Abstract

The second paragraph is focused on a general description of RFA method illustrated through the example of Cyclone Yasi, while it should be personalized to your study. I would describe the method based on your application globally and provide details about the results obtained. This expansion will emphasize the method's capabilities and the valuable resource for coastal risk assessments.

Thank you for your suggestion – we have included another paragraph as follows:
"The methodology presented in this paper is an extension of the regional framework from Sweet et al. (2022), with innovations made to incorporate wave setup and apply the method globally. Water level records from tide gauges and a global reanalysis of tide and surge levels are integrated with a global ocean wave reanalysis. Subsequently, these data are regionalised, normalised, and aggregated, and then fit with a Generalised Pareto distribution. The regional distributions are downscaled to the local scale using the tidal range at every location along the global coastline, obtained through a global tide model. The results show 8cm of positive bias at the 1-in-10-year return level, when compared against individual tide gauges."

INTRODUCTION

I notice the absence of the reference to Calafat et al. () in the text, even though it is included in the references section but not cited in the body of the work. This reference is particularly relevant, representing one of the most significant contributions in recent years where the RFA method is applied.

The Calafat et al. in-text citation must have been removed during a previous edit. We have readded it, thanks for making us aware.

Lines 125-128: "The principle of an RFA is founded on the basis that homogenous region can be identified, throughout which similar meteorological forcings and resultant storm surge or wave events could occur, even if the extreme events have not been seen in part of that region in the historical record (Hosking and Wallis, 1997)."The key to the method lies in the similarity of meteorological forcings. In your global scale application, you divide the entire globe into grid cells of 1 degree and then apply a 400-radius at the grid centroid. Could this definition of homogeneous regions significantly impact your results? Have you conducted a sensitivity analysis concerning meteorological forcings? This question is also related to the one about the Heterogeneity test.

The 400km radius was the same as was used in Sweet et al. (2020) and Sweet et al. (2022). The following is an excerpt from their 2020 paper – "Rather, the regions were bounded with a maximum distance of 400 km (about 250 miles) around a particular military site. The RFAs were conducted using water level records from up to five tide gauges. Our regional range was smaller than the approximately 1000-km synoptic scale of extratropical disturbances (i.e., storm footprint), approximated the diameter (twice the radius) of maximum winds associated with the largest of the 1000+ synthetic hurricanes recently modelled by the U.S. Army Corp of Engineers (Nadal-Caraballo et al. 2015), and was on the order of (Weiss et al. 2014) or smaller than (Hosking 2012) homogeneous regions identified in related RFA-based studies."

We agree that the maximum radius used for the selection of water level records would certainly impact results, but it was decided that such a sensitivity test would be a large amount of work to add to an already complex paper and was therefore beyond the scope of this project. It is certainly something we aim to look at when we update this method. We have added 2 sentences to the paragraph outlining limitations in the discussion to reflect this – "Firstly, changing our definition of a homogeneous region would likely have a great impact on our results. In future iterations of this study, we recommend carrying out a sensitivity analysis to understand how using different maximum radii to select water level records impacts upon estimated extreme water levels within the region."

In the grid cell example shown in Figure 2A, could it be possible to consider that extreme water levels along the east coast of Florida and the west coast of Florida could be generated by common storm track? Perhaps not the extra-tropical, but in the case the tropical cyclones? The way homogeneous regions are defined could potentially explain some of your "undesired results" (lines 556 – 560; lines 619-622: decreases in ESL exceedance probabilities compared to the single site analysis).

