_u; \sigma_u, \xi) = \{T(x_1; \sigma_u, \xi), \ldots, T(x_{n_u}; \sigma_u, \xi)\}$. To incorporate the effect of sampling variability in the data into the threshold choice, the expected (average) deviation over the QQ-plot, calculated for the probabilities $\{p_j = j/(m+1) : j = 1, \ldots, m\}$, is calculated across bootstrapped samples of $\boldsymbol{x}_u$, denoted $\boldsymbol{x}_u^b$ for the b^{th} bootstrap sample, $b = 1, \ldots, B$. For both methods, this results in the overall measure of fit $\hat{d}_E(u) = \sum_{b=1}^{B} d_b(u)/B$, where

$$d_b(u) = \begin{cases} \frac{1}{m} \sum_{j=1}^{m} \mid \frac{\hat{\sigma}_u^b}{\hat{\xi}_u^b}[(1-p_j)^{-\hat{\xi}_u^b} - 1] - Q(p_j; \boldsymbol{x}_u^b, \boldsymbol{q}) \mid & \text{EQD} \\ \frac{1}{m} \sum_{j=1}^{m} \mid -\log(1-p_j) - Q(p_j; \hat{T}(\boldsymbol{x}_u^b; \hat{\sigma}_u^b, \hat{\xi}_u^b), \boldsymbol{q}) \mid & \text{Varty}, \end{cases} \tag{2}$$

and $(\hat{\sigma}_u^b, \hat{\xi}_u^b)$ are the estimated GPD parameters fitted to the bootstrapped sample $\boldsymbol{x}_u^b$. The selected threshold minimises the estimated IAE, $\hat{d}_E$, over a set of candidate thresholds. In Sections 4.2 and 4.3 respectively, we justify the choices made in the formulation of the EQD metric and discuss our recommendation for default values for the tuning parameters (B, m).

In supplementary material S:5.2, we compare the EQD and Varty methods through an extensive simulation study to assess which version of metric (2) performs better for threshold selection and quantile estimation. For threshold selection, the methods perform similarly; each method narrowly achieves the smallest root-mean-square error (RMSE) in two of Cases

1-4, discussed in Section 6. However, for the estimation of high quantiles, the EQD outperforms the Varty method obtaining the lowest RMSE in the majority of cases and quantiles, due to having the smaller variance of estimates. We ultimately aim to estimate high quantiles accurately following threshold selection. Given that this study indicates that the EQD should be preferred for this aim and to avoid unnecessary repetition, we omit the results for the Varty method for the remainder of the paper.

4.2 Investigation of the EQD metric choice

For a given u, $d_b(u)$ evaluates the mean-absolute deviation between the b^{th} bootstrap sample quantiles and the fitted model-based quantiles, i.e., the mean-absolute deviation from the line of equality in a QQ-plot for that particular bootstrap sample. This type of assessment by itself is not radical, as for any observed sample data, QQ-plots are the standard method of assessing model fit (Coles, 2001). The novelty for assessing the validity of a candidate threshold u comes from the way that the EQD metric is constructed.

There are a number of novel choices which we have made in the EQD metric that require justification, in particular; the use of the mean-absolute deviation; the choice of quantiles and their interpolation in the QQ-plot; the use of bootstrap samples; and that the observed data are not explicitly used in the metric. We examine each of these features in the supplementary material, through simulation studies involving the case studies of Section 6. For each feature, we find positive evidence for our selections. Below, we explain why we made these choices and outline how they performed relative to other alternative formulations.

We focus on the mean-absolute deviation on the QQ scale as Varty et al. (2021) found that

this was more effective than using the mean-squared deviation on that scale and either metric on the PP scale. Our simulation studies also found this to be a more robust measure of fit than the maximum deviation, as proposed by Danielsson et al. (2019).

We choose to take $\{p_j\}$ to be equally-spaced and to weight contributions to $d_b(u)$ equally across the corresponding quantiles. Although higher (lower) sample quantiles exhibit greater (less) sampling variability, equal weighting is appropriate when taking the $\{p_j\}$ values to be equally-spaced because for any $\xi > -1$, the GPD density is monotonically decreasing. This leads to dense evaluation for lower sample quantiles and more sparse evaluation in the upper tail. The choice of equal weighting on this scale is motivated and supported by empirical evidence in Varty et al. (2021). Our choice for p_j is based on the expression for plotting points in a QQ-plot assessment we use in previous research and the choice for q_i is the default option for the R `quantile` function. As these choices are subjective, we also consider alternative definitions but find that there is no systematic ordering of the performance over these definitions and any differences in RMSE for the thresholds selected by the EQD are minimal, especially when compared to the differences between the EQD and existing methods.

The average over bootstrapped samples in metric (2) is not a standard use of bootstrapping, i.e., we utilise the bootstraps in a measure of fit rather than to describe the uncertainty in some estimated quantity. Our aim in doing this is to account for the sampling variability in the observed data, thus avoiding over-fitting of the GPD model to the observed dataset which could lead to higher threshold choices than necessary, reduced numbers of exceedances, and extra uncertainty in parameter and quantile estimates. To confirm this, we considered

using only the observed sample values in the metric. This leads to higher and more variable thresholds choices in a variety of cases and an overall performance which is either noticeably worse or at best, comparable to our approach.

