https://doi.org/10.5194/egusphere-2025-1138
Preprint. Discussion started: 6 May 2025

Automated tail-informed threshold selection for extreme coastal sea levels

Thomas P. Collings[1], Callum J. R. Murphy-Barltrop[2,3], Conor Murphy[4], Ivan D. Haigh[1,5], Paul D. Bates[1,6], and Niall D. Quinn[1]

[1]Fathom, Floor 2, Clifton Heights, Clifton, Bristol BS8 1EJ, UK
[2]Technische Universität Dresden, Institut Für Mathematische Stochastik, Dresden, Germany
[3]Center for Scalable Data Analytics and Artificial Intelligence (ScaDS.AI), Dresden/Leipzig, Germany
[4]School of Mathematical Sciences, Lancaster University, Lancaster, LA1 4YF, UK
[5]School of Ocean and Earth Science, University of Southampton, National Oceanography Centre, European Way, Southampton SO14 3ZH, UK
[6]School of Geographical Sciences, University of Bristol, Bristol BS8 1SS, UK

Correspondence: Thomas P. Collings (t.collings@fathom.global)

Abstract. Peaks over threshold (POT) techniques are commonly used in practice to model tail behaviour of univariate variables. The resulting models can be used to aid in risk assessments, providing estimates of relevant quantities such as return levels and periods. An important consideration during such modelling procedures involves the choice of threshold; this selection represents a bias-variance trade-off and is fundamental for ensuring reliable model fits. Despite the crucial nature of this problem, most applications of the POT framework select the threshold in an arbitrary manner and do not consider the sensitivity of the model to this choice. Recent works have called for a more robust approach for selecting thresholds, and a small number of automated methods have been proposed. However, these methods come with limitations, and currently, there does not appear to be a 'one size fits all' technique for threshold selection. In this work, we introduce a novel threshold selection approach that addresses some of the limitations of existing techniques. In particular, our approach ensures that the fitted model captures the tail behaviour at the most extreme observations, at the cost of some additional uncertainty. We apply our method to a global data set of coastal observations, where we illustrate the robustness of our approach and compare it to an existing threshold selection technique and an arbitrary threshold choice. Our novel approach is shown to select thresholds that are greater than the existing technique. We assess the resulting model fits using a right-sided Anderson-Darling test, and find that our method outperforms the existing and arbitrary methods on average. We present and discuss, in the context of uncertainty, the results from two tide gauge records; Apalachicola, US, and Fishguard, UK. In conclusion, the novel method proposed in this study improves the estimation of the tail behaviour of observed coastal water levels, and we encourage researchers from other disciplines to experiment using this method with their own data sets.

1 Introduction

Natural hazards such as flooding, earthquakes and wildfires devastate communities and livelihoods around the world. Extreme value analysis (EVA) applied to the historical records of such events provides a useful tool for describing the frequency and

https://doi.org/10.5194/egusphere-2025-1138
Preprint. Discussion started: 6 May 2025

intensity of these processes, and can be used by practitioners, community leaders, and engineers to prepare in advance for catastrophic events. Example applications include flood risk assessment (D'Arcy et al., 2023), nuclear regulation (Murphy-Barltrop and Wadsworth, 2024), ocean engineering (Jonathan et al., 2014), and structural design analysis (Coles and Tawn, 1994). Furthermore, stakeholders with assets spread across large geographical regions also utilise these tools to understand the hazard across regional, continental, and global scales; see, for instance, Keef et al. (2013), Quinn et al. (2019), and Wing et al. (2020).

Coastal flood events, driven by high tides, surges, or waves, are commonly recorded at tide gauge stations, which cover large proportions of the populated global coastline. When characterising extreme sea level events, these tide gauge records are a primary source of information available to coastal managers. Due to the large number of sites involved, automated techniques for the characterisation of extreme events are preferable.

The earliest EVA techniques used the annual maximum approach, whereby a theoretically motivated distribution is fitted to the observed yearly maxima. However, this approach suffers from the drawback that only one observation is recorded for each year, resulting in some extreme observations being disregarded. In practice, this can lead to an incomplete picture of the upper tail, and consequently, recent consensus has been to move away from the annual maximum approach (Pan and Rahman, 2022). As a result, the POT approach has become the most popular technique for EVA modelling; see Section 3 and Coles (2001) for further details. This approach involves fitting a statistical model to data above some high threshold. However, the choice of this threshold is not arbitrary, and inappropriate choices can result in poor model fits and extrapolation into the tail. Traditional approaches rely on visual assessments of parameter stability above the appropriate threshold. Such approaches suffer from subjectivity (Caballero-Megido et al., 2018) and the time input required to apply such techniques to global tide gauge records is not feasible. Consequently, many efforts have been made to reduce the time burden incurred by manual threshold selection. These include simplifications that allow large amounts of data to be processed, but at the cost of accuracy, e.g., using a static threshold, such as the 0.98 quantile or a fixed number of exceedances per year (Hiles et al., 2019; Collings et al., 2024). We refer to the approach of selecting a static 0.98 quantile across all sites (or variables) as the Q98 approach henceforth. Other approaches aim to automate much of the subjective decision-making process while retaining a flexible method that can capture the underlying behaviour of the physical processes (Solari et al., 2017; Curceac et al., 2020; Murphy et al., 2024).

In this study, our aim is to build upon existing techniques to provide a novel approach to automating threshold selection, which is applicable to a wide range of datasets whereby the extremes are characterised by different drivers. As a motivating example, we apply our method to a global dataset of 417 tide gauge records, demonstrating the performance of our approach over a variety of locations and benchmarking against other commonly used techniques.

The layout of this paper is as follows; in Section 2 we introduce the dataset used in this study and in Section 3 we discuss the common difficulties in using the POT approach across such a large, varied dataset, as well as some of the methods used to simplify the process. In Section 4, we describe our novel approach to automating threshold selection and explain the subjective choices we have made in the method. In Section 5, we present the results of applying our method to the global tide gauge dataset described in Section 2. In Section 6, we discuss our results in the context of uncertainty, bias, and the underlying physical processes and finally, in Section 7 we provide a conclusion to our study.

https://doi.org/10.5194/egusphere-2025-1138
Preprint. Discussion started: 6 May 2025

2 Data

The locations of the considered tide gauge stations are illustrated in Figure 1. These data are obtained from the Global Extreme Sea Level Analysis (GESLA) database (Haigh et al., 2023), version 3.1, which is a minor update to version 3 to include the most recent years (2022-2024). The GESLA database was collated from many organisations that collect and publish tide gauge data. The water level records are prepared using the quality control flags published by the authors alongside the data set, and duplicate timestamps in the records are also removed. The water level records that contain over 40 years of good data (defined as at least 75% complete) are retained. This results in a total of 417 water level records from around the world, which have an average record length of 66 years. The raw time series data are provided on a range of time steps (10, 15, and 60 minutes), and so are interpolated to hourly resolution. A linear trend is calculated and removed to account for mean sea level rise. Daily maxima data are obtained from the hourly records, and the data is subsequently declustered using a 4-day storm window to ensure event independence (Haigh et al., 2016; Sweet et al., 2020). Given the range of oceans and coastlines covered, one would generally expect to observe a wide variety of tail behaviours across the records.

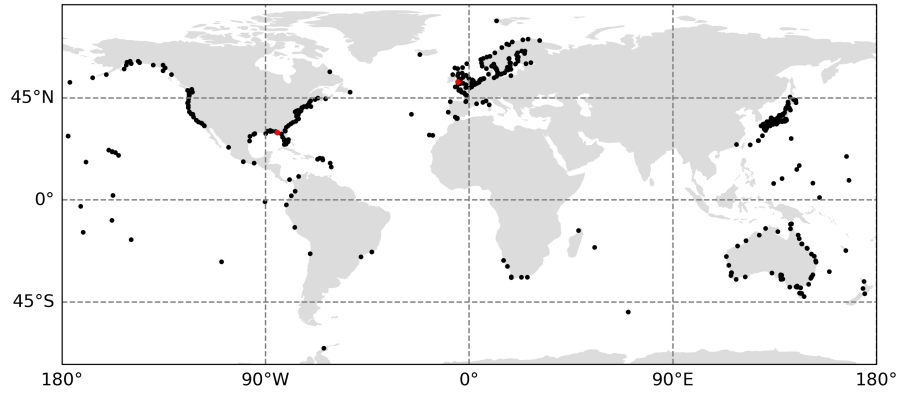

Figure 1. Map of GESLA record locations with record lengths greater than 40 years. The two locations highlighted in red are Apalachicola, US and Fishguard, UK, which are discussed in more detail in Section 5.3.

