Dear Editor,

Thank you for forwarding the comments from the reviews. We are very grateful to the reviewers for their constructive feedback. Below, we describe in turn how we have addressed each one. We believe these changes have greatly strengthen the paper.

Regards,

The authors

Global Coastal RFA Reviewer 1 comments

Abstract

Line 45: The authors refer here to their methodology as innovative, however, there is no mention in the discussion of how the methodology from Sweet et al., 2022 was modified and possibly improved.

We have amended the abstract to reflect that this methodology is innovative in its global application – "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. "

1. Introduction

Line 113-115: while this is true, it is not a limitation that the RFA can overcome. The data used within the paper have a time constraint as well (1979-2018 for the GTSM-ERA5) meaning this is a limit that remains.

Thanks for your comments. Just to clarify, in these sentences we are just referring to the issue of TC representation in short datasets; these sentences do not explicitly refer to the RFA. However, the RFA, while not perfect in TC regions, does act to increase return levels over using site specific EVA. The RFA cannot fully overcome the limitations of the time constraints on the GTSM-ERA5 data, but it can help to address the issue in some areas. We have added the following sentence to the discussion to acknowledge that the RFA is still limited by the input data – 'The RFA aims to address this by using a space-for-time approach, however it is still limited by the bounds of the GTSM-ERA5 data.'

2. Data

Line 211: what is the spatial resolution of the HYBRID-CNES-CLS18-CMEMS2020 MDT dataset?

The resolution is 0.125° x 0.125°. The manuscript has been edited to include this.

3. Methods

Line 243: can the authors explicitly describe what areas are "unsuitable" for the RFA? Although these areas are explained later in the methods, I suggest the authors give some hints already to better understand Figure 1 as well. Another suggestion would be to include the wording "unsuitable" and "suitable" in Figure 3 for coherence with Figure 1, although the colours are already a good indication.

We have added a short definition of an unsuitable area, as follows: "(because there are less than 3 gauges in a region, or the regional water levels are heterogenous)"

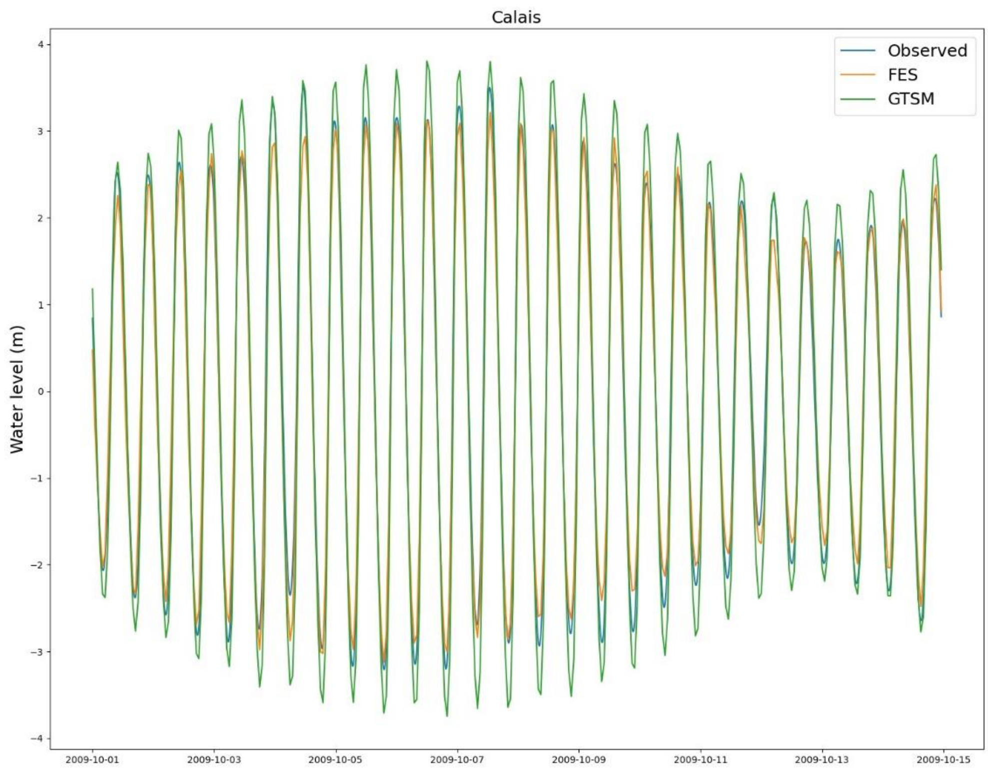
Line 256: were the authors applying a quality check to detect spikes, wrong values, large data gaps, and vertical datum issues?