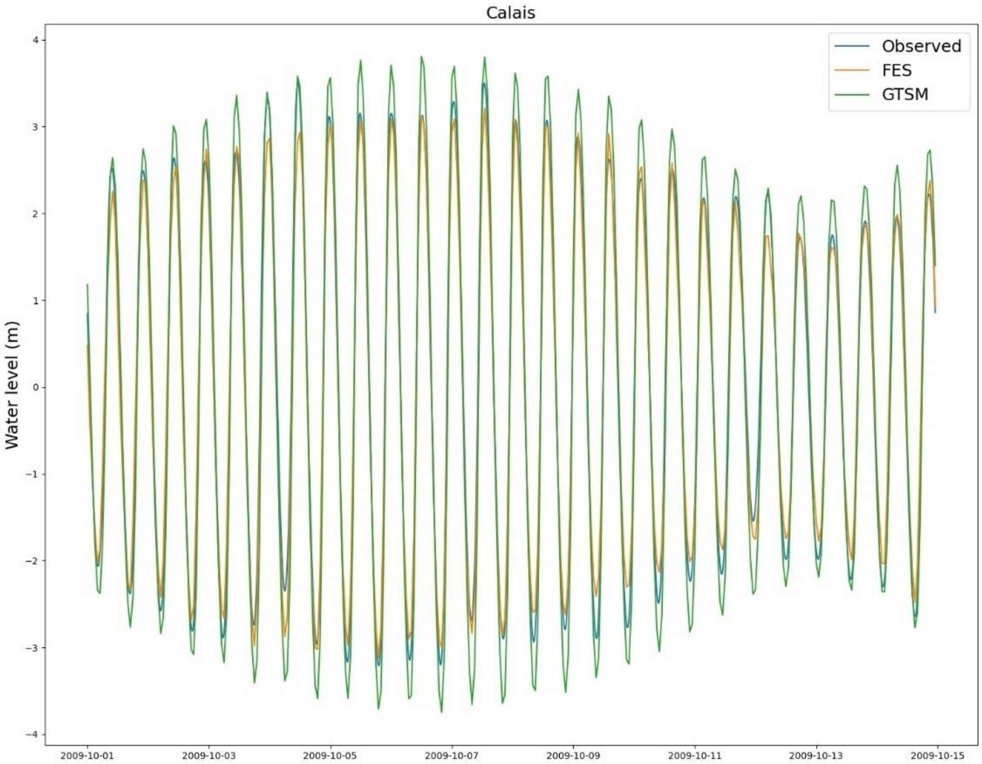
Each grid cell is considered independently, and water levels are declustered around a 4-day storm window. In the example given here, a large TC event could cross over the Florida peninsula and drive a high water event on both the east and west coasts, however this would simply be interpreted as a single event in the RFA of each grid cell.

DATA

FES2014: higher resolution and performed better than those of GTSM (lines 262-262). Have you checked the differences between the tides between GTSM-ERA5 and FES2014? In the case of GTSM-ERA5, the non-linear interactions between tides and surges are included because they have been simulated together. How might these effects influence sea levels (tide + surge)? What implications arise from the overestimation of the linear sum of these dynamics (Arns et al., 2020)?

We identified a few key areas in which the tides in GTSM-ERA5 were particularly erroneous, which initially lead us to using FES2014. The authors of the GTSM-ERA5 data also concluded the following – "It appears that biases increase in regions with a high tidal range, such as the North Sea, northern Australia, and the northwest of the United States and Canada, which could indicate that GTSM is outperformed by the FES2012 model that was used to develop the GTSR dataset."

Below is a figure demonstrating the overprediction of the tidal range by GTSM in Calais, France.



GTSM-ERA5 had two simulations, a tide only simulation and a simulation including tides and atmospheric (ERA5) forcings. The surge component of GTSM-ERA5 was isolated by removing the tide only water levels from the combined model run. Therefore, the surge component of GTSM-ERA5 is inclusive of tide-surge interactions. These tide-surge interactions will be erroneous in areas where the tidal component of GTSM is poor, but as the model was run with tides included it is impossible to remove this. In any areas, where the GTSM tidal heights vary greatly from the FES2014 tidal predictions, then the method will fall foul to the issues highlighted in Arns et al., (2020).

You are using a global wave reanalysis at a resolution of 0.5. Have you considered the potential impact of this coarse resolution, particularly along the coast where the propagation process that modifies waves approaching the coast is not accounted for? How might this simplification influence your results? Have you contemplated applying a simplified method to address this limitation?