One may also be concerned that $\boldsymbol{x}_u$ is not included directly in metric (2). To address this, we additionally explored the effect of using $Q(p_j; \boldsymbol{x}_u, \boldsymbol{q})$ instead of $Q(p_j; \boldsymbol{x}_u^b, \boldsymbol{q})$ within the EQD metric, despite it being unconventional to compare sample quantiles to those of a model fitted using a different (bootstrapped) sample. We found no benefit to doing so. Moreover, using only $\boldsymbol{x}_u$ to estimate the IAE ignores that this estimate would change for another realisation of the same data generating process and that variability in this estimate increases with u. Our approach utilises the bootstrap resamples in the measure of fit to provide more stability in the threshold choice and allow us to account for the increasingly uncertain parameter and quantile estimates as the threshold increases.

4.3 Choice of tuning parameters

An in-depth study in supplementary material S:5.3 demonstrates that the EQD method is robust to the choice of the tuning parameters B and m. Consequently, we take $(B, m) = (100, 500)$ throughout the paper and in the supplementary material, unless stated otherwise.

The number of bootstrapped samples B controls the level of sampling variability that is incorporated into the threshold choice and so we expect higher values of B to lead to more stable threshold choices. The RMSE values for threshold estimation reflect this but also show that computation time increases linearly with B. For a one-off analysis, there is certainly merit in taking as large a value for B that is computationally feasible. For simulation studies,

when the computational implications of the choice of B are more important, we find that $B = 100$ balances accuracy and computation time sufficiently.

The tuning parameter m gives the number of equally-spaced evaluation probabilities used in expression (2). The EQD metric aims to approximate the IAE between model quantiles and quantiles of the data generating process (i.e., not for a particular sample) and a larger choice of m improves this approximation. To compare fairly across a range of candidate thresholds, we choose to keep the quality of the approximation of the IAE fixed across thresholds and bootstraps by fixing the number m of points in the quantile interpolation grid.

For a particular bootstrap sample, this choice of fixed m can lead to under- or over-sampling of the upper tail depending on whether $m < n_u$ or $m > n_u$. We explore the sensitivity of the EQD method to m with $m = cn$ and $m = cn_u$, with $c = 0.5, 1, 2, 10$. For both strategies, we find that increasing m beyond 500 essentially wastes the increased computation time as the RMSE values for threshold estimates showed little sensitivity to m. We also explore the effect of the interpolation grid on the sampling distribution of $d_b(u)$ values, over different thresholds, when evaluated using $m = 500$ or $m = n_u$. We find that there is little effect from the choice of interpolation grid outside of the very highest candidate thresholds, but these differences have no effect on the selected threshold in our examples. We conclude that $m = 500$, is suitable as a default value in practice but we can see merits in also ensuring that $m \geq \max_u(n_u)$, where the maximisation is across all candidate thresholds.

5 Accounting for parameter and threshold uncertainty

Even if the true threshold u is known, relying on point estimates for the GPD parameters results in misleading inference (Coles and Pericchi, 2003). CIs are needed, but as standard errors and profile likelihoods rely on asymptotic arguments, they are not ideal due to the sparsity of threshold exceedances. We prefer parametric bootstrap methods which, as discussed in Section 2, perform similarly to the profile likelihood for large samples. Algorithm 1 details the bootstrapping procedure to account for GPD parameter uncertainty when u is known. A GPD is fitted to the n_u data excesses of u from a sample $\boldsymbol{x}$ of size n ($n \geq n_u$). Using the fitted parameters, B_1 parametric bootstrap samples above u are simulated, each of size n_u, and the GPD is re-estimated for each sample. A summary statistic, e.g., a return level, $s(u, \lambda_u, \sigma_u, \xi)$, may be computed for each of the B_1 bootstrapped values for (σ_u, ξ). This enables the construction of CIs for the GPD parameters and return levels.

Algorithm 1 Parameter uncertainty for known threshold

Require: $(\boldsymbol{x}, u, B_1)$
  Find $n_u = \#\{i : x_i > u\}$, set $\hat{\lambda}_u = n_u/n$, and fit a GPD to $\boldsymbol{x}$ data above u to give $(\hat{\sigma}_u, \hat{\xi}_u)$.
  for $b = 1, \ldots, B_1$ **do**
    Simulate sample $\boldsymbol{y}_u^b$ consisting of n_u excesses of u from GPD$(\hat{\sigma}_u, \hat{\xi}_u)$.
    Obtain parameter estimates $(\hat{\sigma}_b, \hat{\xi}_b)$ for $\boldsymbol{y}_u^b$ and summary of interest $s(u, \hat{\lambda}_u, \hat{\sigma}_b, \hat{\xi}_b)$.
  end for
  return A set of B_1 bootstrapped estimates for the summary statistic of interest.

Algorithm 1 focuses on the uncertainty of the estimates of (σ_u, ξ). We incorporate the additional uncertainty in the estimation of λ_u by replacing the fixed n_u in the loop over b with a random variate from a Bin$(n, \hat{\lambda}_u)$ distribution for each bootstrap sample, with this extension then referred to as Algorithm 1b.

