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 the reviewer for their 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 3

The paper *Automated tail-informed threshold selection for extreme coastal sea levels* by Collings et al. presents a new methodology (called TAILS) to define thresholds when performing Extreme Value Analysis to Peak Over Threshold data. The topic of the research aligns well with the objectives of the journal, the paper is well written and results are in general properly presented and discussed. However, there are some aspects that should be better clarified before the paper is suitable for publication. See the comments below.

We provide detailed responses to the provided comments below.

MAJOR COMMENTS:

1. First and most important, Authors use the p-value of the ADr test to prove that the TAILS method works better than two other methods, i.e., the expected quantile discrepancy and the Q98. It follows that the ultimate goal is to select subsets that are well modeled by a GPD; I therefore believe that other tests should have been employed for a comparative analysis, namely those sharing the same objective - see for instance the work by Solari et al., 2017, which is also mentioned in the paper. On the other hand, the use of a fixed quantile (Q98) seems more like the first step for declustering the exceedances within a two-step selection (e.g., Bernardara et al., 2012), which should lay the ground for a subsequent selection of a statistical threshold - therefore aiming to ensure a proper distribution fit. This should be at least commented in the paper.

   We have selected the ADr test to assess goodness-of-fit (GOF) as it is particularly sensitive to discrepancies in the upper

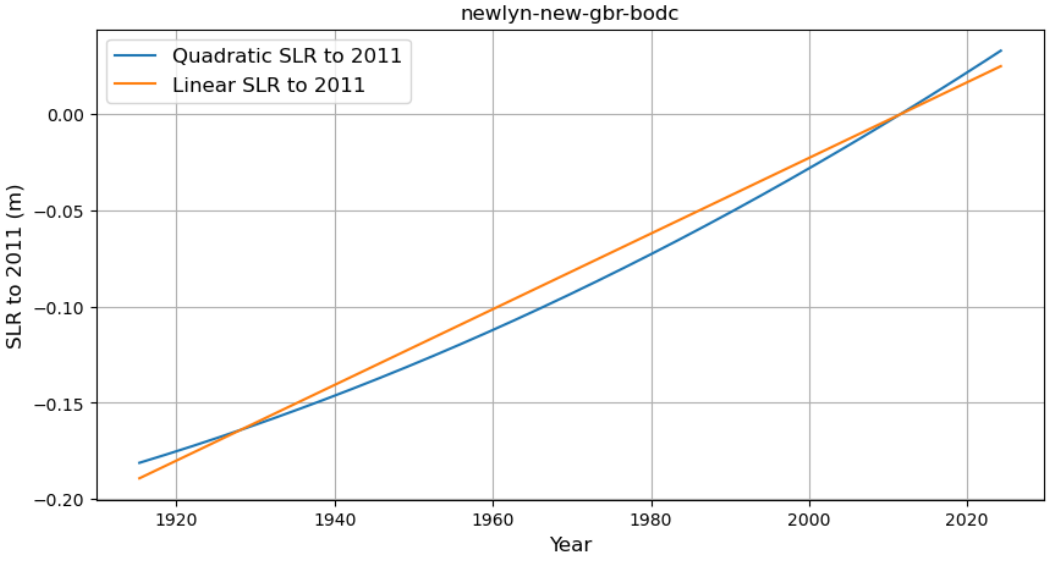
tail, which is what we aim to characterise well with our method. Furthermore, it is one of the most commonly used GOF measures from the literature (Heo et al., 2013; Gharib et al., 2017; Benito et al., 2023). In Solari et al. (2017), they use the ADr p-value as the metric by which to determine the best threshold, but we cannot find any references in the paper to other GOF measures used to assess the resulting GPD fits. The aim of this study was not to be an exhaustive review of GOF measures, but to concisely demonstrate the efficacy of our method against other commonly used methods. Using a single metric that is well suited to the area of the distribution we aim to characterise makes the comparison easier to communicate, without the added ambiguity and complexities of extra GOF measures. We therefore respectfully maintain our choice of the ADr test as the sole comparative metric, while acknowledging that other tests may also provide useful complementary perspectives in future work. As for the use of the Q98 as the first step in a two-step selection process, it is also used as the final threshold in other global studies, as referenced in the manuscript. For the avoidance of doubt, we have added the following sentence to the introduction - 'Note that the use of static thresholds, such as the Q98, are common in two-step threshold selection processes (e.g., Bernardara et al., 2012), and should not be confused with the use of the static threshold as the threshold above which to consider an exceedance.'