We only applied the GESLA quality checks. The manuscript has been amended to reflect this.

Line 262: "the FES2014 tidal elevations performed better than those of GTSM": how was the comparison between GTSM-ERA5 and FES2014 tidal elevations performed?
Were they evaluated against tide gauge data or was it proved in previous studies?

We carried out work to verify the quality of the tidal component of GTSM, when some results of the RFA were found to be different from the single site analyses. We verified the tidal timeseries at a number of tide gauges around Europe, by plotting them against in-situ observations and FES (an example is show below for Calais). We have updated the manuscript to briefly explain how we came to this conclusion.

The authors of the GTSM-ERA5 dataset also concluded in their paper "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 an example of the overprediction of the tidal range at Calais, France.



We have updated the manuscript to include these additions – "The surge component of GTSM-ERA5 at each record location is isolated from the water level timeseries using a tide only simulation and *superimposed upon a tidal timeseries created with FES2014, as the FES2014 tidal elevations performed better than those of GTSM in initial testing against in-situ observation. The decision to use tides from FES2014 is further supported by the conclusion from Muis et al., (2020), in which they state "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." "*

Line 267: Estimating the wave set up as 20% of Hs is indeed very common and convenient and probably the most appropriate choice given the global scale of this application. However, there are limitations to this method and considerations to be made. There needs to be a mention here or in the discussion of the fact that this is a rule of thumb that does not capture the influence on wave setup of the specific coastal geometry, bathymetry and local waves and wind conditions.

This is a fair comment. We have included the following to acknowledge and justify this simplification, and also to acknowledge the simplification highlighted in the comment below. - "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)."

Line 271: The statement suggesting that wave setup is lacking in sheltered areas, such as bays and estuaries, needs clarification. While it might be necessary to make a simplification and consider no wave setup in sheltered areas there are limitations in doing so by only looking at the topography of the coastline. Wave setup in bays and estuaries may indeed be lower compared to exposed coastal regions but is strongly influenced by the specific orientation, shape, and depth characteristics.

This has been amended to "Wave setup is assumed to be absent in sheltered areas (e.g., bays and estuaries)", along with the amendment made in the comment above, which highlights that this approach is a simplification.

Line 282: if a classification of sheltered/exposed coastline was already done why the authors why are the tide gauges "assumed to be located in sheltered regions" and a check was not performed?

The classification of the coastline with respect to wave setup is a first order approximation. When assessing the tide gauges against our classification, we find that over 75% of the gauges are located in areas which we have classified as sheltered. However, as this classification only looks at coastline shape, and not local barriers or defences, we still make the assumption that all tide gauges are located in areas sheltered from wave setup. Historically, tide gauges were positioned inside harbours or sheltered estuaries to be protected from the impact of waves.

Line 292: Why is there an upper limit to the amount of water level records considered in a 400km radius area ("maximum of 10")? Was that the maximum number observed? There is no explanation or sensitivity analysis of how such thresholds were defined.

We chose a maximum of 10, because is the same threshold as was used in Sweet et al. 2022. We have updated the manuscript to make clear we have used the same threshold as Sweet et al.

Line 306: if 836 tide gauges were used in the application of the RFA why in "3.1 Data processing" do the authors state that "A total of 2,223 tide gauges with a mean record length of 21.4 years were used in the RFA."?

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."

Line 308: In Fig. 2A the "example grid cell" colour doesn't match the actual colour due to overlapped layers. Would suggest changing the colour in the legend to match that of the cell.

We have adjusted this.

Line 321: There is no mention of the reason why in some areas the density of record locations from GTSM-ERA5 is too low (which I suppose means a minimum of 3 points?). Especially in Europe where the resolution is even higher (1.25km, as visible in Figure 1 from Muis et al., 2020) why are there green regions (single-site GTSM-ERA5 analysis) in Figure 3?

*In these instances, the RFA fails because of the heterogeneity of water level records. This was especially prevalent around Italy but also occurred in locations in the Baltic Sea. It is also worth noting the resolution of the publicly available GTSM-ERA5 data is much lower (10km in Europe, 50km elsewhere) than the actual model resolution. We have added the word heterogeneous to the manuscript to improve clarity – "… the density of **heterogeneous** record locations from GTSM-ERA5 is also too low…"*

Line 330: Why 19 years? Is it the minimum number of years available from the tide gauges? Is this vertical datum adjustment applied to GTSM-ERA5 data (which only extends until 2018) or only to tide gauges?

