Response to review of '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)

We would like to thank both reviewers for providing detailed and constructive feedback. The comments provided have helped to improve the quality and readability of our article. In the below text, we respond to each comment in turn, highlighting any novel additions in the text.

Reviewer 1

This study proposes a novel automated threshold selection method for modeling extreme coastal sea levels within the Peaks Over Threshold (POT) framework, aiming to better capture tail behavior while addressing the limitations of arbitrary and existing automated threshold choices. The method is applied to global tide gauge data and evaluated using the Anderson-Darling test, demonstrating improved performance over conventional techniques. However, the quality and readability of Figure 4 should be enhanced to ensure clearer communication of results. The discussion section is relatively weak, lacking depth, logical structure, and clarity. It is recommended that the discussion be expanded and subdivided to include specific commentary on the data, methodology, and results of this study, with comparative insights drawn from previous research to highlight the strengths and limitations of the proposed approach. The authors are also encouraged to include a forward-looking perspective outlining directions for future work. In summary, I recommend a major revision.

We thank the reviewer for this useful feedback. We have made several additions to the article to account for these comments:

1. We have removed Figure 4 and replaced it with an alternative histogram figure (see below), alongside detailed spatial figures illustrating the increases in selected threshold levels across different locations. This figure makes it clearer how the use of the TAILS approach always results in high threshold choices compared to the EQD technique. Furthermore, we have also provided a more in-depth spatial analysis of observation sites, showing GPD parameter estimates and return levels obtained using the TAILS method. Such plots illustrate some spatial patterns within the model fits across sites; however, due to the sparsity of sampling locations, we can only clearly identify patterns in regions with large

numbers of observation stations. These figures are shown in the response to reviewer 2.

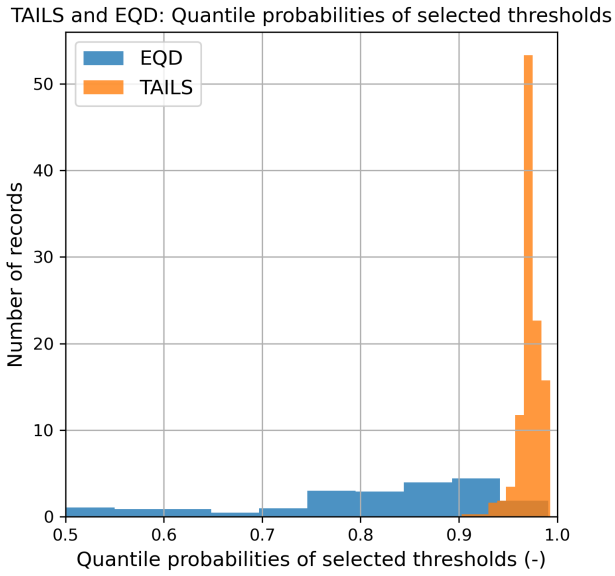


Figure 1. Updated Figure 4 in the manuscript. The results from applying the EQD and TAILS methods to every GESLA record used in this study, showing the distributions of quantile probabilities of the selected thresholds.

2. Following this recommendation, we have greatly expanded the discussion and split sections 5 and 6 into multiple subsections. We provide a more detailed breakdown of the results and conclusions from our study. We also highlight several areas of future work, including accounting for threshold uncertainty and expanding the TAILS approach to account for non-stationary and multivariate data. Moreover, we better highlight the relative strengths and weaknesses of TAILS compared to existing approaches.

Reviewer 2

The manuscript entitled *Automated tail-informed threshold selection for extreme coastal sea levels* by Collings et al. presents a methodological advancement for improving the threshold selection process in the Peaks-Over-Threshold (POT) framework. The authors propose an automated, data-driven approach tailored for extreme coastal sea level analysis, addressing a key bottleneck in both scientific and operational applications of extreme value theory.