3 POT modelling

The POT approach, whereby a theoretically motivated distribution is fitted to the excesses of some high threshold (see, e.g., Coles, 2001), is the most common technique for assessing tail behaviour in environmental settings. Given any random variable X and a threshold u, the results of Balkema and de Haan (1974) and Pickands (1975) demonstrate that under weak conditions, the excess variable $Y := (X - u \mid X > u)$ can be approximated by a *generalised Pareto distribution* (GPD) – so long as the threshold u is 'sufficiently large'. The GPD has the form

$$H(y;\sigma,\xi) = 1 - \left(1 + \frac{\xi y}{\sigma}\right)_+^{-1/\xi}, \quad y > 0, \tag{1}$$

https://doi.org/10.5194/egusphere-2025-1138
Preprint. Discussion started: 6 May 2025

where $z_+ = \max(0, z)$, $\sigma > 0$, and ξ denotes any real number. We refer to σ and ξ as the scale and shape parameters, respectively, and we remark that the latter parameter quantifies important information about the form of tail phenomena; see Davison and Smith (1990) for further discussion. A wide range of statistical techniques have been proposed, including both Bayesian and frequentist frameworks, to fit the model in equation (1) (Dupuis, 1999; Behrens et al., 2004; Scarrott and MacDonald, 2012; Northrop et al., 2017), although we note that maximum likelihood estimation (MLE) remains the most common technique (e.g., Gomes and Guillou, 2015). Consequently, we restrict attention to MLE techniques throughout this paper.

In many practical contexts, equation (1) is used to obtain estimates of *return levels* for some *return period* N of interest. Such values offer a straightforward interpretation: the N-year return level is the value x_N that one would expect to exceed once, on average, every N years. Return levels are easily obtained by inverting equation (1) (see Coles, 2001), and their estimates are often used to inform decision making. For example, in the contexts of flood risk analysis and nuclear infrastructure design, regulators specify design levels corresponding to return periods of $N = 100$ years (D'Arcy et al., 2023) and $N = 10,000$ years (Murphy-Barltrop, 2024), respectively.

The ambiguity of the statement 'a sufficiently large threshold u' requires careful consideration. This is a problem that is commonly overlooked in many applications, and selecting a threshold u is entirely non-trivial. In particular, this selection represents a bias-variance trade-off: selecting a threshold too low will induce bias by including observations that do not represent tail behaviour, while extremely high thresholds will result in more variability due to lower sample sizes. Furthermore, the estimates of return levels are very sensitive to the choice of threshold, and biased estimates can significantly impact the cost and effectiveness of certain infrastructures, such as flood defences (Zhao et al., 2024).

Owing to the importance of threshold choice, a plethora of methods have been proposed which aim to balance the aforementioned trade-off; see Belzile (2024) for an extensive review of the literature. The standard and most-widely used approach for threshold selection involves a visual assessment of the stability of the GPD shape parameter across a range of increasing thresholds (Coles, 2001). This approach suffers from subjectivity in the choice of stable region. Furthermore, visual assessments for individual sites is simply not feasible (within a reasonable time scale) for large scale applications.

Automatic approaches seek to remove this subjectivity by selecting a threshold based on some criterion or goodness-of-fit metric; Wadsworth and Tawn (2012) and Northrop and Coleman (2014) utilise penultimate models and hypothesis testing; Bader et al. (2018) and Danielsson et al. (2019) use goodness-of-fit diagnostics; Wadsworth (2016) utilise a sequential assessment of a changepoint model; and Northrop et al. (2017) create a measure of predictive performance in a Bayesian framework. In the applied literature, Durocher et al. (2018) and Curceac et al. (2020) compare several automated goodness-of-fit approaches for selecting an appropriate threshold in the hydrological setting. Furthermore, Choulakian and Stephens (2001), Li et al. (2005) and Solari et al. (2017) automate goodness-of-fit procedures and apply these techniques to a range of precipitation and river flow data sets.

Recently, Murphy et al. (2024) proposed a novel threshold selection technique building on the work of Varty et al. (2021). This method, termed the *expected quantile discrepancy* (EQD), aims to select a threshold u for which the sample excesses are most consistent with a GPD model. We briefly outline this method below. Let $\boldsymbol{x}_u = (x_1, \ldots, x_{n_u})$ be the sample of excesses of some candidate threshold u, i.e., a sample from Y. For each candidate threshold, the EQD method assesses the expected

https://doi.org/10.5194/egusphere-2025-1138
Preprint. Discussion started: 6 May 2025

deviation between sample and theoretical quantiles at a set of fixed probabilities $\mathcal{P}_m := \{j/(m+1) : j = 1, \ldots, m\}$, where m denotes some large whole number. This assessment is done across a large number of bootstrapped samples, say B, to incorporate sampling variability and stablise the threshold choice. More specifically, letting $\boldsymbol{x}_u^b$ denote the b^{th} bootstrapped sample of $\boldsymbol{x}_u$, with $b = 1, \ldots, B$, Murphy et al. (2024) propose the metric

$$d_b(u) := \frac{1}{m} \sum_{j=1}^{m} \left| \frac{\hat{\sigma}_u^b}{\hat{\xi}_u^b} \left[\left(1 - \frac{j}{m+1} \right)^{-\hat{\xi}_u^b} - 1 \right] - Q\left(\frac{j}{m+1}; \boldsymbol{x}_u^b \right) \right|, \tag{2}$$

where $(\hat{\sigma}_u^b, \hat{\xi}_u^b)$ denote the GPD parameter estimates for $\boldsymbol{x}_u^b$, obtained using MLE, and $Q(j/(m+1); \boldsymbol{x}_u^b)$ denotes the $j/(m+1)$ empirical quantile of $\boldsymbol{x}_u^b$. Considering equation (2) over each bootstrapped sample, an overall measure of fit for u is given by $d(u) = \sum_{b=1}^{B} d_b(u)/B$. Finally, the selected threshold, u^*, is the value that minimises d, i.e., $u^* := \arg\min d(u)$. Through an extensive simulation study, alongside several case studies, Murphy et al. (2024) show that their approach convincingly outperforms the core existing approaches for threshold selection. Therefore, at the time of writing, the EQD technique is the best available approach for automating threshold selection.

In this article, we argue and demonstrate that while the EQD approach appears to work well in a wide variety of cases, it can suffer from drawbacks in certain contexts that result in less than ideal threshold choices. Specifically, the chosen thresholds can result in model fits that do not match up well at the most extreme observations. We briefly explore the reasons for why this may occur below.

To begin, consider two candidate thresholds $u_1 < u_2$ satisfying $\Pr(X > u_1) = 0.5$ (i.e., the median) and $\Pr(X > u_2) = 0.99$. Taking each threshold in turn, the EQD computes quantiles from the (bootstrapped) conditional variables $(X - u_1 \mid X > u_1)$ and $(X - u_2 \mid X > u_2)$ that correspond with the probability set $\mathcal{P}_m$. When considered on the scale of the data, however, this results in very different quantile probabilities. Letting $x_{u_1,j}$ denote the (true) $j/(m+1)$ quantile of $(X - u_1 \mid X > u_1)$ for any $j = 1, \ldots, m$, we have

$$\begin{aligned} \Pr(X \leq x_{u_1,j} + u_1) &= 1 - \Pr(X - u_1 > x_{u_1,j} \mid X > u_1) \Pr(X > u_1) \\ &= 1 - [1 - j/(m+1)]0.5 =: q_{u_1,j}, \end{aligned}$$

with an analogous formula following for u_2, i.e., $q_{u_2,j} := 1 - [1 - j/(m+1)]0.99$. The resulting probability sets $\{q_{u_1,j}\}_{j=1}^m$ and $\{q_{u_2,j}\}_{j=1}^m$, with $m = 100$, are illustrated in Figure 2. This demonstrates clearly that the lower the threshold level u, the lower the quantile probabilities evaluated by the EQD metric. Thus, quantiles lying far out the tail of the data will carry significantly less weight for lower thresholds than for higher thresholds.