19 years is used as it covers the 18.6 year lunar nodal cycle which impacts mean sea level. For tide gauges with less than 19 years the longest continuous period of water level records is used. This vertical datum correction is not applied to GTSM-ERA5. GTSM-ERA5 is referenced to 1986-2005, and so a correction is made (using a linear trend extracted at each record location) to bring the reference period in line with the one used for tide gauges (centred on 2011). An edit has been made to the manuscript to correct this – "GTSM-ERA5 records are referenced to MSL over the period of 1986-2005, and so the timeseries are linearly detrended to reference the same tidal epoch as the tide gauge records, centred on 2011."

Line 338: empty line.

We have removed this.

Line 340: The implications of choosing a unique value of 4-days for the declustering around the global coastline must be addressed, as Haigh et al., 2016 analysed events in the coastline around the UK. The effectiveness of the 4-day window may vary in regions with different climates, meteorological patterns, and coastal configurations and it is important to acknowledge such limitations (see the chapter from Harley 2017 in Coastal storm definition, "Coastal storms: processes and impacts", for the meteorological independence criterion variability). Was a sensitivity analysis performed to understand how the declustering method affects the resulting return period curves? Moreover, what do the authors mean by "moving window of the storm"? How is a storm defined?

The 4-day storm window is the same as was used in both Sweet et al. 2022 and Sweet et al. 2020, which covered the US and Pacific Basins respectively. The results of a sensitivity analysis would be interesting, but we think it would be a large analysis and beyond the scope of this study to include. We have amended the manuscript to include that the 4-day storm window is also used by Sweet et al 2022 and Sweet et al. 2020. The 'moving window of the storm' refers to the storm window, i.e. the 4-day period over which a storm impacts the coastline. We agree the clarity of this can be improved and have made edits to this effect. – "This window length was used by Sweet et al., 2020 and Sweet et al., 2022, and is a similar length to the storms that cause surge events in the UK (Haigh et al., 2016)."

Line 341: "This window is selected…"

We have addressed this above.

Line 342: No reference for estimating the index flood as the 98th percentile of the declustered daily highest water levels?

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

Line 365: By applying a fixed threshold without local considerations isn't there the risk to exclude exceptionally extreme events from the analysis? Was a sensitivity analysis performed on this value? Once again, extreme water level events can exhibit considerable variability based on the local characteristics of the region, including bathymetry, coastal morphology, and storm dynamics. Applying a fixed threshold globally may not capture this local variability effectively. In fact, both the Gulf of Mexico and Japan are subjected to extreme weather events and storm surges.

Further work was completed on the results from the RFA to identify grid cells that contained unrealistically large EWLs. This work identified regions in the US and Japan that had estimated 1-in-1000-year water levels of over 50m, and subsequently led to the decision to use a limiter on the *shape parameter of 0.35. 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 acknowledge that using a static limiter is a limitation of the study and updating this part of the methodology is a priority for future work on this method.*

To reflect this, we have added the following to the discussion – "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."

Line 380: "The index u is then estimated…". First, the u index is estimated, and then the LEWLs can be estimated as well. The order of the sentences must be inverted.

This has been corrected.

Line 398: Would substitute "Q99 tidal elevations" with 99th percentile for coherence with line 343 "98th percentile".

This has been amended.

Line 394: It isn't clear nor mentioned why a bias correction is needed.

An increase in the high frequency return levels was observed in our results as a consequence of the regionalisation process. This is similar to the results of Sweet et al. 2022 and Sweet et al. 2020, as well as Bardet et al. 2011. In our study, we took the view that the single site EVA was accurate in characterising the high frequency return levels, and therefore a correction term should be applied so that the high frequency return levels from the RFA are aligned with the high frequency return levels of a single site analysis.

We have added the following to the manuscript – "Other surge RFA studies also concluded that the approach generally yields higher estimated surge heights when compared to single site analysis, because during the regionalisation process an extreme event that occurred in one location is assumed to have the same probability of occurring at another location within the homogeneous region. (Bardet et al., 2011; Sweet et al., 2022)."

Line 420-423: I would move this part of the methods in the "2. Data" section as it describes how COAST-RP was derived.

We have removed this as it is already described in the Data section.

4. Results

Line 487: the example of cyclone Yasi wasn't under sampled in the historical record. Two gauges out of 10 recorded it because those were at the impacted locations. I would remove line 485 to 487 from here.

We have removed this.