The proposed method is potentially useful for practitioners, particularly in settings where arbitrary or fixed thresholds are problematic and require substantial expertise and technical judgment from the user. The manuscript is timely and relevant for NHESS, as it contributes to a sensitive topic and to the ongoing discussions around robust quantification of coastal hazards. It is generally well-written and understandable.

However, the discussion of the results should be expanded and more clearly structured to enhance the scientific impact of the work. The validation of the method—although supported by two illustrative case studies—remains somewhat limited. A more comprehensive evaluation of the model's performance across a wider range of sites or under varying data availability conditions would help assess its robustness and provide stronger support for broader applicability.

For these reasons, I recommend a major revision before the manuscript can be considered for publication.

We thank the reviewer for this useful feedback. As mentioned above, we have greatly expanded our discussion of the results, splitting sections 5 and 6 into multiple in-depth subsections. We also considered applying the approach to a wider range of sites, but decided against this for several reasons. Firstly, we believe our case study is already comprehensive, covering a significant area of the world with 417 study locations. Finding additional data sets of comparable quality and study length is also not straightforward and would represent a significant research project in itself without any guarantee of success. We therefore opted to instead vary the data availability using the GESLA database, and assess how this impacts the performance of TAILS.

We provide detailed responses to the provided comments below.

MAJOR COMMENTS:

1. It would be useful to explore TAILS approach sensitivity to the length of the calibration sample, i.e., the number of available years of observation. Since the method is appealing for practical applications by professionals and practitioners, understanding its robustness under limited data availability would greatly increase its usability in poorly gauged sites;

   Given the focus of our approach is on adjusting the EQD approach to focus on tail observations, we would naturally expect a reduction in performance for lower sample sizes, since this in turn reduces the number of observed extreme events. To test the sensitivity in this case, we applied the TAILS approach to a subset of sites and randomly varied the size of the observation period. These results are presented in the appendix. Encouragingly, even though the tuning parameters were selected using the full sample, the chosen GPD thresholds and parameters do not change significantly when the sample size is reduced, suggesting our approach can be robust to the size of the observation period. The same could not be said for the EQD method, which varied more with the sample size. Of course, this trend would not continue indefinitely, and one should always aim, when possible, to use data-rich samples when applying extreme value modelling techniques.