GPD inferences are sensitive to the choice of threshold (Davison and Smith, 1990) but

uncertainty about this choice is not represented in Algorithms 1 or 1b. This omission would be important when return levels inform the design of hazard protection mechanisms, where omitting this source of uncertainty could lead to over-confidence in the inference and have dangerous consequences. Algorithm 2 provides a novel method to propagate both threshold and parameter uncertainty through to return level estimation, using a double-bootstrap procedure. To focus on the threshold uncertainty and to forgo the need for a parametric model below the threshold, we employ a non-parametric bootstrap procedure on the original dataset. We resample with replacement n values from the observed data B_2 times, estimate a threshold for each such bootstrap sample using the automated selection method of Section 4, and fit a GPD to the excesses of this threshold. For each one of the B_2 samples, we employ Algorithm 1 to account for the subsequent uncertainty in the GPD parameters. Calculating a summary statistic for each of the $B_1 \times B_2$ samples leads to a distribution of bootstrapped estimates that accounts for uncertainty in the threshold selection as well as in the GPD and threshold exceedance rate parameters. We use $B_1 = B_2 = 200$. To run this algorithm using the EQD method for the threshold selection step (which itself has B bootstraps), it would require $B_2(B + B_1)$ bootstrap samples to be generated. Specifically, for the B_2 samples initially generated for Algorithm 2, we have $B_2 \times B$ in selecting the threshold values and $B_2 \times B_1$ in capturing the GPD parameter uncertainty above these selected thresholds. this can res In Section 6, we illustrate how using Algorithm 2 improves the coverage probability of CIs, and in Section 7 how it widens CIs for return levels of the River Nidd.

Algorithm 2 Parameter uncertainty for unknown threshold

Require: $(\boldsymbol{x}, n, B_2, B_1)$
  for $b = 1, \ldots, B_2$ **do**
    Obtain sample $\boldsymbol{x}_b$ of size n by sampling n times with replacement from $\boldsymbol{x}$.
    Estimate threshold $\hat{u}_b$ for $\boldsymbol{x}_b$ and record number of excesses as $n_{\hat{u}_b}$.
    Employ Algorithm 1 with inputs: $(\boldsymbol{x}_b, \hat{u}_b, B_1)$.
  end for
  return A set of $B_1 \times B_2$ bootstrapped estimates for the summary statistic of interest.

6 Simulation study

6.1 Overview

We illustrate the performance of the EQD method against the Wadsworth (2016) and Northrop et al. (2017) approaches, which we term the *Wadsworth* and *Northrop* methods respectively. Danielsson et al. (2001, 2019) approaches perform considerably worse than all others in threshold selection and quantile estimation; so results for these methods are only given in supplementary material **??**. We utilised the following R code for Wadsworth, Northrop and EQD methods respectively: code given in the supplementary materials of Wadsworth (2016), *threshr* (Northrop and Attalides, 2020), and `https://github.com/conor-murphy4/automated_threshold_selection` (Murphy et al., 2023).

The performance of all of the methods depends somewhat on the choice of the set of candidate thresholds which we denote by:

$$C_u = \{u_i, i = 1, \ldots, k : u_1 < \ldots < u_k\}, \tag{3}$$

where we restrict the u_i to be sample quantiles evaluated at equally-spaced probabilities. The range $[u_1, u_k]$, the number of candidates k and the inter-threshold spacing are all potentially

important in terms of how they affect the performance of the methods. As emphasised in Section 1, we are aiming for an automated threshold selection method which can achieve accurate results without any user inputs, so a key element of our study is to investigate how these features of the set C_u impact on the methods' relative performance. When fitting a GPD with decreasing density (i.e., for $\xi > -1$), it would be inadvisable to use a threshold which clearly lies below the mode of the distribution. As we want to avoid the requirement of user estimation of the mode, our standard choice for the range of the candidate grid is $[u_1, u_k]$: $(u_1, u_k) = (0\%, 95\%)$ sample quantiles of all the data. However, we also explore several cases where only the data lying above the mode are used with $[u_1, u_k]$: $(u_1, u_k) = (0\%, 95\%)$ now sample quantiles of the remaining data. To remove the uncertainty arising from the choice of estimator of the mode, we use the true mode which has a unique value in our simulated cases. Results in supplementary material **??** indicate that our original choice for the candidate threshold set does not unfairly favour the EQD method in any way.

We consider two scenarios: Scenario 1 and Scenario 2 where the true threshold is known and unknown respectively. Here, we present the results using a candidate threshold grid across the whole distribution for Scenario 1 and above the sample median for Scenario 2, with the latter chosen as the Wadsworth method fails when applied across the default range in that setting. The Wadsworth method relies on asymptotic arguments, which limits how large k can be relative to the sample size, n', above the mode, with $n' \leq n$, where n is the total sample size. To assess how the Wadsworth method performs as a fully automated method, we apply the method despite the value of k not always aligning with the guidance in Wadsworth (2016) about its size relative to n'.