Line 503: I do not agree. The Cardwell tide gauge is in a unique location compared to the rest of the gauges in the area, as it is located at the back of a semi-enclosed bay and at the northern mouth of the Hinchinbrook Channel. The specific shape and morphology around the Cardwell tide gauge can contribute to the funnelling and amplification of the storm surge on the left side of the storm track (where the onshore winds push water toward the coast). Moreover, not "any local effects due to surge" can be accounted for using the index flood but only those variations that are within the scale of resolution of the analysis. That would correspond to 1 km when considering the interpolated points, but 1.5/2.5 km if considering instead the original resolution of GTSM-ERA5 (which is using bathymetric data from EMODnet and GEBCO therefore even lower resolutions). It is important to clarify the spatial scale of the application of extreme water level values derived from global models, as overconfidence in the results can lead to errors at the local level.

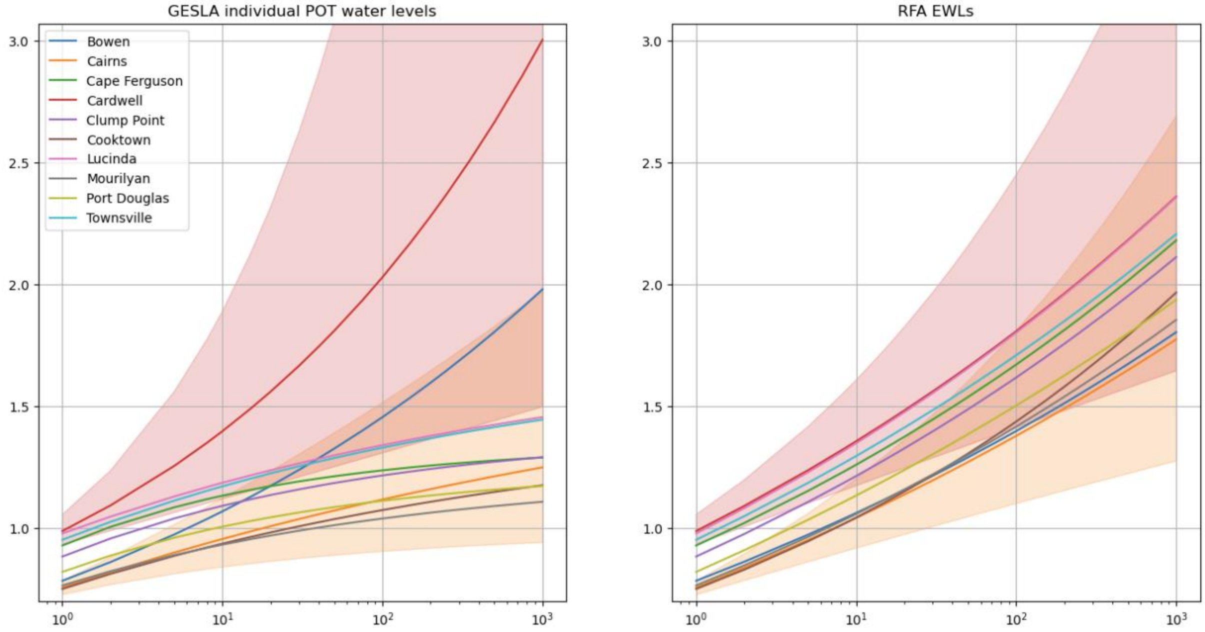
We have removed this statement. However, it is worthing noting that the water levels are normalised by the index flood, u, which is the 98th percentile of daily highest water levels of the tide gauge in question. At Cardwell, the index flood is 10cm above the average for all the gauges in the area, so the normalisation process would help reduce the local effects of the surge compared to nearby gauges. We agree that the resolution of the FES2014 (from which the tidal range is used to downscale the regional water levels) is important when considering the confidence of results at a local level, and we have added a statement to this effect in the discussion – "This downscaling process is, however, limited by the resolution of the tide model used to obtain the tidal range values. In the case of this study, FES2014 is output at 1/16th of a degree (approximately 7km at the equator).".

Line 519: I would adjust for coherence the plot in Fig. 5D and 5E so that the return periods on the x-axis extend from 1 to 1000 (see Fig. 4C and 4D).

We have adjusted this.

Line 520: To demonstrate the increase in robustness from the RFA would have been interesting to include a mention of the confidence intervals, how do they compare with those from the individual tide gauges?