2. Second, I found some parts of the methods hard to follow. I am not familiar with the work by Murphy et al., (2025), so I apologize in advance if some questions may look naive. At page 5, line 111, n exceedances xu are considered. As such, shouldn't there be as many associated probabilities P? In other words, isn't m equal to n? By looking at lines 118-119 it seems so (i.e., Q is associated to xu b). If that is the case, does it make sense to use a fixed value of m to compute the probabilities q for increasing thresholds? (see the last equation at page 5, which by the way should be numbered). Moving to the TAILS approach, I do not understand why in Equation (3), at the numerator, there are both pi and pj. Is that a typo? Moreover, if xu b is defined as in Section 3, that implies that it is based on a sample of excesses of a candidate threshold u. If so, I do not understand what do probabilities πu represent (clearly not the chance that an excess of a threshold is lower than the threshold itself). In summary, I think that the whole methodology should be better explained to avoid any confusions. We appreciate that the content referenced here is rather technical in nature and may be confusing at first, especially to those unfamiliar with such techniques. We have therefore added some additional comments to clarify the approaches discussed. Firstly, for each threshold u, there will be n_u exceedances, with n_u varying over u and the sample in question. However, the set $\mathcal{P}_m$ (and value m) is defined independently of the threshold choice/sample, and as such the same probabilities are considered for every threshold. In the case when $m > n_u$, the quantile function Q linearly interpolates between the observed quantiles. We have adjusted the text on page 5 (now line $\sim$115) to make it clear $\mathcal{P}_m$ is fixed over threshold and n_u is sample + threshold dependent. In practice, we do exactly as you suggested and keep m fixed. We also have labelled the final equations on page 5 (now on page 6)

In equation 3 (now equation 5), we originally opted for different indices in the numerator and denominator to make it clear these sums are computed separately before their ratio is evaluated (i.e., we don't take the sum of ratios, but rather the ratio of the sums). However, to avoid confusion, we have updated the indices to both be i.

Finally, the probabilities π_u represent (empirical) non-exceedance probabilities for each threshold u. We are only able to evaluate quantiles at probabilities great than π_u, since the GPD is obviously not valid below the threshold u. This explains the rather dense and complicated equation 3 (now equation 5), and we have a comment on page 8, line ~175, stating that this equation 'accounts for cases when the threshold probability, π_u, exceeds a subset of $\mathcal{P}$'. We have also added the comment 'In other words, this ensures the fitted GPD is only evaluated above the candidate threshold' to better explain the proposed method.

3. Finally, at page 3, line 66, Authors claim that SLR is accounted for by removing a linear trend to all data. Is this a reliable assumption on a global scale? Given the length of the time series, would not an exponential trend be better suited for this purpose at least to some locations? Could you elaborate on this point in the text?; Whilst SLR can be modelled as an quadratic trend, modelling it as a linear trend is common in other regional and global studies (Sweet et al., 2020; Frau et al., 2018). We tested this a sample of sites globally and show that difference between an quadratic trend and a linear trend is small (see figure below for Newlyn, UK). We have added the following sentence to section 2 - "Although some tide gauge stations show an accelerating sea level change, a linear trend is judged to be sufficient to model sea level change in this study."



MINOR COMMENTS:

1. In Figure 1, the two test sites are hard to see. Consider adding an inset with a close-up on the area of interest; Thank you for your suggestion. We have added two insets to improve clarity of the figure. Please see the updated figure below.

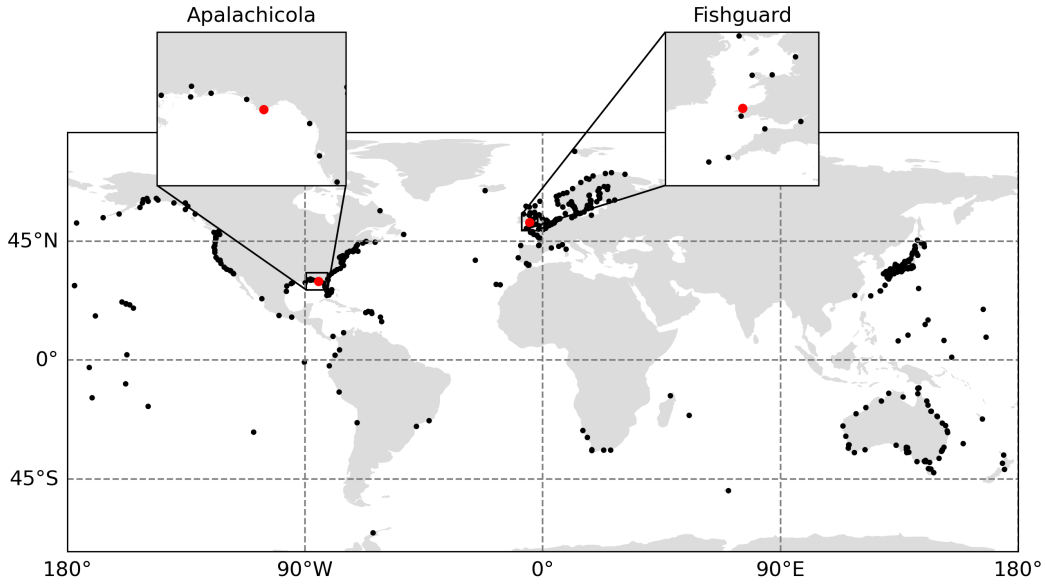


Figure 1. New Figure 1 in the manuscript, showing the locations of the GESLA records and the highlight Fishguard and Apalachicola using map insets.