We assess the methods' ability to estimate the true threshold (when it exists) and the true return levels, using the RMSE to measure performance. The true quantiles and all bias-variance components of RMSE, discussed in this section, are given in supplementary materials S:4 and **??** respectively. We also investigate the merits of including the uncertainty in threshold selection in our inference, as discussed in Section 5, in terms of how they improve the coverage levels of CIs relative to their nominal values.

6.2 Scenario 1: True threshold for GPD tail

We consider Cases 0-8, with different properties above and below the true threshold of $u = 1.0$ and various sample sizes. Case 0, where all of the data are from a GPD, is reported in supplementary material **??**, with the EQD performing notably better than the existing methods. Here, we present detailed results for Cases 1-4, with Table 1 providing outline model and sample size properties, with full details and density plots given in supplementary material S:4. Cases 5-8 are considered briefly after discussing Cases 1-4 below.

Cases 1-3 all have a distinct changepoint in the density and density mode both at the true threshold which should make all methods of threshold selection perform better than in situations without either of these features. Cases 1 and 2 have the same distribution, with $\xi > 0$, with Case 2 having a smaller sample size. We find that the Wadsworth method fails to estimate a threshold in samples with $\xi < -0.05$ irrespective of sample size, so Case 3 is selected near that boundary where the method works reasonably and has double the sample size of Case 1. Case 4 provides a more difficult example with a continuous density and a small number of exceedances of the true threshold. The data are derived from a partially

observed GPD, denoted GPD_p, with data drawn from a GPD above 0 and rejected if less than an independent realisation from a Beta(1,2) distribution.

For each case, the results are based on 500 replicated samples, for which we test the candidate thresholds C_u, with $k = 20$, as given in (3), with the true threshold being the 16.67% quantile for Cases 1-3 and the 72.10% quantile for Case 4.

Models	Below threshold	Sample size	Above threshold	Sample size
Case 1	$\text{U}(0.5, 1.0)$	200	$\text{GPD}(0.5, 0.1)$	1000
Case 2	$\text{U}(0.5, 1.0)$	80	$\text{GPD}(0.5, 0.1)$	400
Case 3	$\text{U}(0.5, 1.0)$	400	$\text{GPD}(0.5, -0.05)$	2000
Case 4	$\text{GPD}_\text{p}(0.5, 0.1)$	721	$\text{GPD}(0.6, 0.1)$	279

Table 1: Model specifications for Cases 1-4.

Cases 1-4, Threshold recovery: Table 2 shows the RMSE of the chosen thresholds for each method in Cases 1-4, with the EQD achieving RMSEs 1.2-7.7 (1-11.2) times smaller than the Wadsworth (Northrop) method. The EQD has the lowest bias by a considerable margin in Cases 1-3 and shows the lowest variance in threshold estimation in all cases. In fact, the variance is reduced by a factor of at least 20 relative to both the Wadsworth and Northrop methods (see Table **??**). The very strong performance of the EQD relative to both the Wadsworth and Northrop methods is particularly noteworthy in Cases 1-3, and is also seen for Case 0 and later for Cases 5-7. We believe that the key reason for this is the discontinuity in the density, a feature common to all of these cases, as that appears to lead to a very small bias for the EQD method relative to the other methods. Specifically, the variance penalty of the EQD metric seems to push the threshold as low as possible, but its complementary goodness-of-fit measure almost entirely stops the threshold being selected below the clear discontinuity in the density. For Case 4, which has a continuous density, the

EQD achieves the smallest RMSE almost entirely due to it having the smallest variance but with a bias component broadly comparable with the other methods.

	EQD	*Wadsworth*[1]	*Northrop*
Case 1	**0.048**	0.349	0.536
Case 2	**0.060**	0.461	0.507
Case 3	**0.060**	0.221	0.463
Case 4	**0.526**	0.628	0.543

Table 2: RMSE of the threshold choices for each method-case combination. The smallest values for each case are highlighted in bold.

Cases 1-4, Quantile recovery: Table 3 presents the RMSEs for the $(1 - p_{j,n})$-quantiles where $p_{j,n} = 1/(10^j n)$ for $j = 0, 1, 2$ for sample size n, which ensures that extrapolation is equally difficult over n for a given j. When $j = 0$, no extrapolation is required so the choice of u should not be too important; the similar RMSEs across methods reflect this. As j increases, all RMSEs increase and the differences between methods become clear. The EQD method is best uniformly, followed by the Wadsworth and then the Northrop methods. This pattern reflects the findings in Table 2, although here with differential performances sensitive to j. However, in terms of quantile estimation, the EQD method does not retain the large differential relative to the other methods which was seen for threshold selection in Cases 1-3. In contrast, the differences between methods in Case 4 are now more apparent, as controlling the variance is more important than any small differences in bias when we are concerned with a RMSE assessment of quantiles which lie far into the tail. The EQD achieves the lowest bias in the majority of cases and leads to quantile estimates with considerably less variance in all cases, particularly as j increases.

Summary for Cases 5-8: Cases 5-8 are very similar in form to Cases 1-3 but with different

[1]Results for Wadsworth are calculated only on the samples where a threshold was estimated. It failed estimate a threshold for 2.4%, 26.4%, 0%, 4.4% of the simulated samples in Cases 1-4, respectively.