On a similar note, we remark that the metric described in equation (2) is equally weighted across all probability levels. We argue that this somewhat disagrees with intuition in the sense that many practitioners mainly care about a models' ability to estimate very extreme return levels, and one only wants observations in the tail to be driving this estimation. Including non-extreme observations will bias the estimation procedure and therefore assessing quantile discrepancies mainly for lower

https://doi.org/10.5194/egusphere-2025-1138
Preprint. Discussion started: 6 May 2025

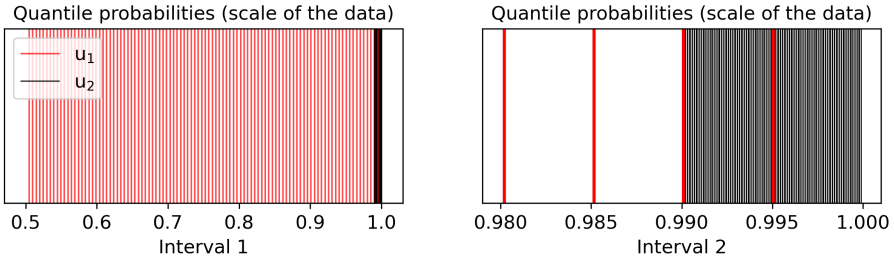

Figure 2. The probability sets $\{q_{u_1,j}\}_{j=1}^{m}$ and $\{q_{u_2,j}\}_{j=1}^{m}$ illustrated in red and black vertical lines, respectively. The left and right plots are given on different intervals to illustrate the fact the quantile probabilities exist in entirely different subregions of $[0, 1]$.

quantile levels, as will occur for lower candidate thresholds, provides little to no intuition as to how the fitted model will perform at the most extreme levels.

Taking these points into account, we propose an extension of the EQD procedure to improve the model fit to the most extreme observations. Our proposed extension results in models fits which more accurately capture the upper tail of the data in contexts where the EQD method struggles. Specifically, in the context of coastal modelling, we demonstrate that the EQD approach selects thresholds that do not appear appropriate for capturing the most extreme observations across many coastal sites; such issues do not arise for our extended approach. Consider the example illustrated in Figure 3 for a tide gauge record located in Penscola Bay, US, which is in the Gulf of Mexico. This record was selected as it is located in a region impacted by tropical cyclones, where the uncertainty in the model fits using the historical records is typically large. As demonstrated in the left panel of this figure, the model fit obtained using the EQD approach performs poorly within the upper tail. For this particular example, this indicates that the overall model fit is being driven mainly by lower observations, biasing the fit in the upper tail. Such findings were replicated across many coastal sites, indicating that this is not an unusual phenomenon. We also illustrate the model fit that arises from our proposed method (see Section 4) in the right panel of Figure 3. One can observe that even though the updated model fit has a higher disrepency value $d(u)$, the model quantiles appear better able to capture the upper tail in the data.

These findings indicate that whilst the EQD approach outperforms many existing techniques, it can, in some cases, result in model fits that fail to capture the most extreme observations. This drawback motivates novel developments, and in this work we propose an adaptation of the EQD technique, which we term the ***Ta**il-**i**nformed thresho**l**d **s**election methodology with quantile matching for extreme value modelling* (TAILS) approach. Unlike the EQD approach, our technique focuses exclusively on quantiles within a pre-defined upper tail of the data, independent of the choice of threshold. Furthermore, we demonstrate in Section 5 that TAILS results in improved model fits across a wide range of tide gauge records. Code for implementing the TAILS approach is freely available online at https://github.com/callumbarltrop/TAILS.

https://doi.org/10.5194/egusphere-2025-1138
Preprint. Discussion started: 6 May 2025

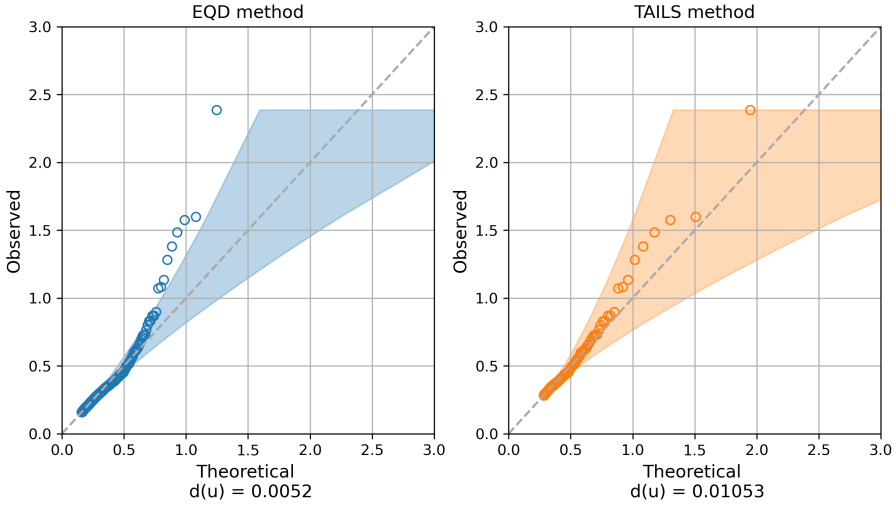

Figure 3. QQ plots for the thresholds selected using the EQD (left) and TAILS (right) approaches; see Section 4 for more details of the TAILS method. The sub captions in both cases gives the EQD score $d(u)$ at the threshold chosen by both methods.

4 The TAILS approach

In this section, we introduce the TAILS approach for GPD threshold selection. To begin, let $\mathscr{P} := \{p_i : i = 1, \ldots, m\}$ denote a set of increasing quantile levels close to 1: the selection of $\mathscr{P}$ is subsequently discussed. Given a candidate threshold u, let $\boldsymbol{x}_u^b$, $b = 1, \ldots, B$, be defined as in Section 3 and let $\pi_u = \Pr(X \leq u)$. We propose the following metric

$$\tilde{d}_b(u) := \frac{\sum_{i=1}^{m} \mathbb{1}(\pi_u < p_i) \left| \frac{\hat{\sigma}_u^b}{\hat{\xi}_u^b} \left[\left(\frac{1-p_i}{1-\pi_u} \right)^{-\hat{\xi}_u^b} - 1 \right] - Q\left(1 - \frac{1-p_j}{1-\pi_u}; \boldsymbol{x}_u^b\right) \right|}{\sum_{j=1}^{m} \mathbb{1}(\pi_u < p_j)}, \tag{3}$$

with $Q(\cdot\,;\cdot)$ and $(\hat{\sigma}_u^b, \hat{\xi}_u^b)$ defined as before. For each threshold u, this metric ensures that the same quantile probabilities are evaluated, when considered on the scale of the data. Furthermore, observe that equation (3) accounts for cases when the threshold probability, π_u, exceeds a subset of $\mathscr{P}$; in such instances, the metric is only evaluated on probabilities greater than the threshold non-exceedance probability, corresponding to the region where the given GPD model is valid. Analogous to the original approach, an overall measure of fit for a candidate threshold u is given by $\tilde{d}(u) = \sum_{b=1}^{B} \tilde{d}_b(u)/B$, and the selected threshold, u^*, is the value that minimises $\tilde{d}$, i.e., $u^* := \arg\min \tilde{d}(u)$.

The motivation behind (3) is to only evaluate quantile differences within the tail of the data, independent of the threshold candidate. This ensures that the threshold choice is driven entirely by the model fit within the most extreme observations. However, prior to applying the method, one must select a probability set $\mathscr{P}$. This choice is non-trivial, and is crucial for ensuring the proposed method selects a sensible threshold. For instance, selecting probabilities very close to one is meaningless

https://doi.org/10.5194/egusphere-2025-1138
Preprint. Discussion started: 6 May 2025

in a practical setting, since the corresponding quantiles cannot be estimated empirically from data of a finite sample size. On the other hand, selecting probabilities too low will defeat the objective of our proposed technique.

With this in mind, we term p_1 the *baseline probability*, i.e., the smallest probability in $\mathscr{P}$. This corresponds to the 'baseline' observation frequency below which one treats any events to be extreme relative to the sample size. Naturally, this represents a subjective choice, and the best choice of baseline probability is likely to be context dependent. In practice, we recommend selecting p_1 based on expert or domain-specific knowledge; for example, what magnitude of return period normally results in a relatively low-impact, but significant event within a given context? Take coastal flood risk mitigation and the occurrence of 'nuisance' flooding as an example. Nuisance flooding is defined as '*low levels of inundation that do not pose significant threats to public safety or cause major property damage, but can disrupt routine day-to-day activities, put added strain on infrastructure systems such as roadways and sewers, and cause minor property damage*' (Moftakhari et al., 2018). Although the exact return period of these events varies by location, a study carried out in the US demonstrated that these events generally occur at sub-annual frequencies, and that the median across their study sites was 0.5 years (Sweet et al., 2018). In this study, we chose to use a return period of 0.25 years for p_1, to include events below the median obtained in the study above. This choice was further supported by a sensitivity analysis, the results of which are presented in the Appendix. Note that this does not imply that the optimum threshold choice will lie close to the baseline event, since this choice is driven exclusively by the asymptotic rate of convergence to the underlying tail distribution.