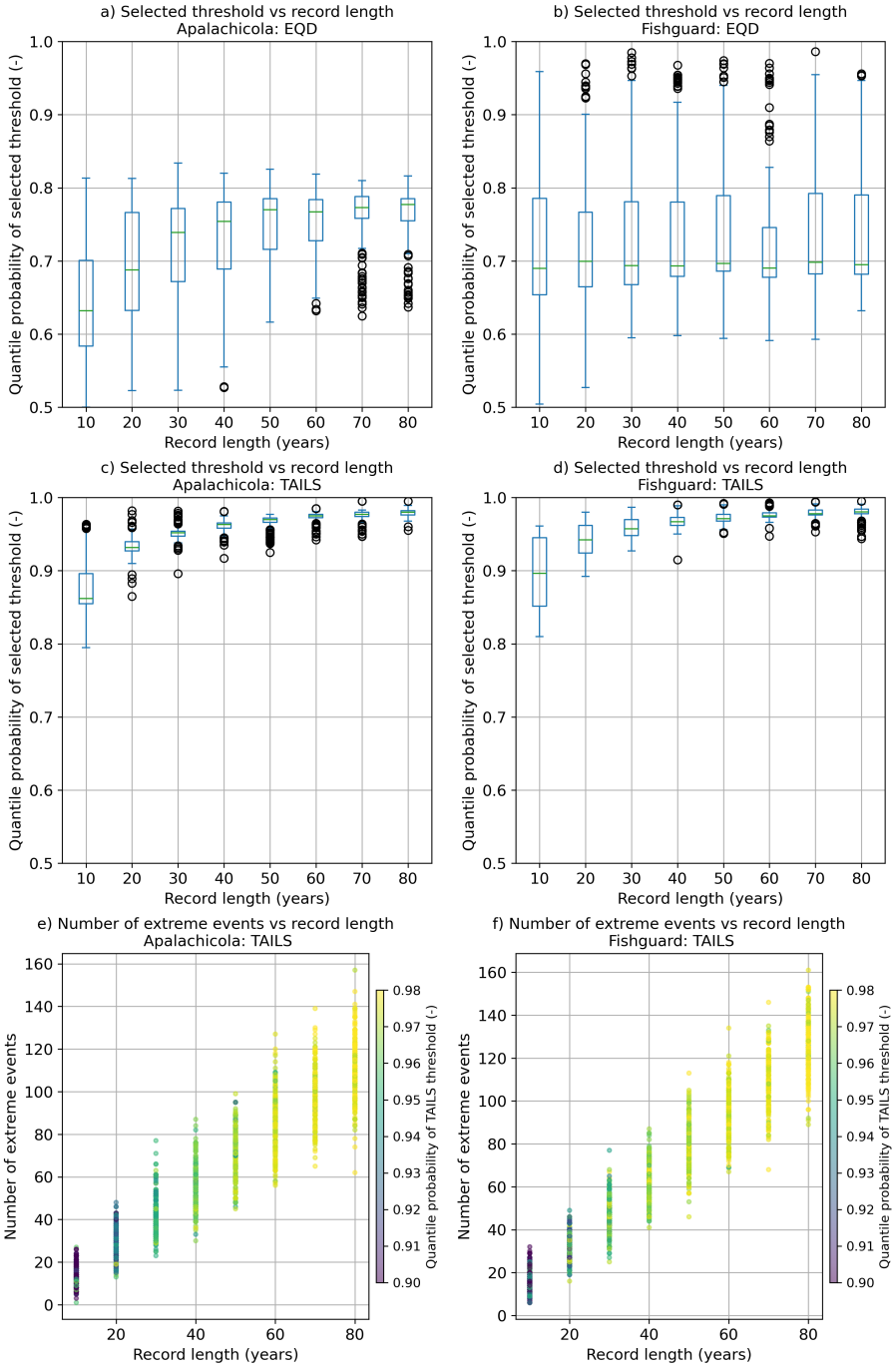
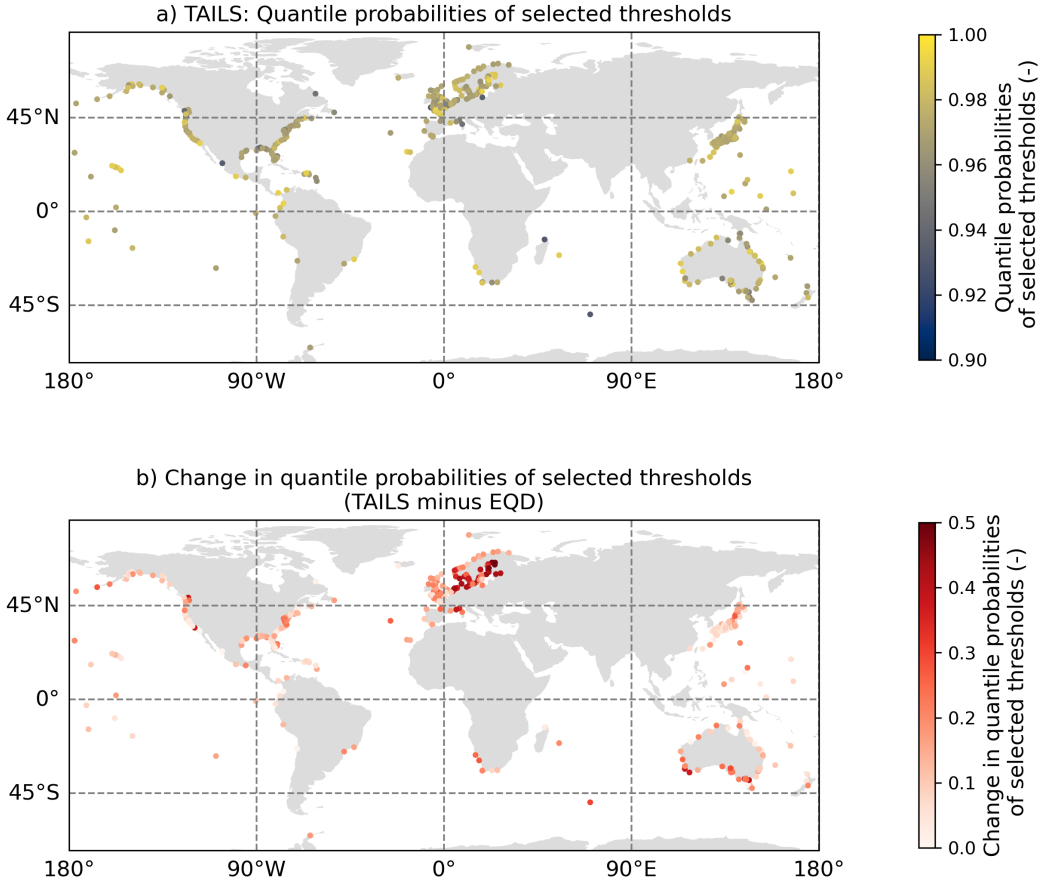