	EQD	*Wadsworth*[1]	*Northrop*	*EQD*	*Wadsworth*[1]	*Northrop*
j	**Case 1**			**Case 2**		
0	**0.563**	0.594	0.755	**0.599**	0.631	0.736
1	**1.258**	1.391	2.376	**1.488**	1.644	3.513
2	**2.447**	2.717	7.097	**3.119**	3.484	22.916
	Case 3			**Case 4**		
0	**0.190**	0.195	0.230	**0.677**	0.800	0.791
1	**0.323**	0.344	0.450	**1.563**	2.059	2.217
2	**0.483**	0.516	0.744	**3.043**	4.485	5.568

Table 3: RMSEs in the estimated quantiles in Cases 1-4 based on fitted GPD above chosen threshold. The smallest RMSE for each quantile are highlighted in bold.

shape parameters and sample sizes. The results for these cases are presented in supplementary material **??**, with a brief summary given here. Specifically, for Cases 5-7, we find that the EQD exhibits the strongest performance and the Wadsworth method consistently fails due to the small sample sizes or computational issues with numerical integration when $\xi < -0.05$. Case 8 is parameterised similarly to Case 1 but with an unrealistic sample size of $n = 20000$. Although the data in Case 8 are more suited to a method reliant on asymptotic theory, the EQD performs comparably with the Wadsworth method, with both performing better than the Northrop method.

Case 4, True quantile coverage: We apply Algorithms 1, 1b and 2 to data from Case 4, the hardest case for threshold selection. Table 4 presents the coverage probabilities of the nominal 80% and 95% CIs of the estimated $(1-p_{j,n})$-quantiles as well as the average ratio of the CI widths (based on *Alg 2* relative to *Alg 1*) over the 500 samples, termed CI ratio. Results for extra quantile levels, as well as coverage for the 50% CI, are given in supplementary material **??**. Overall, incorporating only parameter uncertainty (*Alg 1* and *Alg 1b*) leads to underestimation of interval widths and inadequate coverage of the true quantiles, especially as we extrapolate further. The additional uncertainty, given in *Alg 1b*, by also accounting for

uncertainty in the rate of threshold exceedance, typically makes a very small improvement in coverage, and for some quantiles, this actually leads to a reduction in coverage due to Monte Carlo variation in the simulations. In contrast, the inclusion of the additional threshold uncertainty (*Alg 2*) leads to much more accurate coverage of the true quantiles across all exceedance probabilities. The CI ratios show that this highly desirable coverage is achieved with only 43-62% increase in the CI widths on average.

	80% confidence			95% confidence		
j	0	1	2	0	1	2
Alg 1	0.646	0.618	0.606	0.834	0.804	0.794
Alg 1b	0.656	0.638	0.612	0.830	0.814	0.794
Alg 2	0.798	0.772	0.758	0.954	0.948	0.944
CI ratio	1.430	1.452	1.475	1.484	1.546	1.621

Table 4: Coverage probabilities for estimated quantiles using Algorithms 1, 1b and 2 for 500 replicated samples from Case 4 with sample size of 1000. CI ratio gives the average ratio of the CIs for Algorithm 2 relative to Algorithm 1 over the 500 samples.

6.3 Scenario 2: Gaussian data

In applications, there is no true or known value for the threshold above which excesses follow a GPD, so we explore this case here. We select the standard Gaussian distribution as it has very slow convergence towards an extreme value limit (Gomes, 1994), so threshold selection is likely to be difficult. We assess threshold selection methods based on estimation of the true quantiles $\Phi^{-1}(1-p_{j,n})$ where $p_{j,n} = 1/(10^j n)$, for $j = 0, 1, 2$. We simulate 500 samples, for $n = 2000$ and 20000, with C_u, given in (3), now having range $[u_1, u_k] : (u_1, u_k) = (50\%, 95\%)$ sample quantiles of the data and $k = 10$ and 91 (i.e., steps of 5% and 0.5%) for the two choices of n respectively. As with Case 8 in Section 6.2, $n = 20000$ is unrealistic, but we include it to illustrate the slow convergence.

Quantile recovery: Table 5 shows the RMSEs of the estimated quantiles. For $n = 2000$, the EQD method achieves the smallest RMSE with the Northrop method a close second, with the reverse when $n = 20000$. The median and 95% CI of the chosen thresholds are given in supplementary material **??**. The Northrop method tends to choose slightly higher thresholds than the EQD method, leading to a small reduction in bias, but for only the smaller n is the additional variability relative to the EQD a disadvantage. The Wadsworth method performs the worst, selecting lower thresholds and so incurring the most bias.

	$n = 2000$			$n = 20000$		
j	*EQD*	*Wadsworth*[2]	*Northrop*	*EQD*	*Wadsworth*	*Northrop*
0	**0.214**	0.239	0.225	0.187	0.214	**0.172**
1	**0.430**	0.529	0.461	0.368	0.422	**0.331**
2	**0.703**	0.890	0.765	0.594	0.672	**0.533**

Table 5: RMSEs of estimated $(1-p_{j,n})$-quantiles for 500 replicated samples from a Gaussian distribution for samples of size n. The smallest RMSE are highlighted in bold.