Alongside the baseline probability, we also set p_m (the largest probability in $\mathscr{P}$), such that we ensure we observe 10 exceedances above the corresponding quantile, on average, over the observation period. Extrapolating beyond this level is unlikely to be meaningful, since we cannot estimate empirical quantiles outside of the range of data. Furthermore, we impose that all candidate thresholds (i.e., the values of u for which equation (3) is evaluated) are less than the 1 year return level. This upper threshold is used in similar automated threshold selection studies, such as Durocher et al. (2018).

Finally, for the remaining probabilities in $\mathscr{P}$, we set $p_j := p_1 + (j-1)(p_m - p_1)/(m-1)$, $j = 2, \ldots, m-1$, corresponding equally spaced values from the p_1 to p_m. For the number of quantile levels m, we follow Murphy et al. (2024) and set $m = 500$; such a value ensures a wide range of probabilities are evaluated without too much linear interpolation between observed quantile levels. Similar to Murphy et al. (2024), we found that the choice of m made very little difference to the thresholds selected by their approach. See the Appendix for more details.

5 Results

We now assess the performance of the TAILS approach using the dataset introduced in Section 2. In Section 5.1, we apply both the EQD and TAILS approaches over all locations with $m = 500$ and $B = 100$, to obtain thresholds above which we can consider an exceedance. The same values for m and B were used by Murphy et al. (2024). In Section 5.2 we assess with a right-sided Anderson-Darling (ADr) test the GPD model fits obtained using the selected thresholds from each approach, as well as the model fits using a static quantile threshold of the Q98. Lastly, in Section 5.3 we show the distance metrics from the

https://doi.org/10.5194/egusphere-2025-1138
Preprint. Discussion started: 6 May 2025

EQD and TAILS approaches for two tide gauge records, and present the resulting return levels from the two methods, as well as the results obtained using the Q98 as the threshold.

5.1 Selected thresholds

Since the scales of data at different locations vary, we present the selected threshold probabilities rather than the threshold magnitudes; these are illustrated in Figure 4. The TAILS approach clearly selects higher thresholds compared to the EQD approach, as expected. The lowest threshold selected by the TAILS and EQD methods is 0.903 and 0.501, respectively, and the highest threshold selected by the TAILS and EQD methods is 0.993 and 0.991, respectively. The lowest threshold selected by the EQD approach is very close to the lower limit, which was the median (i.e., 0.5).

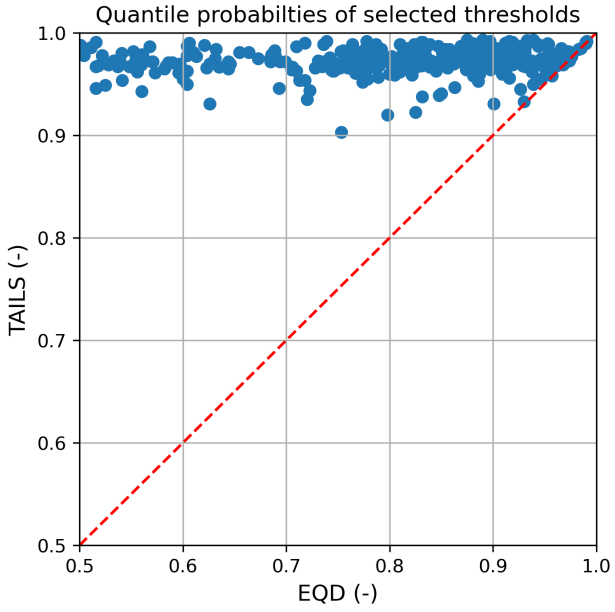

Figure 4. The results from applying the EQD and TAILS methods to every GESLA record used in this study, showing the quantile probability of the selected thresholds.

5.2 Right-sided Anderson-Darling test

The ADr test statistic (Sinclair et al., 1990; Solari et al., 2017) is used to measure the goodness-of-fit of the exceedances over the thresholds selected using both the EQD and TAILS methods, as well as the model fits computed using the Q98 approach. The test compares the theoretical quantiles against the empirical distribution, with more weight placed on the tails of the distribution (hence right-sided). The statistic quantifies the deviation of the data from the specified distribution. A p value is obtained by bootstrapping the test statistic, with p indicating the probability of observing such a deviation under the null hypothesis that

https://doi.org/10.5194/egusphere-2025-1138
Preprint. Discussion started: 6 May 2025

the threshold exceeding data cannot be modelled by a GPD. The null hypothesis is typically rejected for p values exceeding 0.05, corresponding to a 5% significance level.

A larger test statistic (equivalently, a lower p-value) indicates more deviation from the model distribution being tested, which in this case, is a GPD. As shown in Figure 5 a, the EQD approach yields larger ADr test statistics than the TAILS method. The range of test statistics computed using the TAILS method are all less than 1, whereas the EQD approach has many values exceeding 1. This indicates the EQD method could be selecting a threshold over which the exceedances are not well characterised by a GPD. This is further corroborated by the p-values obtained for each method, plotted in Figure 5 b. The median p-value across all model fits obtained using the TAILS method is 0.615, compared with 0.312 for the EQD approach. The TAILS method also outperforms the Q98 approach, with a smaller test statistic average and greater average p-value. While all the methods achieve adequate fits for most of the dataset, in some of the cases where the EQD and Q98 method lead to poor model fits (p-value less than 0.05), the TAILS method can significantly improve results. Of the 417 tide gauge records that were assessed, 89 records had an ADr p-value of less than 0.05 when using the EQD method. By comparison, using the TAILS approach, we obtain only 17 model fits with ADr p-values less than 0.05.

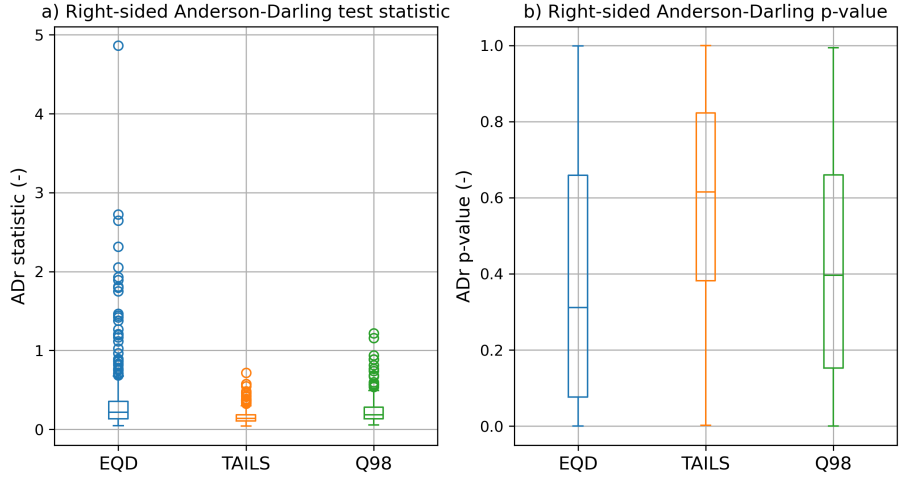

Figure 5. Box and whisker plots showing the results from applying an ADr test to all the exceedances over the thresholds selected using the EQD and TAILS approaches, as well as using a static Q98 threshold.

5.3 Distance metrics and return levels

As a further illustration, consider the model fits for two sites; Apalachicola in the US and Fishguard in the UK, both shown in Figure 6. The two sites have been selected based on the differences in geographic location and the associated extreme water level drivers, which lead to contrasting return level estimates. Apalachicola, located on the western coast of Florida in the Gulf of Mexico, is subjected to violent tropical cyclones which drive huge storm surges due to the large and shallow continental shelf (Chen et al., 2008; Zachry et al., 2015). The GPD model fit that characterises the return levels of the water

https://doi.org/10.5194/egusphere-2025-1138
Preprint. Discussion started: 6 May 2025

level record therefore has a large positive shape parameter, which displays a steep and exponentially increasing return period curve. In contrast, Fishguard is located on the southern side of Cardigan Bay, near the inlet of the Irish Sea. The events driving extreme sea levels in this location are a combination of strong extratropical storms and astronomical tidal variation, which are characterised by a wholly different return period curve (Amin, 1982; Olbert and Hartnett, 2010). The GPD model fit for this record has a negative shape parameter, which means that the return levels plateau as the return period increases.