Figure 2. New Figure A5 in the appendix. Sensitivity test of record length and the number of extreme events present in the record against the quantile probabilities of the selected thresholds. The left column shows the results for Apalachicola, US, and the right column is Fishguard, UK. Panels a) and b) compare record length against the EQD method. Panels c) and d) compare record length against the TAILS method.

Panels e) and f) show the results comparing record length against the number of extreme events present in the records, along with the quantile probabilities of the selected thresholds using the TAILS approach.

2. Given the more weight attributed to the tail of the events population, that is the novelty and authors contibution to overcome existing approaches limitations, the more extreme events presence in data records effects are even more important in the proposed technique. I am wondering how sensitive the model threshold selection is to very rare events compared to the other methods. Artificially removing/adding adding large enough events or, consistently with previous point, years containing large enough events, and testing obtained thresold/quantiles can give some useful insights; In line with the previous comment, we applied the TAILS approach across a subset of sites while randomly varying the size of the observation period. This has the same effect as randomly removing and/or including certain extreme events from the time series, thus allowing us to assess the sensitivity to certain events. As expected, there was some discrepancy as we altered the size of the observation window, but this was not significant, suggesting our approach appears relatively robust to one-off extreme events. The results of this are shown in panels e) and f) of the figure above. In this test, an extreme event is defined as a water level in the bootstrapped sampled record that is greater than the 0.99 quantile of the original declustered record.

3. Considering the global scale of the analyses conducted, it would be valuable to assess whether any spatial patterns emerge in the performance of TAILS relative to conventional methods. Identifying systematic spatial behaviours, if any, could offer useful insights for practitioners and strengthen the case for its broader adoption, especially where spatially consistent behavior is observed; We have added a range of spatial plots (see below) illustrating several features of the model fits, including the quantile probabilities of the selected thresholds from the TAILS approach, and the change in threshold from the EQD to the TAILS method. This figure has been added into the manuscript as Figure 5. We have also included spatial plots of the shape and scale parameters of the GPD, as well as the differences between the EQD and TAILS which has been added to the appendix. All thresholds selected using the TAILS method are greater than the thresholds selected by the EQD. Strong spatial patterns are present particularly at tide gauge locations in north-eastern Europe. The tide gauge records with the largest increases are located in the Baltics, showing changes of nearly 0.5. Spatial trends are also visible around Australia, with the TAILS approach selecting higher threshold probabilities around the south of the country compared with the north. Similar patterns appear evident for northern Japan and the north-west US. The scale parameters are generally quite small, with the exception of German/Danish coastlines, which have values around 0.3 - 0.4. Overall, we see a reduction in the scale parameter obtained using the TAILS approach when compared with the EQD. Some locations show increases, such as along the German/Danish coast, Japan and North East US. The shape parameters have more variability globally, with strong positive values present along the US east coast and Caribbean. Europe generally exhibits negative shape parameters, which are more common for areas impacted by extratropical storms, although some outliers persist. When comparing the differences between the TAILS and EQD approaches, we see increases in the shape parameter in the vast majority of locations. This supports the observation that using the TAILS approach results in heavier tail estimates of the GPD. We acknowledge that the sparsity of observation stations makes it difficult to identify clear patterns in many regions, and as such, all identified patterns tend to occur around Europe, Japan, USA or Australia.