True quantile coverage: For assessing the coverage of true quantiles using Algorithms 1, 1b and 2 for Gaussian data, Table 6 presents the coverage probabilities of the nominal 80% and 95% CIs of the estimated quantiles, when $n = 2000$, as well as the average ratio of the CI widths (again, of *Alg 2* relative to *Alg 1*) over the 500 samples, with more results given in supplementary material **??**. Across the $p_{j,n}$, both *Alg 1 and 1b* give very low coverage probabilities in both cases, with performance deteriorating as j increases. The added threshold uncertainty from *Alg 2* results in large increases in coverage though still somewhat less than required, with this achieved through increases in CI widths by 45-66% on average. This weaker performance than we find in Section 6.2 suggests that no sample threshold (for realistic sample sizes) is large enough to overcome bias in making

[2]Results for the Wadsworth method, which failed on 0.4% of the samples here, are calculated only for samples where a threshold estimate was obtained.

extreme value approximations for Gaussian data, but the improvement in coverage using *Alg 2* demonstrates the importance of including the additional threshold uncertainty.

	80% confidence			95% confidence		
j	0	1	2	0	1	2
Alg 1	0.588	0.450	0.366	0.750	0.618	0.510
Alg 1b	0.592	0.442	0.364	0.746	0.620	0.508
Alg 2	0.718	0.598	0.492	0.866	0.814	0.756
CI ratio	1.457	1.480	1.509	1.495	1.576	1.665

Table 6: Coverage probabilities for estimated quantiles using Algorithms 1, 1b and 2 for 500 replicated samples from a Gaussian distribution with sample size of 2000. CI ratio gives the average ratio of the CIs for Algorithm 2 relative to Algorithm 1 over the 500 samples.

7 Application to river flow data

The River Nidd dataset consists of 154 storm event peak daily river flow rates that exceeded 65 m^3/s in the period 1934-1969, i.e., an average exceedance rate of 4.4 events per year. Each observation can be deemed "extreme" and iid, though not necessarily well-described by a GPD. Davison and Smith (1990) identify the difficulties these data present for threshold selection and parameter uncertainty, which we reiterated in discussion of Figure 1. Given the small sample size for the River Nidd, any increase in the threshold value is more significant in terms of parameter uncertainty, than for larger datasets studied in Section 6.

Table 7 shows the selected thresholds of each of the methods for a range of candidate grids[3]. The remarkable robustness of the EQD (evaluated with $B = 200$ bootstrap samples) across grids stems from the method's novel incorporation of data uncertainty. The Wadsworth method fails to estimate a threshold unless the grid is made very coarse, and even then

[3]In marked cases, the Northrop method outputted a chosen threshold with some convergence warnings.

exhibits considerable sensitivity (varying between 0% and 90% sample quantiles) over grids of equal size but different endpoints and increments. This is problematic as a coarse grid is likely to remove the most appropriate threshold from consideration. The Northrop method critically depends on the validation threshold, and we find that increasing this level above the 90%-quantile leads to failure or convergence warnings. The thresholds selected by this method are quite variable (between 0% and 80% sample quantiles) over the grids.

Estimated thresholds for the River Nidd dataset			
Grid (% quantile)	*EQD*	*Wadsworth*	*Northrop*
0 (1) 93	67.10 (3%)	NA	68.45[3] (6%)
0 (1) 90	67.10 (3%)	NA	65.08 (0%)
0 (1) 80	67.10 (3%)	NA	100.28 (75%)
0 (20) 80	65.08 (0%)	NA	109.08 (80%)
0 (30) 90	65.08 (0%)	149.10 (90%)	65.08 (0%)
0 (25) 75	65.08 (0%)	100.28 (75%)	81.53 (50%)
0, 10, 40, 70	65.08 (0%)	65.08 (0%)	69.74 (10%)

Table 7: River Nidd dataset selected thresholds (and quantile %) for each method for different grids of candidate thresholds. The Grid column gives *start (increment) end* for each grid.

Comparing thresholds selected between the methods is complicated due to the sensitivity of the Wadsworth and Northrop methods to the grid choice. For the EQD, it is natural to use the densest and widest grid, giving $\hat{u} = 67.10$. This threshold, which is lower than previously found, gives far more data for the extreme value analysis. As all the River Nidd data are "extreme", we believe taking u so close to the lower endpoint of the data is not problematic, and it may indicate that the pre-processing level used to produce these data was too high. The first estimated threshold from the Wadsworth (Northrop) methods, without convergence or warning issues, is $\hat{u} = 149.10$ ($\hat{u} = 65.08$). For these three threshold choices, the corresponding GPD parameter estimates (and 95% CIs) are: $\hat{\sigma}_{u:EQD} = 23.74$ $(17.78, 29.70)$ and $\hat{\xi} = 0.26$ $(0.06, 0.46)$ for the EQD; $\hat{\xi} = -0.15$ $(-1.00, 0.70)$ for the Wadsworth method; and

for the Northrop method, $\hat{\xi} = 0.20\ (0.02, 0.38)$, where we omit the latter two scale parameter estimates as they are estimating different quantities which depend on the threshold, see Section 2. Provided all estimated thresholds are high enough for the GPD to be appropriate, the values of $\hat{\xi}$ should be similar across methods, due to the threshold stability property (see supplementary material S:2). The Wadsworth method leads to an extremely wide CI, which results in meaningless inference. However, the EQD and Northrop findings about ξ are very similar, but the sensitivity to the candidate grid is still a problem for the Northrop method.