The simplified method for applying wave setup is definitely a limitation of our study, especially given the resolution of the global wave reanalysis. Other methods (such a Stockdon) were considered at the time, but given the global application were deemed beyond the scope of the project. We have added the following to the manuscript to highlight this simplification – "Applying wave setup using this approach is an obvious simplification that has been used for the ease of global application. In reality *wave setup is impacted by local bathymetry and coastal geometry, as well as local wind and wave conditions. There are other more complex methods for estimating wave setup that incorporate some aspects of bathymetry and coastal geometry, such as Stockdon et al. (2006)."*

We have also added this as a constraint of the study and recommended that updating this be a focus of future research – "The method used to incorporate wave setup is another constraint, as it has been greatly simplified for ease of global application. Improving upon this should also be a focus of future studies."

Regarding the definition of COAST-RP dataset, It is not very clear for me the sentence "in extra-tropical regions, a 38-year timeseries of ERA5 is used" (lines 226-227).

We have changed the sentence to improve the clarity as follows – "in regions impacted only by extra-tropical storms, a 38-year timeseries of ERA5 is used"

<u>METHODS</u>

I believe Figure 3 contributes to a better understanding of the methodology. However, I suggest indicating and structuring it based on the five key steps enumerated in lines 240-249. Similarly, I propose adopting the same structure for the subsequent subsections. For example, 3.1 Data Processing, which may correspond to i) and ii), and 3.2 RFA, aligning with the remaining steps, could be further divided. For example, 3.2 subsection could be divided based on the steps and the scheme presented in Figure 1. For instance, downscaling and bias correction could be more effectively presented in separate subsections for clarity and coherence.

We are assuming Figure 2 is referenced here, rather than Figure 3. Figure 2 is intended to illustrate the key steps through the regionalisation process and the fitting of the GPD to the aggregated data, rather than to illustrate the whole process. The steps outlined in the line 240-249 are high level, and so to incorporate them all into figure would mean losing important detail in the current figure, to make space for processes which do not benefit hugely from being illustrated in a figure (e.g. pre-processing of data). However, we agree splitting the method up into 5 sections as is indicated in lines 240-249 would improve clarity and help sign post readers to specific sections, and so we have implemented this change.

<u>Lines 270-271:</u> "wave setup is interpolated to the nearest record location using a nearest neighbour approach". I think a nearest-neighbour approach cannot be considered as an interpolation method, you are directly assigning the closest node

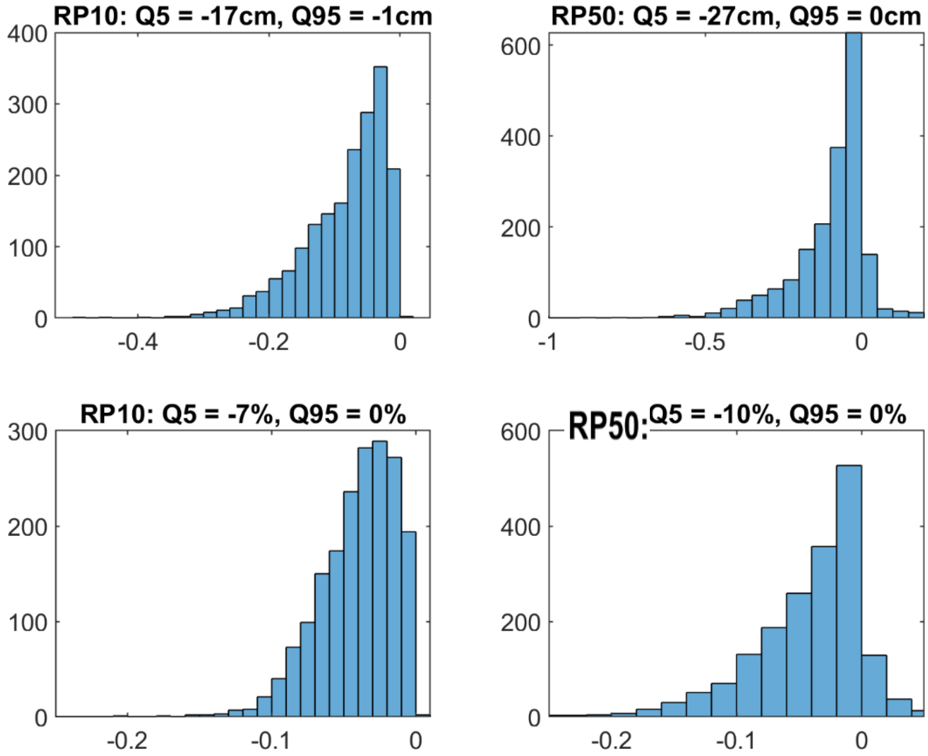
We have changed "interpolated" to "assigned".