We have compared the confidence intervals of individual tide gauges with the confidence intervals from the RFA. Please see the figure below. We only show the confidence intervals for two tide gauges (Cardwell and Cairns) for clarity. For tide gauges that have a positive shape parameter, we see a decrease in the spread between the 5th and 95th percentile. Interestingly, however, the spread between the 5th and 95th percentile for tide gauges in this region that have a negative shape parameter is larger using the RFA compared to single site analysis. An explanation of this would be that the regional EWLs propagated from the RFA contains the EWL hazard observed at Cardwell, and so can be assumed to be inclusive of the risk of TC generated surges. We'd therefore expect the *uncertainty bounds to increase, as we are moving from a regime that does not incorporate TC risk, to one that does.*



Line 538: The results shown in Fig. 5E show an increase in the return period curves and therefore a "spread" of the extreme event due to the regionalization process. However, the Cardwell curve significantly decreased and the information relative to the 3 m extreme event was lost. The authors should not only reinforce the strength of the RFA in propagating the risk in the entire region considered but should discuss a) the fact that at Cardwell and Clump Point, there is an underestimation (or "smoothing out", as the authors wrote in Line 559) of the extreme water levels from the return period curves estimated with the RFA approach and b) the possible negative implications of increasing the return levels at the other gauge locations. While RFA is a valuable tool, it is crucial to consider its limitations and potential challenges, especially when interpreting results for coastal management and safety assessments. An overestimation of extreme water levels can have several implications, potentially resulting in unnecessary costs for risk mitigation measures and infrastructure design that often needs to be optimized.

We have added the following sentence to the paragraph on Cyclone Yasi in the discussion – "On the other hand, the damping of the return levels in the RFA output at Cardwell and Bowen could mean an underprediction of the risk from surges in these locations." Furthermore, we've added a more general paragraph to the discussion covering a number of other suggestions. See the paragraph below.

"The outputs from the RFA should be supplemented with local knowledge wherever possible, and the uncertainties in the results should be considered before the data are used. The RFA is a powerful tool for estimating return levels in ungauged locations or in locations where the historical records are short or incomplete, but there are risks associated with both overpredicting and underpredicting surge heights. Underprediction can lead to complacency among coastal managers and the potentially dangerous assumption that communities are safe from surge risk. Conversely, overprediction can result in unnecessary cost for risk mitigation measures and potential economic loss driven by a lack of investment in a region deemed at risk. Disseminating the risk of TC generated surges over a region could lead to overprediction in some locations, and so conducting sensitivity analyses to understand

the robustness of findings is recommended, especially in the context of coastal management and safety assessments. The RFA has been developed in this study as a method for regional to continental to global scale risk analyses from globally available data, and not local studies. The results give a first order approximation of extreme water levels in ungauged locations. It is not expected that they would be used in the design for local flood defences, for example."

Line 555: After looking at the example in Fig. 5, I would say that the increase in the RFA is not only related to the under sampling of rare surge events. During cyclone Yasi there were 10 tide gauges in the area actively recording and only two of those were impacted by the rare surge event. Therefore, the fact that in Fig. 6A there are positive differences is associated with the RFA regionalizing extreme events. The wording "under sampling" gives an idea of a lack of in situ measurements or short records, but the GTSM-ERA5 data are also available for 40 years only and with 1.5/2.5 km resolution. I would write a disclaimer about the fact that extreme events are under sampled because of their nature, they are rare and might have never occurred at all at a specific location apart from not being recorded because of the scarcity of in situ tide gauges.

Thanks for the comment. A disclaimer has been added as follows – "Extreme surge events can be undersampled for two reasons. Firstly, by their very nature, they are rare and might never have occurred at a specific location. Secondly, as a result of a scarcity of in-situ tide gauges, surges can occur and remain unrecorded."

Line 557: Why were the authors not assessing the impact of the limiter on the shape parameter? It is a relevant difference whether the decreases observed in the return levels are associated with a threshold that could be modified or rather by the regionalization method itself.

The locations of the gauges which lie in grid cells with limited shape parameters have been identified and the paper has been updated to reflect this with the inclusion of this sentence – "Of the gauges shown in the Fig. 6A, only 5 had limited shape parameters, which were located in the Gulf of Mexico."

Line 573: Here and along the paper I would use "quantification of the increase" or quantification of the incorporation of tropical cyclones as their associated risk is in fact better captured. Assuming that an increased extreme water level exceedance probability by the RFA is an improvement is a step further that needs to be discussed. The authors assume from the Introduction onwards that increasing the ESL exceedance probabilities through the RFA approach is an improvement before discussing the implications of this assumption. Just because a region could possibly be affected by e.g. tropical cyclones it doesn't mean that the solution to account for that is to have an increased return period curve of extreme water levels.

We have accepted this suggestion and made edits to all the references to "quantifying the improvements", as well as adding a sentence in the discussion – "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."

Line 583: Why is there an oversampling of extreme events and how was that checked?

This statement is incorrect. We can't be sure that it is in fact oversampling that occurs here. We have amended this sentence to "Sporadic negative differences are also observed in Fig. 7A, which are driven by a smoothing of ESL exceedance probabilities at locations which have experienced anomalously high ESL compared to the local region".

Line 615: For consistency use under sampled not "under-sampled".