Figure 2 shows a QQ-plot for the GPD model using the EQD estimate $\hat{u} = 67.10$. The tolerance bounds show a reasonable agreement between model and data. For $\hat{u}$, Figure 2 also shows the T-year return level estimates, with $1 \leq T \leq 1000$. The 95% CIs incorporate parameter uncertainty alone and both parameter and threshold uncertainty via Algorithms 1 and 2 respectively, with an increase in uncertainty from the latter for larger T; e.g., for the 100- and 1000-year return levels, the CI width increases by a factor of 1.38 and 1.52 respectively. This reiterates how vital it is to incorporate threshold uncertainty into inference.

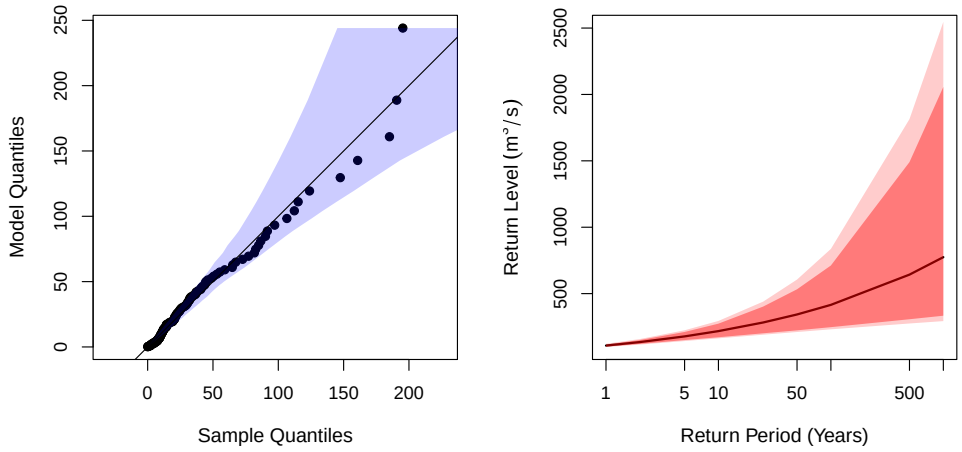

Figure 2: River Nidd analysis: QQ-plot [left] showing model fit with 95% tolerance bounds [shaded] and return level plot [right] based on EQD threshold choice with 95% CIs incorporating parameter uncertainty [dark-shaded] and additional threshold uncertainty [light-shaded].

8 Conclusion and discussion

We proposed two substantial developments to univariate extreme value analysis. Firstly, we addressed the widely-studied problem of how to automatically select/estimate a threshold above which an extreme value, generalised Pareto, model can be fitted. We presented a novel and simple approach, which we termed the EQD method, that minimises an approximation to the IAE of the model quantiles and quantiles of the data generating process. Secondly, we proposed a new approach to improve the calibration of confidence intervals for high quantile inference, addressing an important but under-studied problem. We achieve this through an intuitively simple, but computationally intensive, double-bootstrapping technique which propagates the uncertainty in the threshold estimation through to quantile inference.

Regarding the threshold selection component of the work, we compared the EQD method to the leading existing threshold selection methods in terms of both threshold selection and

consequent high quantile estimation. This was conducted using data from iid continuous univariate random variables and the superiority of the EQD method was illustrated across a range of examples using various metrics. Relative to existing approaches, we showed that the EQD exhibits greater robustness to changes in the set of candidate thresholds, to tuning parameters, and avoids a reliance on asymptotic theory in existing likelihood methods. The EQD method is applicable for all data set sizes and for any set of candidate thresholds.

So why does the EQD method perform much better than the existing approaches? Our analysis has identified two core reasons: the choice of a robust measure of goodness of fit for a given (bootstrapped) sample, which controls bias; and the use of bootstrapped replicates, which leads to reduced variance and also appears to reduce bias. Specifically, in comparison to existing methods, the use of our goodness-of-fit measure, over simply exploiting the GPD threshold stability property, ensures better model fits and hence better threshold selection, and the bootstrapping removes the variation that arises if only the observed sample is used, as that may not be a typical realisation from the underlying data generating process.

In assessing our suggested improvement for the calibration of confidence intervals, we compared the coverage of true quantiles using our proposed approach and the widely-adopted approach of incorporating the GPD parameter uncertainty alone in quantile inference once a threshold has been selected. We showed that the coverage of the existing approach was substantially less than the nominal confidence levels and our proposed approach led to much more reliable confidence intervals without an undue increase in their width.