Lines 271-272: "to account for the lack of wave setup in sheltered areas…" I would phrase it the other way around; it is the type of coastline that determines whether the wave setup is considered or not.

Thank you for the suggestion. We have amended the manuscript as follows – "Wave setup is assumed to be absent in sheltered areas (e.g., bays and estuaries). To account for this, the global coastline is classified as either sheltered or exposed…"

Lines 278-279: Could an overestimation of water levels result from adding the daily maximum wave setup to the daily highest water level? Have you checked if the "total water level," calculated as the sum of storm surge, astronomical tide, and wave setup at an hourly scale, and then selecting the extreme water levels, exhibits peaks that are very similar (in the same order) to the sum of the maximum daily values of surge-tide and maximum daily significant wave height?

We decided to use daily max waves because we were concerned about the noise in the results from fitting a copula to sub-daily surge heights to extend the wave hindcast beyond the limit of the reanalysis years, to cover some of the older tide gauge records. In hindsight this was probably an oversight. We have run an analysis to compare the difference it would have made if we had used max hourly wave height compared to daily max. Below is a figure showing the difference at RP10 and RP50 at a selection of GTSM sites around the world. As a percentage, the mean bias this introduces is 4% for both RP10 and RP50. As this is relatively small we don't think it's worth rerunning the entire analysis.



Lines 282-284: Tide gauges are assumed to be located in sheltered regions such as bays and estuaries, and consequently, wave setup is not considered necessary. Why is it necessary to fit a copula between daily peak water levels and daily maximum significant wave heights when tide gauge records fall outside the temporal range of the ERA5 data for providing predictions of daily maximum significant wave heights (lines 279 – 281)? If this fitting is required, which database are you utilizing for surge + tide and significant wave heights? Is it GTSM-ERA5, FES2014, and ERA5?

Tide gauges are assumed to be in sheltered areas, but the RFA needs to be applied on both sheltered and exposed coastlines. We therefore apply wave setup to all tide gauge water level records, so that we can produce local extreme water levels on sections of exposed coastline that are inclusive of wave setup. The ERA5 wave reanalysis used extents back to 1979 and so does not cover many older tide gauge records. We therefore fit a copula to the daily peak water levels (from the tide gauge in question) and daily max Hs (from the nearest output point from the ERA5 wave reanalysis). We can then use the copula to estimate the daily max Hs for peak events that occurred before 1979.

Lines 306-307: why the final number of tide gauges are 836 while at the beginning the total number were 2223 (line 258)? I am not clear on how the discretization of regions leads to this reduction. (same comment as the other reviewer)

The difference between the two numbers is that there are 2,223 gauges which pass the QC checks and are not located on rivers, but only 836 tide gauge records which remain after spatially discretising and removing records that cover a period of less than 10 years. As this is clearly confusing, we have removed the first statement of the number of gauges, and amended the remaining statement to include the average number of years of the tide gauges actually used in the RFA – "This spatial discretisation of regions results in a total of 836 tide gauge records (with a mean record length of 17 years) and 18628 GTSM-ERA5 records for use in the application of the RFA."