We have updated this.

Line 620: In Fig. 8C I mostly see an increase (in some cases by 50%) in ESL exceedance probability from the RFA when compared to the single gauges, with the distribution being overall positively biased (overestimation of extreme water levels from the RFA).

We know the RFA underestimated the water level when compared to a single gauge analysis from looking at the data. It is difficult to discern this from the figure as many points are overlayed and the magnitude of the decreases are comparatively small compared to the magnitudes of the increases.

Line 651: The water level at the gauge is not mischaracterized, that is an in-situ measurement. What can be mischaracterized is the overall extreme event in case there aren't tide gauges in the location where the maximum water levels are reached.

An amendment has been made – "As a result, measured records can easily miss the maximum of an extreme event, thus mischaracterising extreme water levels ***of the event****"*

Line 652: Change "undersampled" to "under sampled".

We have made this change.

Line 654: I think would be more correct to state that the issues were improved, not overcome.

We have updated as suggested.

Line 664: While it is true that a single site analysis of tide gauge data can underpredict the regional risk of ESL generated by tropical cyclones, overpredicting it has also its drawbacks. Please discuss the implications of disseminating the hazards over a larger area.

This has been included in a more general paragraph added to the discussion. Please see the response given to Line 538.

Line 668: "high degree of accuracy"

We have corrected thanks.

Line 677: In addition, Hudson Bay showed an unexplained increase in extreme water levels that does not reflect any cyclone activity.

The largest increases in water levels in Hudson Bay are seen in the COAST vs GTSM-ERA5 comparison. As we didn't have access to the full COAST dataset, we were not able to look into why this occurred. As the area is not subject to cyclone activity, we theorised it could have been a result of changes to tidal dynamics between COAST and GTSM-ERA5. There are some locations in the GTSM-ERA5 RFA results which also show increases in areas which aren't impacted by cyclone activity. However, in the original GTSM paper, the authors state "Output locations in high-latitude regions are excluded because the model performance is expected to be insufficient. This is because GTSM does not include sea ice physics and because the bathymetry in the Arctic areas is generally poor." The increases here are likely driven by inconsistencies and anomalies in the GTSM output. As flood risk is generally low in these areas, identifying the cause of these increases was not a priority, but may be included in a future study.

Line 681: While RFA is designed to handle short and incomplete tide gauge records, the reliability of the analysis is still contingent on having a sufficient number of extreme events for robust parameter estimation. In regions with very limited data, uncertainty in estimates may be high. However, while it is true that GTSM-ERA5 has a temporal limit, saying that "As such, the RFA has little basis upon which to draw data from when characterising rare extreme events." seems in contrast with what was previously said:
"672 [….] Once again, the RFA provides a solution to this problem. As demonstrated in
673 Fig. 7, the distribution of increases to local return levels made by the RFA broadly follows
674 the same patterns globally as the differences between COAST-RP and GTSM-ERA5. This
675 highlights the ability of the RFA to characterise tropical cyclone hazard which is typically
676 underrepresented as a result of short records." Please consider rewriting this section to avoid contradictions when exploring the limits and advantages of the RFA.

We have taken this onboard and rewritten the latter section with a focus on the inability of ERA5 to capture the full intensity of TC events, and therefore the impact this can have on surge modelling. The paragraph now reads:

"While the RFA is capable of identifying areas of increased risk from tropical cyclone TC activity, it is still constrained by the training data available. This is demonstrated in Fig. 7. Two distinct areas lack increased water levels in the RFA difference plot (Fig. 7A), namely: the Bay of Bengal and Northwestern coasts of India and Pakistan. ERA5, the forcing data used for GTSM-ERA5 has been found to consistently underestimate TC intensity in both minimum sea level pressure and maximum windspeed (Dulac et al., 2023). Consequently, the intensity of extreme events in GTSM-ERA5 in these regions could underrepresent the potential hazard from TC activity. If the maximums of extremes are not captured in the reanalysis data, then the full magnitude of the surge cannot be simulated by GTSM-ERA5. As such, the RFA will have smaller or fewer extremes with which to draw data from when characterising rare extreme events, therefore leading to a persistent underestimation of the return levels."

Line 683: The term “tropical cyclone” was used 41 times in the manuscript, consider adopting the commonly used “TC” abbreviation.

Thanks for the comment – we have made this change as you suggest.

Line 698: The downscaling process assumes that the extreme event observed at one location can be used to estimate the risk for nearby areas. This may be less reliable in regions with diverse coastal characteristics and complex topography. At specific coastal areas a much finer resolution analysis than the RFA is needed before applying inundation models. It should be made clear that it is advisable to supplement RFA with local knowledge, consider uncertainties in the results, and conduct sensitivity analyses to understand the robustness of the findings, especially in the context of coastal management and safety assessments.