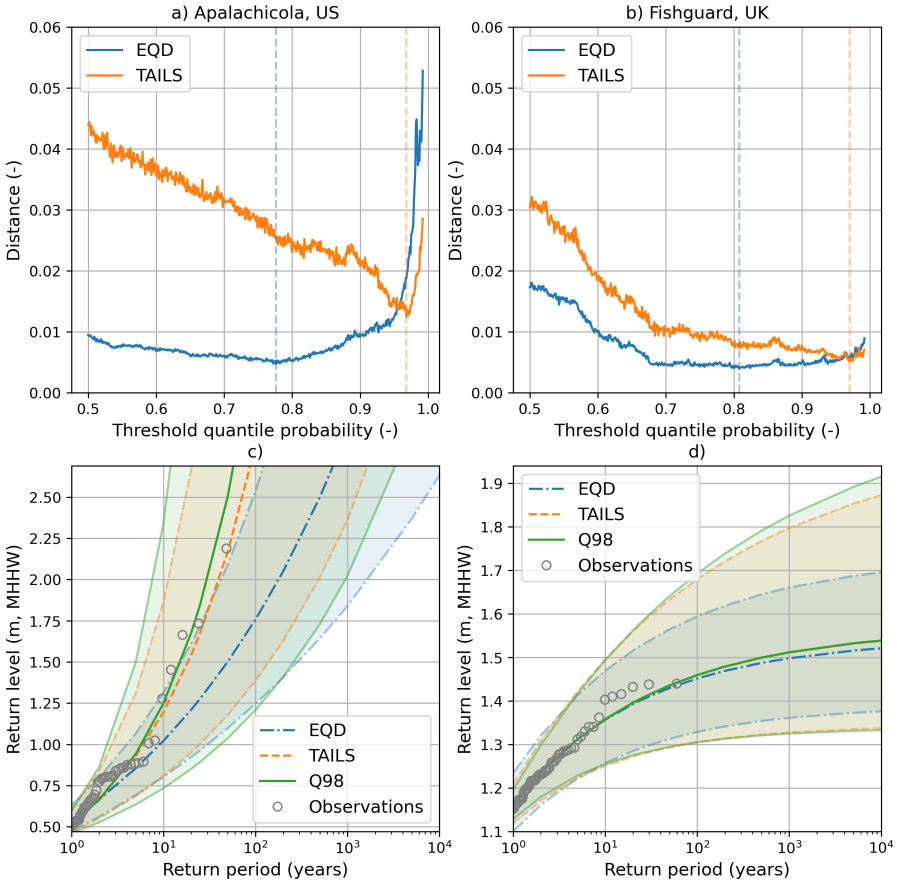

Figure 6. Model fits for two locations. Left column: Apalachicola, US (a and c). Right column: Fishguard, UK (b and d). The top row (a and b) shows the TAILS and EQD distance metrics, plotted as a function of the threshold probability. The vertical dashed lines indicate the distance minima, and therefore the selected threshold quantile probability. The bottom row (c and d) displays the return level plots for both methods, alongside the empirical plot and model fit obtained by using the Q98 approach. The shaded areas indicate the 95% confidence interval, calculated using bootstrapping of the GPD model parameters.

In the top row of Figure 6 (panels a and b), one can observe the EQD and TAILS distances metrics (i.e., equations (2) and (3)) plotted as a function of the threshold probability for both tide gauge records. Clearly the global minimums of both approaches are starkly different, representing the different quantile estimates evaluated by either approach. Panels c and d of Figure 6 show the estimated return levels and 95% confidence intervals from each of the TAILS, EQD and Q98 methods, at Apalachicola and Fishguard, respectively.

https://doi.org/10.5194/egusphere-2025-1138
Preprint. Discussion started: 6 May 2025

In the case of Apalachicola, the minimum distance (panel a) obtained using the TAILS method (0.012) is greater than double the minimum distance obtained using the EQD approach (0.005). Compare this with the return level estimates from each of the 3 methods presented (panel c). Despite having a larger minimum distance, the TAILS approach captures the empirical observations much better than the EQD method. In fact, four of the historical events even lie outside of the 95% confidence interval for the EQD method, highlighting the need for the TAILS method.

Contrast this with the results from Fishguard (panel b), where the minimum distances obtained using each approach are much more comparable; 0.005 for TAILS and 0.004 for the EQD approach. The resulting return level estimates (panel d) are also similar, with very small differences in the mean return levels between each of the three methods. The key difference observed in panel d is the uncertainty bounds, with the EQD method having better constrained uncertainty in the higher return periods than the other two methods.

6 Discussion

In this work, we have introduced an automated threshold selection technique that addresses certain limitations of the leading existing approach. Using a global tide gauge dataset, both methods are rigorously compared in Section 5 alongside a commonly used static threshold. We demonstrate that in many cases, the TAILS approach better captures the most extreme observations compared to the EQD technique, and outperforms the static Q98 threshold when assessed using an ADr test.

The TAILS method guarantees that the resulting model fits will be driven by data observed in the tail, which is desirable for practical applications where estimation of extreme quantities (e.g., return levels) is required. We also believe that calibrating threshold selection to focus on the tail will encourage more practitioners to adopt our approach, since we are more likely to obtain a model fit that accurately captures the tail behaviour.

However, focusing on model fits within the tail comes at the cost of additional uncertainty, since by definition, less data is available for inference. Since uncertainty quantification is a key focus of the approach proposed by Murphy et al. (2024), the EQD technique will generally offer lower model uncertainty compared to TAILS. In some applications, this may be more desirable than capturing the most extreme observations. Thus, when deciding whether to use EQD or TAILS, one must consider the following question: is it more important that the model is more certain and robust, or that the model better captures the most extreme observations? We recommend that practitioners consider this question within the context of their application before selecting a technique.

For the application demonstrated in this paper, acknowledging and embracing uncertainty is key for any practitioner. Take the example of Apalachicola, US given in Section 5.3. This region is impacted by tropical cyclones, making the return level estimates made from the historical record very uncertain. To illustrate this point, two major Category 4 hurricanes (Helene and Milton) made landfall on the west coast of Florida in September and October 2024, after the GESLA 3.1 update was collated. Preliminary data recorded during the event suggest that Hurricane Helene broke the highest recorded water levels at three tide gauges located in Florida, and Hurricane Milton set the second highest water level ever recorded at the tide gauge located in Fort Myers, US (Powell, 2024a, b). Fitting distributions to these records pre and post these events would likely result in

https://doi.org/10.5194/egusphere-2025-1138
Preprint. Discussion started: 6 May 2025

different mean return levels being estimated, especially when considering the most extreme return periods (e.g., the 1 in 500 year event). We tested this and found that, when using the TAILS approach, the mean return level for the 1 in 500 year event increased by 55 cm if the tide gauge record is extended beyond the GESLA 3.1 update, to include these events. By recognising the uncertainty in the underlying processes and the uncertainty inherent in the estimates made from observations, we can be more confident that our models will be able to capture extreme events which are yet to occur.

Future work could include a variable baseline event, which is linked to the underlying forcing mechanisms in an area. As discussed in Section 5.3, tide gauges around the world are characterised by different patterns of extreme water levels. It might be possible to link a dominant forcing type to the baseline event, which could further improve the ability of TAILS to capture the tail behaviour in the estimated return levels. Another direction of future work could be to extend the method to include non-stationary data by allowing the GPD parameters to be functions of time or covariates (e.g., Eastoe and Tawn, 2009; Youngman, 2019). Relevant covariates are those that impact the number of extreme events that occur within a given year; for example, indices related to the ENSO and NAO phenomena, which affect the likelihood of temperature and precipitation extremes (Dong et al., 2019), can be incorporated into the POT modelling framework. Continuing to develop automated threshold selection approaches to suit a wide range of different data structures represents an important line of future research.