Figure 2. I suggest to plot the tide gauges selected (Figure 2A) in the same color as the time series in Figure B.

We have amended as suggested, thanks.

Lines 342-343: the index flood u defined as the 98th percentile for all locations – same comment as the other reviewer

The reference has been added – Sweet et al. 2022.

How does the Heterogeneity test operate in terms of assessing the homogeneity of a region? What characteristics are considered to deem time series within a region as homogeneous? If the test fails due to an anomalous record (lines 354), at what point could the anomalous record be considered an extreme event resulting from a tropical cyclone?

The heterogeneity test (Hosking and Wallis, 1997) assesses the regional L-moment ratios against those synthetic L-moment ratios randomly sampled from a Kappa distribution which has been fit using the L-moment ratios as parameters. The Kappa distribution provides a reference distribution for the L-moments under the assumption of homogeneity. If the original data is homogeneous, the synthetic L-moments derived from the Kappa distribution should be consistent with the observed L-moments. As the test assess the L-moment ratios of the region, it is difficult to assess the impact of a single extreme event in the record. In our experience of using the method, a region tends to fail if a gauge has multiple events of a greatly different magnitude impacting a just 1 gauge. This could potentially be a gauge which has been hit by a disproportionately high number of extreme events within its record but is more likely to be a gauge which isn't representative of the surge characteristics of the other gauges in the region.

Lines 366:367: How could the empirical threshold of 0.35 for the shape parameter impact the results of high return periods? Are you aware of typical values for the shape parameter in the case of tropical cyclones (TCs)? While you mention "expert judgment" (line 368), could you provide a reference supporting this approach? As you mention in the results, section 4.3 (lines 555-557), some gauges show decreases in the return levels, and this could be driven by the shape parameter that maybe limit to much large water levels?

There is the possibility that the limit on the maximum shape parameter could potentially lead to an underestimation of return levels for high return periods. However, in the work we did to assess the empirical threshold, in some instances we saw return levels of 50m+ for 1-in-1000 year return period. It was results like this that lead us to implement the limit.

In Sweet et al. (2020), they provide shape parameter estimates for 3 TC regions in the Pacific. The maximum (median) shape parameter they obtained was 0.228.

We have also added a section to the discussion following comments from the other reviewer, highlighting this as a limitation of the study – "Another limitation of the approach used in this study is the static shape parameter limiter. It is probable that the maximum shape parameter varies by location around the world, and that by implementing a fixed threshold globally we are perhaps limiting some of the most extreme events in some regions. Improving this section of the methodology is a high priority for future updates."

Lines 394-396: Regarding the last stage, which involves removing the bias in the high frequency portion of the exceedance probability curves, the bias is quantified based on the divergence in the 1-in-1-year return period at each tide gauge/GTSM-ERA5 location. With this approach, a constant bias correction is applied at each location; however, the bias can be higher for higher return periods, as shown in Figures 6 and 8. How might this simplification impact the results?

We only adjust for bias at the 1-in-1 year return period because we are confident the historical record can be used to accurately estimate such a high frequency return level, and therefore any increase in *our results can be called bias. The increase in the high frequency return levels is acknowledged as a consequence of the regionalisation process, and Sweet et al. (2022) and Sweet et al. (2020) find similar results. The increase in return levels at low frequency return periods is assumed to be because of better representation of rare, extreme events from across the region. We have added the following in response to a comment from the other reviewer - – "As TC hazard is typically underrepresented due to short records, it can be inferred that the increases observed across these regions are an improvement on a single site analysis."*

Additionally, why do you use the 99th percentile of tidal elevations to interpolate the bias across coastal locations? Why not use tidal range, as in the interpolation of index flood u, or the 98th percentile, as the peak over threshold to select the extreme events?

We used monthly Q99 from a 3-year period (centred on 2011) to determine the extents over which to apply bias corrects from each tide gauge. This has been explained clearly in the manuscript, so we've added the following - "...the correction term is interpolated across all coastal LEWL points based on correlation between monthly values of the 99th percentile of tidal elevations produced over a 3-year period centred on 2011..."