2. In Figure 5, using fewer colors would help appreciate the spatial differences in the results; We have reduced the number of colours in this plot, as well as in Figure A4 which shows similar global plots. The updated Figure 5 is shown below.

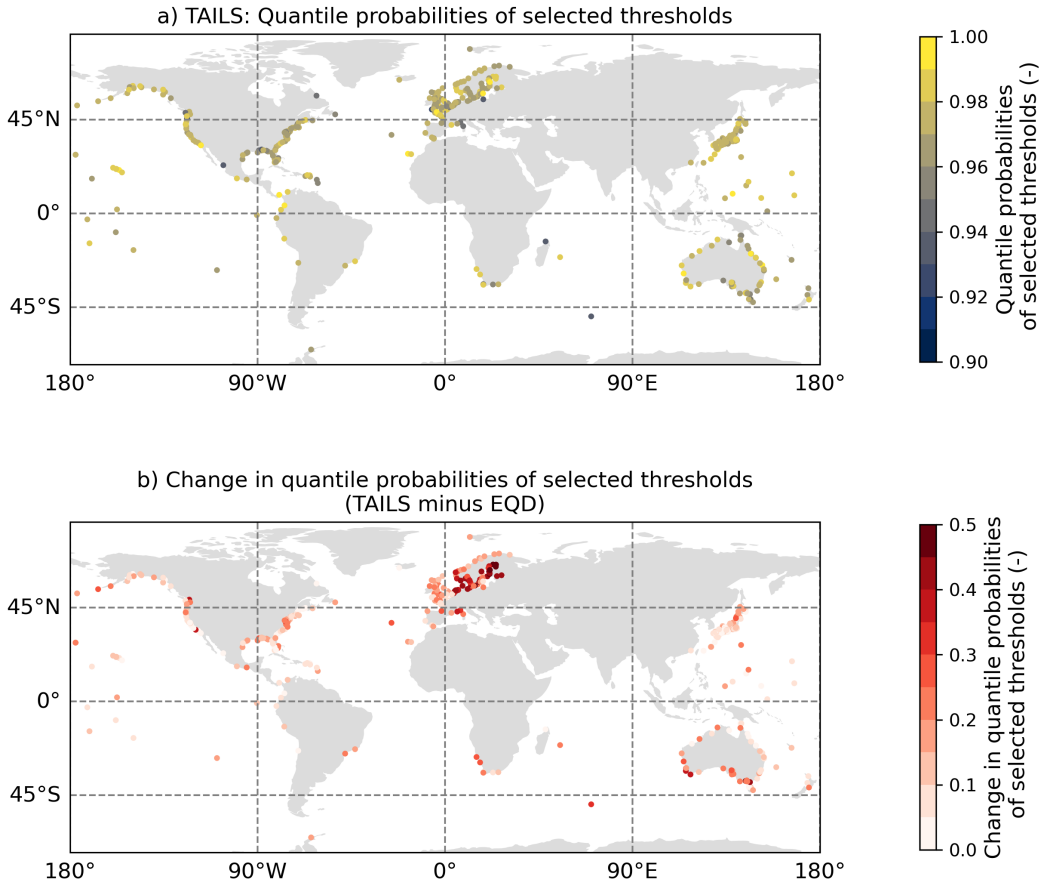


Figure 2. New Figure 5 in the manuscript, with 10 colours used in each colour map.

3. At page 11, I like the use of the p-value as a GOF measure. In this respect, Authors may want to cite the seminal work by Wasserstein & Lazar (2016); Thank you for the information. We have added the reference to the manuscript on page 11.

4. Panels c and d in Figure 7 are very busy and hard to interpret; We have removed the shading of the confidence intervals as we believe this improves the clarity of the figure. Please see the updated figure below.

