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.

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[1]), 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.

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 x_u 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 $x_u{}^b$). 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 p_i and p_j. Is that a typo? Moreover, if $x_u{}^b$ 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 π_u 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.

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?

See below other minor comments:

- In Figure 1, the two test sites are hard to see. Consider adding an inset with a close-up on the area of interest;
- In Figure 5, using fewer colors would help appreciate the spatial differences in the results;
- 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)[2];
- Panels c and d in Figure 7 are very busy and hard to interpret;

[1] Bernardara, P., Mazas, F., Weiss, J., Andreewsky, M., Kergadallan, X., Benoît, M., & Hamm, L. (2012). On the two step threshold selection for over-threshold modelling. *Coastal Engineering Proceedings*, *1*(33), 1-6.

[2] Wasserstein, R. L., & Lazar, N. A. (2016). The ASA statement on p-values: context, process, and purpose. *The American Statistician*, *70*(2), 129-133.