While results may indicate in certain examples that the Q98 approach outperforms the EQD, the benefits of a data-driven approach can not be understated. When relying on TAILS or the EQD, not only is the threshold justified by a goodness-of-fit measure but sampling variability has also been taken into account. This leads to a well-justified threshold choice and an easier characterisation of the uncertainty in resulting estimates. It also allows for the uncertainty in the threshold choice to be incorporated when making inference; see Murphy et al. (2024). Furthermore, when applying methods to a large number of sites, employing an automated procedure avoids the need for manual checks on individual threshold choices.

Finally, we note that the selection of the probability set $\mathcal{P}$ is non-trivial, as discussed in Section 4. We therefore recommend that practitioners experiment with both the baseline and maximal probabilities to assess whether such values have a practical effect on the resulting model, using diagnostics such as QQ and return level plots to guide this procedure. The code has been written in such a way as to make it easily parallelised, allowing for fast testing of multiple baseline and maximal probabilities across a variety of datasets. We encourage and invite fellow researchers to utilise this method on other perils, such as rainfall or river flow measurements.

7 Conclusions

Accurately estimating the extreme tail behaviour of historical observations is of great importance to researchers and practitioners working in natural hazards. POT methods are regularly used in these fields for this purpose, but selecting the threshold above which to consider an exceedance requires careful consideration. In this paper, we present TAILS, a new method for automating the threshold selection process building upon the recently published EQD method (Murphy et al., 2024).

We apply two key innovations to improve upon the EQD method in the context of extreme coastal sea levels. Firstly, we fix the quantiles that we consider when computing the distance metrics. This avoids oversampling the most extreme quantiles

https://doi.org/10.5194/egusphere-2025-1138
Preprint. Discussion started: 6 May 2025

when assessing higher thresholds. Secondly, we limit the quantiles considered for our distance metric to be only above a predetermined baseline probability. This means that when optimising the distance metric to select a threshold, we are only considering quantiles that we deem to be extreme, and hence worth considering when selecting a threshold. In this study, the baseline probability was decided using the literature and a sensitivity test.

We show that the TAILS approach selects, on average, higher thresholds than the EQD method. When the resulting model fits are evaluated using an ADr test against the EQD method and the Q98 method, we show that the TAILS method outperforms both with respect to the ADr test statistic and the p-value. We also illustrate that the TAILS method typically results in larger uncertainty bounds, but argue that when considering water level records located in regions that experience tropical cyclones, this is positive.

Although a large number of records are assessed, this study is limited in scope as it only considers tide gauge records. We hope that the method can be widely used to better estimate the intensities and frequencies of other natural hazards. The code has been written in such a way as to make it easily accessible and easily parallelised so as to encourage uptake from fellow researchers.

Code and data availability. The code for implementing the TAILS approach is freely available online at https://github.com/callumbarltrop/TAILS, along with an example data set.

The GESLA 3 tide gauge database is available at https://doi.org/10.5285/d21a496a-a48f-1f21-e053-6c86abc08512 (Haigh et al., 2023)

Appendix A: Sensitivity test of baseline probability, p_1

A range of baseline probabilities were tested across the whole dataset, and the resulting threshold and model fits were used to calculate a right-sided Anderson-Darling (ADr) test statistic and the p-value (Sinclair et al., 1990; Solari et al., 2017). For more details on the ADr test, see the main text. The return periods that were tested for the baseline probabilities were 0.083, 0.167, 0.25, 0.33, 0.5, 0.667, and 1.0 years. These equivalate to the 1 in 1, 2, 3, 4, 6, 8 and 12 month events.

The results of this sensitivity test are shown in Figure A1. Panel a presents the ADr test statistic for the 7 return periods tested. When looking at the median and interquartile ranges of the ADr test statistics, the threshold selection looks relatively insensitive to the return period chosen, with very little differences between the 0.167, 0.25, 0.333, and 0.5 year return periods. When considering the ADr test p-value (panel b), there is also only small differences between the 0.167, 0.25, 0.33 and 0.5 year return periods. We take this, along with the value obtain from the literature (presented in main text), as evidence that any one of these values would suffice as the baseline probability, p_1.

https://doi.org/10.5194/egusphere-2025-1138
Preprint. Discussion started: 6 May 2025

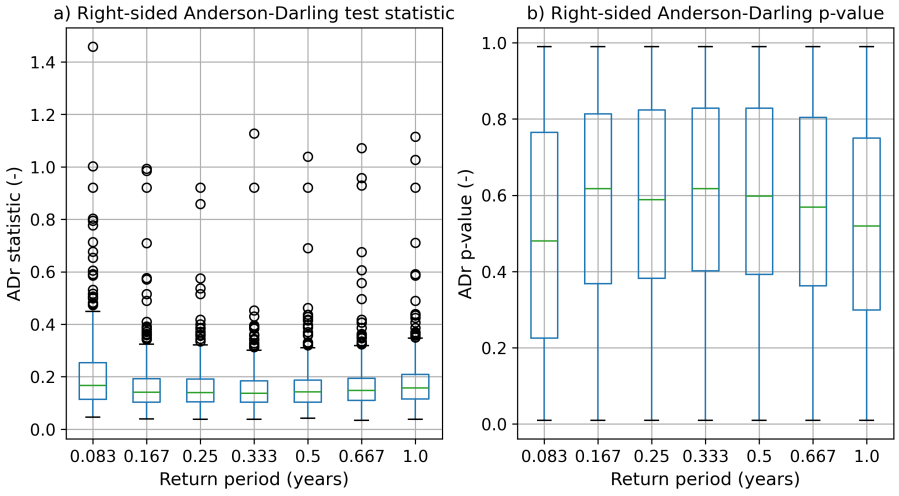

Figure A1. The results from the sensitivity test of different baseline probabilities.

Appendix B: Sensitivity test of number of quantile levels, m

Following Murphy et al. (2024), a sensitivity test to the number of quantile levels, m was carried out. The values of m tested were 10, 50, 100, 200, 500, 1000 and 'n_exceedances', which is equal to the number of exceedances over the baseline probability for each tide gauge record. The range m values that are used by the 'n_exceedances' are shown below in Figure B1. The full range spreads between 161 to 811, and the median is centred on 231.

https://doi.org/10.5194/egusphere-2025-1138
Preprint. Discussion started: 6 May 2025

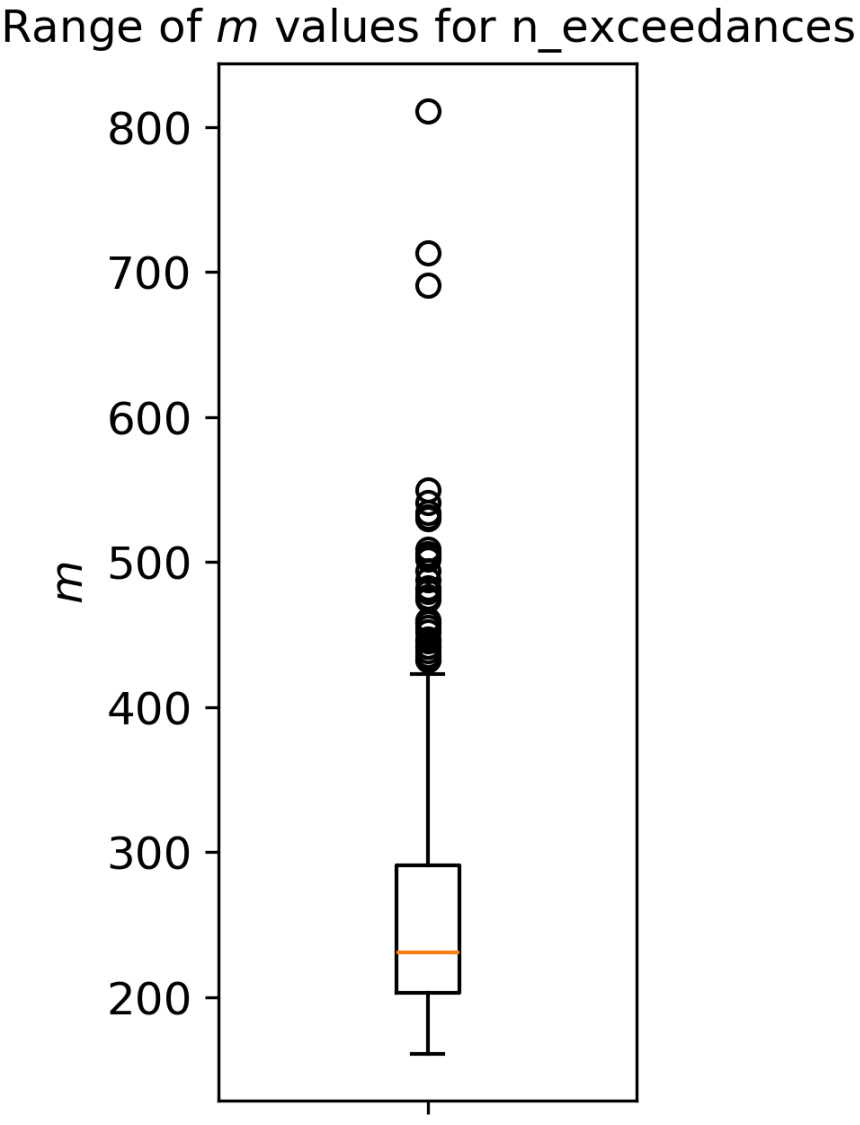

Figure B1. The range of m values used by 'n_exceedances', which is equal to the number of exceedances over the baseline probability for each tide gauge record.