We could not have used tidal range as we needed to carry out a correlation coefficient, and we therefore needed a timeseries representative of the general characteristics of each location. It was an oversight not to use Q98 for consistency with other aspects of the paper.

Lines 421-424: description of COASTAL-RP dataset that should be included in section 2.

It is already included in the section 2 and so we have removed it from here. Thanks for the suggestion. The description reads – "COAST-RP uses the same hydraulic modelling framework as GTSM-ERA5 but simulates extra-tropical and tropical surge events separately using different forcing data. In areas prone to TC activity, synthetic TCs representing 10,000 years under current climate conditions are used from the STORM dataset (Bloemendaal et al., 2020)."

Lines 436-438: not completely clear, maybe 1000 grid cells which have between 3 and 10 GTSM-ERA5 record locations?

We have amended the sentence to improve clarity, as follows – "To do this, we identified 1000 grid cells which use 10 GTSM-ERA5 records for the RFA and contain 3 GTSM-ERA5 record locations inside the grid cell (and therefore the RFA can be used to directly estimate ESLs at the record locations)."

RESULTS

4.1 Global application of RFA

I wonder if in areas with tropical cyclones, a Generalized Extreme Value (GEV) distribution could be the best curve to represent the behavior of extreme water levels. What happens if there are two families of extremes due to extratropical storms and tropical storms with significantly different magnitudes? Perhaps considering the use of a mixed extreme value distribution, as suggested by O'Grady et al. (2022), might be beneficial.

The study by O'Grady et al. (2022) looks very relevant and interesting. Thank you for bringing it to our attention. We have added a sentence to the paragraph on future work – "Future updates could also include an assessment of using different extreme value distributions, perhaps following the mixed climate approach of O'Grady et al., (2022)."

4.4 Tropical Cyclone Yasi

I don't believe the case study of Cyclone Yasi illustrates how the RFA methodology enhances the representation of rare extreme events in the ESL exceedance probabilities. I think this section aims to demonstrate the regionalization process (as mentioned in lines 538-539). In some ways, these results suggest that the RFA method behaves like an interpolation of the shape parameter?

We have amended this sentence to remove the phrase "…improves the representation of rare extreme events…" so that it reads "…is to illustrate how the RFA methodology previously described can draw on few, rare events, to provide more realistic representation of low frequency ESL exceedance probabilities across a region, using the case study of cyclone Yasi…"

Line 495: Figure 5B.

Corrected, thank you.

Lines 503-507. It's not clear to me what is intended to be communicated with these sentences.

Following this comment and a comment from the other reviewer, we have removed these lines.

4.3 Comparison with GESLA

Areas where there is an improvement using RFA, should be also reflected in a better agreement with COAST-RP (section 4.4)?

Yes this is true, and we see agreement around the United States and in areas of Europe. Other areas which we see large increases in COAST-RP return levels are China, the Bay of Bengal and the western Gujarat region of India, however we don't have many/any tide gauges in these regions, and so it's difficult to quantify the increases from the tide gauge RFA in these areas.

4.4 Quantifying the improvements made by the RFA when compared to single site analysis

Why not plot the difference with COAST (GTSM with STORM) instead of GTSM-ERA? This would provide an overview of the magnitude of the differences and how your method approaches the COAST-RP results, not just in terms of the spatial pattern. On the other hand, the comparison with COAST-RP should only informs about the improvement in those areas prone to tropical cyclones.

The publicly available COAST-RP data is in the form of return levels. We do not have access to the raw data output from the model. The return levels have been computed using a different method that a GPD fit to POT exceedances (as we've used in our study). To compare the magnitude of return levels computed using different methods would incorporate differences/biases based on the methodological choice rather than the actual data. We therefore chose to use this method assessing spatial patterns because to redo the RFA using another method for estimating return levels would have been a vast quantity of work.