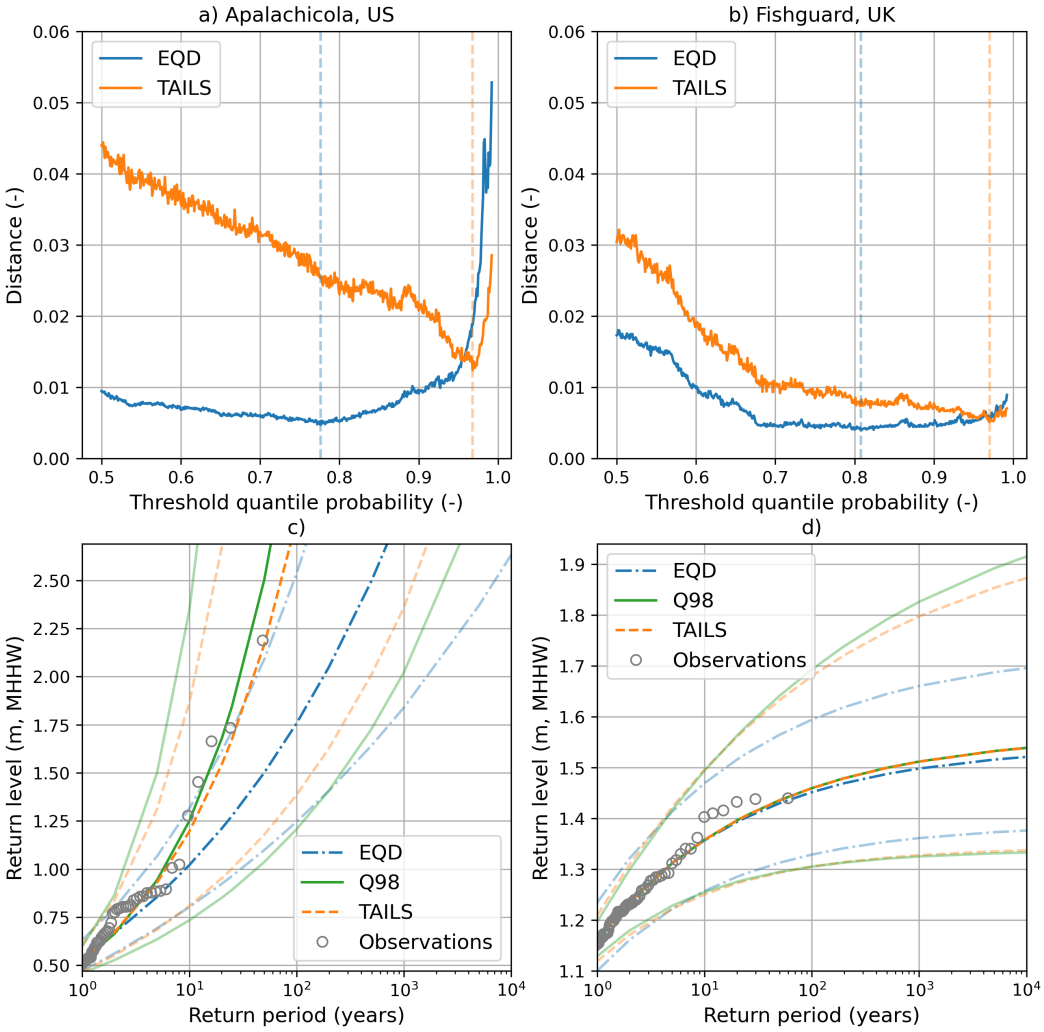


Figure 3. New Figure 7 in the manuscript, with the shading of the CI removed.

References

Benito, S., López-Martín, C., and Navarro, M. A.: Assessing the importance of the choice threshold in quantifying market risk under the POT approach (EVT), Risk Management, 25, 1743–4637, https://doi.org/10.1057/s41283-022-00106-w, 2023.

Bernardara, P., Mazas, F., Weiss, J., Andreewsky, M., Kergadallan, X., Benoît, M., and Hamm, L.: On the two step threshold selection for over-threshold modelling, Coastal Engineering Proceedings, 1, management.42, https://doi.org/10.9753/icce.v33.management.42, 2012.

Frau, R., Andreewsky, M., and Bernardara, P.: The use of historical information for regional frequency analysis of extreme skew surge, Natural Hazards and Earth System Sciences, 18, 949–962, https://doi.org/10.5194/nhess-18-949-2018, 2018.

Gharib, A., Davies, E. G. R., Goss, G. G., and Faramarzi, M.: Assessment of the Combined Effects of Threshold Selection and Parameter Estimation of Generalized Pareto Distribution with Applications to Flood Frequency Analysis, Water, 9, https://doi.org/10.3390/w9090692, 2017.

Heo, J.-H., Shin, H., Nam, W., Om, J., and Jeong, C.: Approximation of modified Anderson–Darling test statistics for extreme value distributions with unknown shape parameter, Journal of Hydrology, 499, 41–49, https://doi.org/https://doi.org/10.1016/j.jhydrol.2013.06.008, 2013.

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., 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.