100

Figure 3. New Figure 5 in the manuscript. Spatial plots of a) the quantile probabilities of selected thresholds using the TAILS methods, and b) the difference in the quantile probabilities of the selected thresholds between the TAILS and EQD approaches.

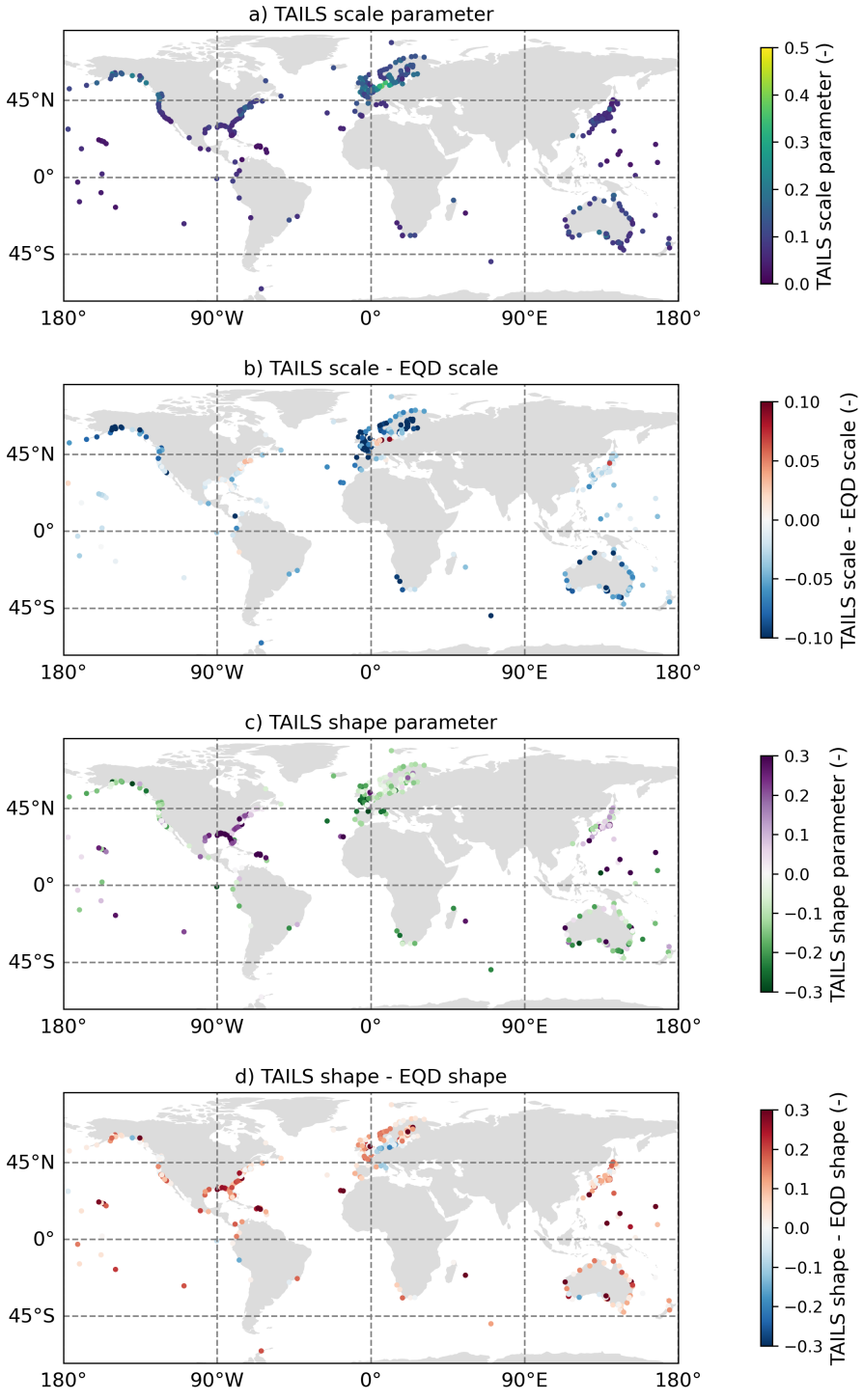

a) TAILS scale parameter
45°N
0°
45°S
180°
90°W
0°
90°E
180°
0.5
0.4
0.3
0.2
0.1
0.0
TAILS scale parameter (-)
b) TAILS scale - EQD scale
0.10
0.05
0.00
−0.05
−0.10
TAILS scale - EQD scale (-)
c) TAILS shape parameter
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
TAILS shape parameter (-)
d) TAILS shape - EQD shape
TAILS shape - EQD shape (-)

Figure 4. New Figure A4 in the appendix. Spatial plots of the scale parameter (a) and the shape parameter (c) of the GPD when using the TAILS method. The difference in scale parameter obtained using the TAILS approach vs the EQD approach is shown in panel b) and the difference in the shape parameter is shown in panel d).

4. The study provides a good validation of the proposed method based on the exceedences over the thresholds goodness-of-fit. Given the practical relevance of the proposed method, I would suggest to include a more comprehensive benchmark of the model performance on the annual maxima, in addition to the two case studies currently presented. Applying the method to a larger, eventually selected, set of stations could offer a more comprehensive assessment of its return levels predictive potential, that is of primary interest from an engineering perspective.

   We thank the reviewer for this suggestion, but respectfully argue from a statistical perspective that the use of peaks-over-threshold (POT) is more appropriate and widely accepted in modern extreme value analysis than annual maxima (AM). This selection is particularly relevant when the goal is to make efficient use of the available data. In our case, we have some records with only 40 years of observations available, which would give just 40 observations for the AM model. Quantities computed from this model, such as return levels, would have far higher variability compared to those obtained using POT techniques, reducing their usability and reliability.