This limitation is mentioned in the discussion when discussing using 1° by 1° tiles, and the suggestions of supplementing the RFA with alternative data has been added to a more general paragraph. Please see the response for Line 538.

Line 721: This is why can be useful to have an additional plot (e.g. in the Appendix) that could complement Fig. 4A and 4B with confidence intervals or with some sort of indication of the amount of data that was used for the RFA.

Thank you for the suggestion, we have added the following figure and caption to the appendix as requested.

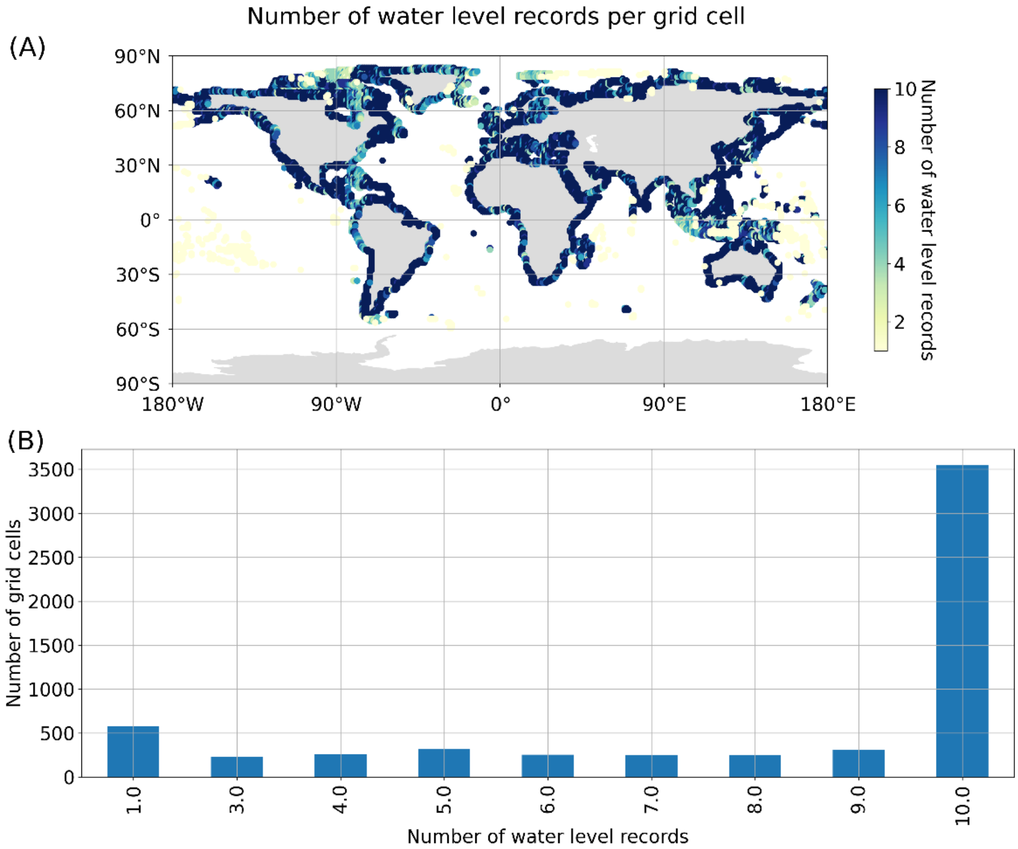


Figure A1: The number of water level records used per grid cell (A) as a scatter plot showing the distribution globally, and (B) as a bar plot showing the number of water level records vs the number of grid cells.

Line 734: There is a wave buoy dataset distributed by CMEMS that integrates observations from around the world: the "INSITU_GLO_WAVE_REP_OBSERVATIONS_013_045" at https://doi.org/10.17882/70345).

Thanks for the information.

Global Coastal RFA Reviewer 2 comments

<u>Abstract</u>

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).