While this paper has demonstrated the effectiveness of both the EQD method and our proposed approach for confidence interval construction in the univariate iid setting, we believe

that the findings suggest that these approaches could have much wider utility. For example, the Varty et al. (2021) method, which motivated the structure of the EQD method, was originally developed for non-identically distributed data, with the transformation of excesses of a time-varying threshold to a common marginal Exponential(1) distribution. As such cases typically find that excesses have a common shape parameter ξ (Chavez-Demoulin and Davison, 2005), we could use the EQD variant of Varty et al. (2021) by transforming instead to a common GPD with parameters $(1, \xi)$ given we have seen here that by retaining the scale of the original data, the EQD out-performs the Varty et al. (2021) approach. We also believe that the strategy of our new methods could be used to improve threshold estimation in multivariate extremes, in cases of multivariate regular variation assumptions (Wan and Davis, 2019) or for asymptotically independent variables (Heffernan and Tawn, 2004), and allow for the uncertainty in this threshold estimation to be incorporated in the subsequent joint tail inferences. Such developments would naturally have similar implications for spatial extreme value modelling as the threshold selection in this context currently comes down to a multivariate (at the data sites) threshold selection process.

Acknowledgements

We are grateful to Ross Towe (Shell) and Peter Atkinson (Lancaster University) for their support and comments. We also thank the referees and editors for very helpful comments that have improved the presentation and computational evidence of the work.

Supplementary Materials

The software to reproduce Figures 1 and 2 as well as Tables 5 and 7 is provided as separate files in the online supplement. The online supplement also contains a PDF document which provides further information to accompany the main paper including derivations of key properties, further description of methods and additional simulation experiments.

References

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. *Annals of Applied Statistics*, 12(1):310–329.

Belzile, L., Dutang, C., Northrop, P., and Opitz, T. (2023). A modeler's guide to extreme value software. *Extremes*, 26:1–44.

Chavez-Demoulin, V. and Davison, A. C. (2005). Generalized additive modelling of sample extremes. *Journal of the Royal Statistical Society: Series C*, 54(1):207–222.

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

Coles, S. G. and Pericchi, L. R. (2003). Anticipating catastrophes through extreme value modelling. *Journal of the Royal Statistical Society: Series C*, 52(4):405–416.

Coles, S. G., Pericchi, L. R., and Sisson, S. (2003). A fully probabilistic approach to extreme rainfall modeling. *Journal of Hydrology*, 273(1-4):35–50.

Danielsson, J., de Haan, L., Peng, L., and de Vries, C. G. (2001). Using a bootstrap method to choose the sample fraction in tail index estimation. *Journal of Multivariate Analysis*, 76(2):226–248.

Danielsson, J., Ergun, L., de Haan, L., and de Vries, C. G. (2019). Tail index estimation: quantile-driven threshold selection. Staff Working Papers 19-28, Bank of Canada.

Davison, A. C. and Smith, R. L. (1990). Models for exceedances over high thresholds (with discussion). *Journal of the Royal Statistical Society: Series B*, 52(3):393–425.

Gomes, M. I. (1994). Penultimate behaviour of the extremes. *Extreme Value Theory and Applications: Proceedings of the Conference on Extreme Value Theory and Applications, Gaithersburg Maryland 1993*, 1:403–418.

Heffernan, J. E. and Tawn, J. A. (2004). A conditional approach for multivariate extreme values (with discussion). *Journal of the Royal Statistical Society Series B*, 66(3):497–546.

Murphy, C., Tawn, J. A., Varty, Z., and Towe, R. (2023). Software for threshold selection.

Naveau, P., Huser, R., Ribereau, P., and Hannart, A. (2016). Modeling jointly low, moderate, and heavy rainfall intensities without a threshold selection. *Water Resources Research*, 52(4):2753–2769.

Northrop, P. J. and Attalides, N. (2020). *threshr: Threshold Selection and Uncertainty for Extreme Value Analysis*. R package version 1.0.3.

Northrop, P. J., Attalides, N., and Jonathan, P. (2017). Cross-validatory extreme value

threshold selection and uncertainty with application to ocean storm severity. *Journal of the Royal Statistical Society: Series C*, 66(1):93–120.

Northrop, P. J. and Coleman, C. L. (2014). Improved threshold diagnostic plots for extreme value analyses. *Extremes*, 17(2):289–303.

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

Scarrott, C. and MacDonald, A. (2012). A review of extreme value threshold estimation and uncertainty quantification. *REVSTAT–Statistical Journal*, 10(1):33–60.

Smith, R. L. (2003). Statistics of extremes, with applications in environment, insurance, and finance. In *Extreme Values in Finance, Telecommunications, and the Environment*, edited by Finkenstadt, B. and Rootzén, H., pages 20–97. Chapman and Hall/CRC.

Varty, Z., Tawn, J. A., Atkinson, P. M., and Bierman, S. (2021). Inference for extreme earthquake magnitudes accounting for a time-varying measurement process. *arXiv:2102.00884*.

Wadsworth, J. L. (2016). Exploiting structure of maximum likelihood estimators for extreme value threshold selection. *Technometrics*, 58(1):116–126.

Wan, P. and Davis, R. A. (2019). Threshold selection for multivariate heavy-tailed data. *Extremes*, 22(1):131–166.