The results of this sensitivity analysis are presented in Figure B2, showing that the method is quite insensitive to the m value used. This is similar to the findings of Murphy et al. (2024). We recommend using any value over 10, and choose to use $m = 500$ in this study for consistency with Murphy et al. (2024).

https://doi.org/10.5194/egusphere-2025-1138
Preprint. Discussion started: 6 May 2025

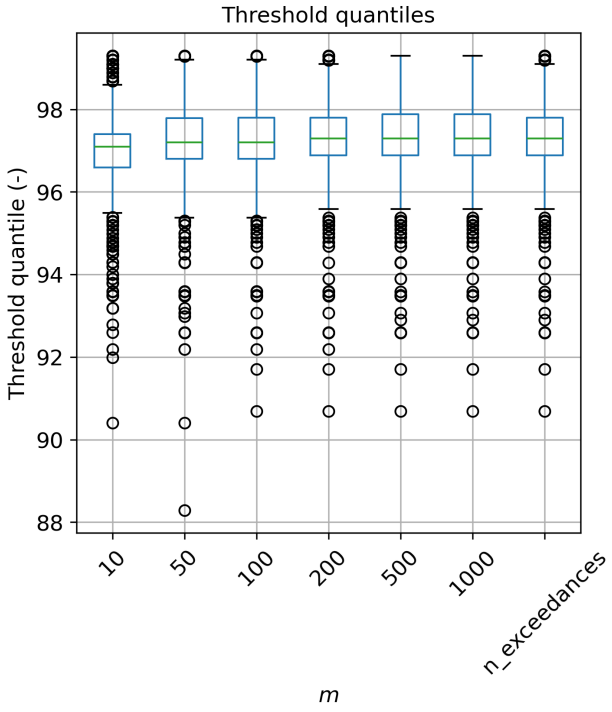

Figure B2. The results of the sensitivity test using different m values. 'n_exceedances' refers to the number of exceedances over the baseline probability, at each tide gauge record.

Author contributions. TPC was responsible for making the edits to the original code, preparing the data for use in the study, validating the results and drafting the introduction, data, results, discussion and conclusion sections of the manuscript, as well as reviewing the manuscript. CJRMB provided guidance and expertise in interpreting the results, helped with ideas on how to improve the original method, and contributed to the manuscript, including writing parts of the abstract, POT modelling, the TAILS approach and discussion sections, as well as carrying out thorough reviews. CM kindly provided the original code that underlies the EQD method, and has contributed to the manuscript by writing parts of the introduction, POT modelling and the TAILS approach sections, whilst also helping with thorough reviews of the manuscript throughout. IDH provided the GESLA 3.1 update, initial guidance, and ideas about how to start, as well as reviewing the manuscript. PDB provided guidance and also reviewed the manuscript. NDQ provided guidance, advice, and support throughout the study, offering insight in interpreting the results and ideas on how to proceed, as well as reviewing the manuscript.

Competing interests. The contact author has declared that none of the authors have any competing interests.

https://doi.org/10.5194/egusphere-2025-1138
Preprint. Discussion started: 6 May 2025

References

Amin, M.: On analysis and forecasting of surges on the west coast of Great Britain, Geophysical Journal International, 68, 79–94, https://doi.org/10.1111/j.1365-246X.1982.tb06963.x, 1982.

Bader, B., Yan, J., and Zhang, X.: Automated threshold selection for extreme value analysis via ordered goodness-of-fit tests with adjustment for false discovery rate, Annals of Applied Statistics, 12, 310–329, https://doi.org/10.1214/17-AOAS1092, 2018.

Balkema, A. A. and de Haan, L.: Residual Life Time at Great Age, The Annals of Probability, 2, 792–804, https://doi.org/10.1214/aop/1176996548, 1974.

Behrens, C. N., Lopes, H. F., and Gamerman, D.: Bayesian analysis of extreme events with threshold estimation, Statistical modelling, 4, 227–244, https://doi.org/10.1191/1471082X04st075oa, 2004.

Belzile, Léo R., H. A. Y. R.: A utopic adventure in the modelling of conditional univariate and multivariate extremes, Extremes, pp. 1–35, https://doi.org/10.1007/s10687-024-00493-1, 2024.

Caballero-Megido, C., Hillier, J., Wyncoll, D., Bosher, L., and Gouldby, B.: Technical note: comparison of methods for threshold selection for extreme sea levels, Journal of Flood Risk Management, 11, 127–140, https://doi.org/10.1111/jfr3.12296, 2018.

Chen, Q., Wang, L., and Tawes, R.: Hydrodynamic response of northeastern Gulf of Mexico to hurricanes, Estuaries and Coasts, 31, 1098–1116, https://doi.org/10.1007/s12237-008-9089-9, 2008.

Choulakian, V. and Stephens, M. A.: Goodness-of-fit tests for the generalized Pareto distribution, Technometrics, 43, 478–484, https://doi.org/10.1016/j.cam.2019.04.018, 2001.

Coles, S.: An Introduction to Statistical Modeling of Extreme Values, Springer London, ISBN 978-1-84996-874-4, https://doi.org/10.1007/978-1-4471-3675-0, 2001.

Coles, S. G. and Tawn, J. A.: Statistical Methods for Multivariate Extremes: An Application to Structural Design, Applied Statistics, 43, 1–48, https://doi.org/10.2307/2986112, 1994.

Collings, T. P., Quinn, N. D., Haigh, I. D., Green, J., Probyn, I., Wilkinson, H., Muis, S., Sweet, W. V., and Bates, P. D.: Global application of a regional frequency analysis to extreme sea levels, Natural Hazards and Earth System Sciences, 24, 2403–2423, https://doi.org/10.5194/nhess-24-2403-2024, 2024.

Curceac, S., Atkinson, P. M., Milne, A., Wu, L., and Harris, P.: An evaluation of automated GPD threshold selection methods for hydrological extremes across different scales, Journal of Hydrology, 585, https://doi.org/10.1016/J.JHYDROL.2020.124845, 2020.

Danielsson, J., Ergun, L., de Haan, L., and de Vries, C. G.: Tail index estimation: quantile-driven threshold selection, Staff Working Papers 19-28, Bank of Canada, https://ideas.repec.org/p/bca/bocawp/19-28.html, accessed: 10/03/2025, 2019.

Davison, A. C. and Smith, R. L.: Models for Exceedances Over High Thresholds, Journal of the Royal Statistical Society. Series B: Statistical Methodology, 52, 393–425, https://doi.org/10.1111/j.2517-6161.1990.tb01796.x, 1990.

Dong, X., Zhang, S., Zhou, J., Cao, J., Jiao, L., Zhang, Z., and Liu, Y.: Magnitude and frequency of temperature and precipitation extremes and the associated atmospheric circulation patterns in the Yellow River basin (1960–2017), China, Water, 11, 2334, https://doi.org/10.3390/w11112334, 2019.

Dupuis, D.: Exceedances over High Thresholds: A Guide to Threshold Selection, Extremes, 1, 251–261, https://doi.org/10.1023/A:1009914915709, 1999.

Durocher, M., Mostofi Zadeh, S., Burn, D. H., and Ashkar, F.: Comparison of automatic procedures for selecting flood peaks over threshold based on goodness-of-fit tests, Hydrological processes, 32, 2874–2887, https://doi.org/10.1002/hyp.13223, 2018.

https://doi.org/10.5194/egusphere-2025-1138
Preprint. Discussion started: 6 May 2025

D'Arcy, E., Tawn, J. A., Joly, A., and Sifnioti, D. E.: Accounting for seasonality in extreme sea-level estimation, The Annals of Applied Statistics, 17, 3500–3525, https://doi.org/10.1214/23-AOAS1773, 2023.