   This has been a discussion point in many significant works; for example, Davison and Smith (1990), Coles (2001) and Scarrott and MacDonald (2012) all advocate for the use of POT over AM on the basis of data efficiency and estimation accuracy. Therefore, we believe comparing our approach with AM would not add meaningful value to our study and could potentially dilute the clarity of our methodological focus.

   We have updated the discussion in Section 1, page 2 to account for this comment.

MINOR COMMENTS:

1. I would also mention the work from Tancredi et al. (2006) in the POT modelling section as a Bayesian study that explore how to integrate uncertainty in the threshold selection; We have updated the literature review in Section 3 (page 4) to include this additional reference.

2. The authors mention the GESLA 3.1 update as a minor revision of the GESLA 3 dataset, which is provided by one of the authors. As far as I am aware, this updated dataset is not yet publicly available. The authors are encouraged to clarify whether they plan to release it, to ensure reproducibility and broader adoption of the proposed methodology. We have updated the data availability statement to clarify that the revised GESLA 3.1 database is available from the corresponding author upon request. We have been told the revised GESLA 3.1 database will be released publicly in the next couple of weeks as well.

References

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.

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.

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

Tancredi, A., Anderson, C., and O'Hagan, A.: Accounting for threshold uncertainty in extreme value estimation, Extremes, 9, 87–106, 2006.