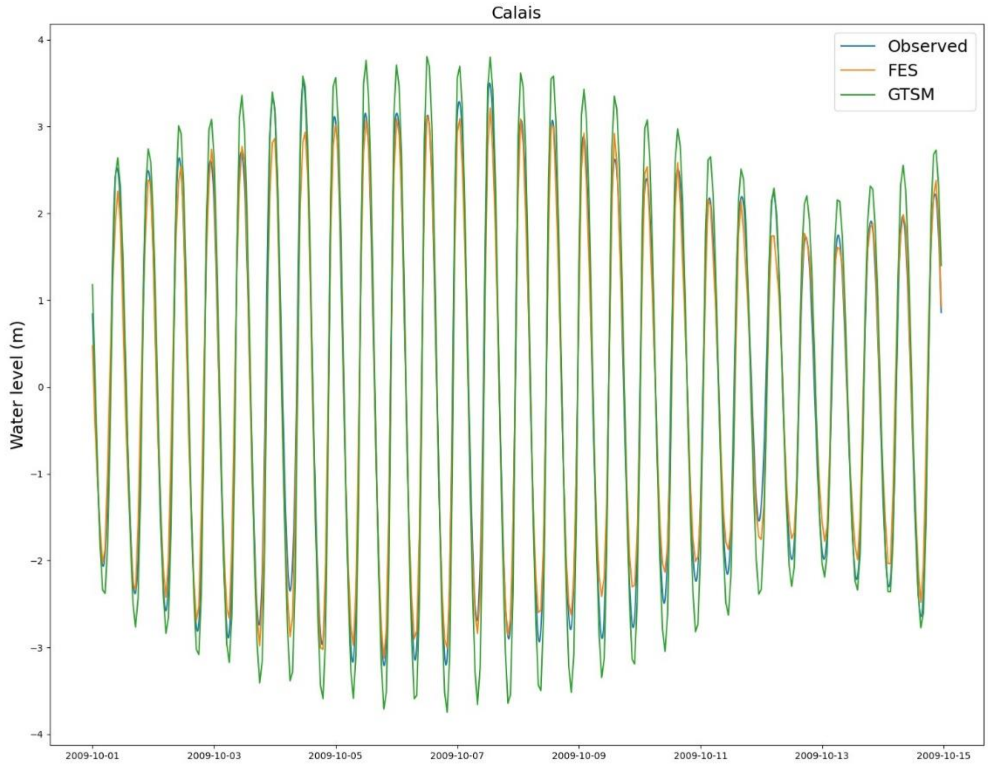
Each grid cell is considered independently, and water levels are declustered around a 4-day storm window. In the example given here, a large TC event could cross over the Florida peninsula and drive a high water event on both the east and west coasts, however this would simply be interpreted as a single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

METHODS

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.

Lines 270-271: "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

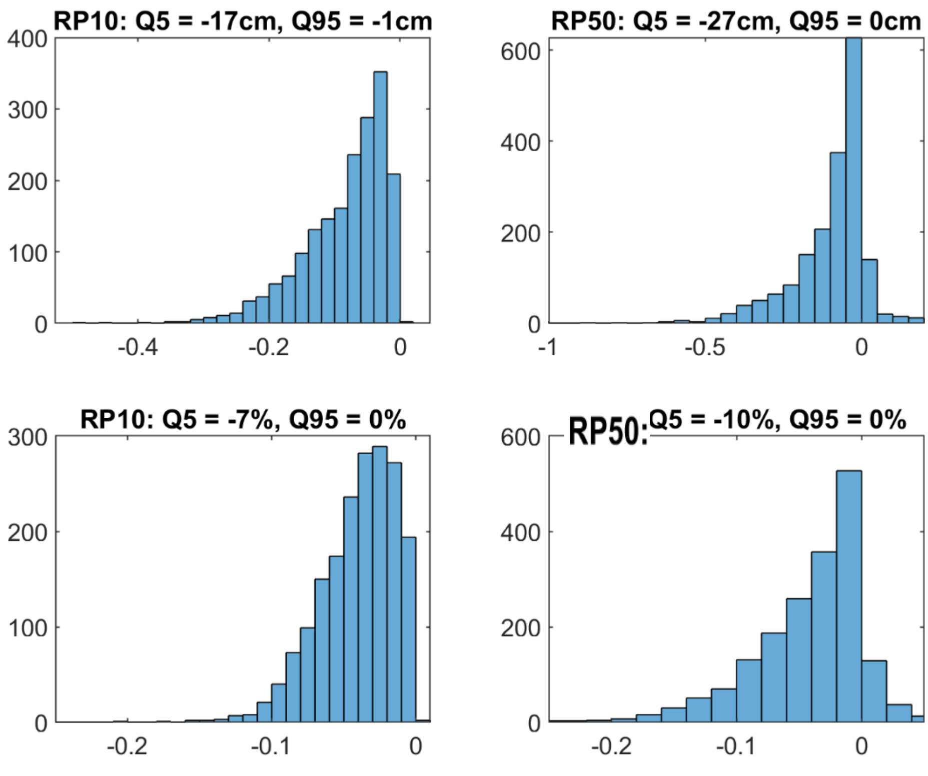
We have changed "interpolated" to "assigned".

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.