Eastoe, E. F. and Tawn, J. A.: Modelling non-stationary extremes with application to surface level ozone, Journal of the Royal Statistical Society. Series C: Applied Statistics, 58, 25–45, https://doi.org/10.1111/j.1467-9876.2008.00638.x, 2009.

Gomes, M. I. and Guillou, A.: Extreme Value Theory and Statistics of Univariate Extremes: A Review, International Statistical Review, 83, 263–292, https://doi.org/10.1111/INSR.12058, 2015.

Haigh, I. D., Wadey, M. P., Wahl, T., Ozsoy, O., Nicholls, R. J., Brown, J. M., Horsburgh, K., and Gouldby, B.: Spatial and temporal analysis of extreme sea level and storm surge events around the coastline of the UK, Scientific data, 3, 1–14, https://doi.org/10.1038/sdata.2016.107, 2016.

Haigh, I. D., Marcos, M., Talke, S. A., Woodworth, P. L., Hunter, J. R., Hague, B. S., Arns, A., Bradshaw, E., and Thompson, P.: GESLA Version 3: A major update to the global higher-frequency sea-level dataset, Geoscience Data Journal, 10, 293–314, https://doi.org/10.1002/gdj3.174, 2023.

Hiles, C. E., Robertson, B., and Buckham, B. J.: Extreme wave statistical methods and implications for coastal analyses, Estuarine, Coastal and Shelf Science, 223, 50–60, https://doi.org/10.1016/j.ecss.2019.04.010, 2019.

Jonathan, P., Ewans, K., and Flynn, J.: On the estimation of ocean engineering design contours, Journal of Offshore Mechanics and Arctic Engineering, 136, 1–8, https://doi.org/10.1115/1.4027645, 2014.

Keef, C., Tawn, J. A., and Lamb, R.: Estimating the probability of widespread flood events, Environmetrics, 24, 13–21, https://doi.org/10.1002/env.2190, 2013.

Li, Y., Cai, W., and Campbell, E.: Statistical modeling of extreme rainfall in southwest Western Australia, Journal of climate, 18, 852–863, https://doi.org/10.1175/JCLI-3296.1, 2005.

Moftakhari, H. R., AghaKouchak, A., Sanders, B. F., Allaire, M., and Matthew, R. A.: What Is Nuisance Flooding? Defining and Monitoring an Emerging Challenge, Water Resources Research, 54, 4218–4227, https://doi.org/10.1029/2018WR022828, 2018.

Murphy, C., Tawn, J. A., and Varty, Z.: Automated threshold selection and associated inference uncertainty for univariate extremes, arXiv, 2310.17999, http://arxiv.org/abs/2310.17999, 2024.

Murphy-Barltrop, C.: ONR-RRR-079, Tech. rep., Office for Nuclear Regulation, 2024.

Murphy-Barltrop, C. and Wadsworth, J.: Modelling non-stationarity in asymptotically independent extremes, Computational Statistics & Data Analysis, 199, 108 025, https://doi.org/10.1016/j.csda.2024.108025, 2024.

Northrop, P. J. and Coleman, C. L.: Improved threshold diagnostic plots for extreme value analyses, Extremes, 17, 289–303, https://doi.org/10.1007/s10687-014-0183-z, 2014.

Northrop, P. J., Attalides, N., and Jonathan, P.: Cross-validatory extreme value threshold selection and uncertainty with application to ocean storm severity, Journal of the Royal Statistical Society. Series C: Applied Statistics, 66, 93–120, https://doi.org/10.1111/RSSC.12159, 2017.

Olbert, A. I. and Hartnett, M.: Storms and surges in Irish coastal waters, Ocean Modelling, 34, 50–62, https://doi.org/https://doi.org/10.1016/j.ocemod.2010.04.004, 2010.

Pan, X. and Rahman, A.: Comparison of annual maximum and peaks-over-threshold methods with automated threshold selection in flood frequency analysis: a case study for Australia, Natural Hazards, pp. 1–26, https://doi.org/10.1007/s11069-021-05092-y, 2022.

Pickands, J.: Statistical Inference Using Extreme Order Statistics, The Annals of Statistics, 3, 119–131, https://doi.org/10.1214/aos/1176343003, 1975.

https://doi.org/10.5194/egusphere-2025-1138
Preprint. Discussion started: 6 May 2025

Powell, E.: Hurricane Helene Post-Storm Summary Report, https://climatecenter.fsu.edu/images/docs/Hurricane-Helene-Summary-Report.pdf, accessed: 07/01/2025, 2024a.

Powell, E.: Post-Storm Summary Report on Hurricane Milton, https://climatecenter.fsu.edu/images/docs/Hurricane-Helene-Summary-Report.pdf, https://climatecenter.fsu.edu/images/docs/Hurricane-Helene-Summary-Report.pdf, accessed: 07/01/2025, 2024b.

Quinn, N., Bates, P. D., Neal, J., Smith, A., Wing, O., Sampson, C., Smith, J., and Heffernan, J.: The Spatial Dependence of Flood Hazard and Risk in the United States, Water Resources Research, 55, 1890–1911, https://doi.org/10.1029/2018WR024205, 2019.

Scarrott, C. and MacDonald, A.: A review of extreme value threshold estimation and uncertainty quantification, Revstat Statistical Journal, 10, 33–60, 2012.

Sinclair, C., Spurr, B., and Ahmad, M.: Modified anderson darling test, Communications in Statistics - Theory and Methods, 19, 3677–3686, https://doi.org/10.1080/03610929008830405, 1990.

Solari, S., Egüen, M., Polo, M. J., and Losada, M. A.: Peaks Over Threshold (POT): A methodology for automatic threshold estimation using goodness of fit p-value, Water Resources Research, 53, 2833–2849, https://doi.org/10.1002/2016WR019426, 2017.

Sweet, W. V., Dusek, G., Obeysekera, J., and Marra, J. J.: Patterns and Projections of High Tide Flooding Along the U.S. Coastline Using a Common Impact Threshold, Tech. rep., National Oceanic and Atmospheric Administration, https://www.tidesandcurrents.noaa.gov/publications/techrpt86_PaP_of_HTFlooding.pdf, accessed: 10/03/2025, 2018.

Sweet, W. V., Genz, A. S., Obeysekera, J., and Marra, J. J.: A regional frequency analysis of tide gauges to assess Pacific coast flood risk, Frontiers in Marine Science, 7, 1–15, https://doi.org/10.3389/fmars.2020.581769, 2020.

Varty, Z., Tawn, J. A., Atkinson, P. M., and Bierman, S.: Inference for extreme earthquake magnitudes accounting for a time-varying measurement process, arXiv, 2102.00884, http://arxiv.org/abs/2102.00884, 2021.

Wadsworth, J. L.: Exploiting structure of maximum likelihood estimators for extreme value threshold selection, Technometrics, 58, 116–126, https://doi.org/10.1080/00401706.2014.998345, 2016.

Wadsworth, J. L. and Tawn, J. A.: Likelihood-based procedures for threshold diagnostics and uncertainty in extreme value modelling, Journal of the Royal Statistical Society: Series B, 74, 543–567, https://doi.org/10.1111/j.1467-9868.2011.01017.x, 2012.

Wing, O. E., Quinn, N., Bates, P. D., Neal, J. C., Smith, A. M., Sampson, C. C., Coxon, G., Yamazaki, D., Sutanudjaja, E. H., and Alfieri, L.: Toward Global Stochastic River Flood Modeling, Water Resources Research, 56, https://doi.org/10.1029/2020WR027692, 2020.

Youngman, B. D.: Generalized Additive Models for Exceedances of High Thresholds With an Application to Return Level Estimation for U.S. Wind Gusts, Journal of the American Statistical Association, 114, 1865–1879, https://doi.org/10.1080/01621459.2018.1529596, 2019.

Zachry, B. C., Booth, W. J., Rhome, J. R., and Sharon, T. M.: A National View of Storm Surge Risk and Inundation, Weather, Climate, and Society, 7, 109 – 117, https://doi.org/10.1175/WCAS-D-14-00049.1, 2015.

Zhao, F., Lange, S., Goswami, B., and Frieler, K.: Frequency bias causes overestimation of climate change impacts on global flood occurrence, Geophysical Research Letters, 51, e2024GL108 855, https://doi.org/10.1029/2024GL108855, 2024.