Detailed commentary

L20: remove 'that'

L25-28 : wat do you mean by linear approximation? In the 2016 reanalysis, tides where also linearly added to storm surges as far as I know

L36: "The GTSM-derived timeseries…" you mean the reanalysis in particular? More generally, you mention 'The dataset' often in this paragraph but it's not clear which dataset exactly you refer to. For example, you mention the study of climate change (Muis et al. 2023a), but this study uses projections not reanalyses. Please be more specific, and focus on the reanalysis dataset if that's the intention (e.g. as in L42 "The GTSM-ERA5 reanalysis…").

L40-42 "*A recent assessment on the drivers of shoreline changes of the global coastline shows that despite clear changes in storm surges, there is no clear link with shoreline changes (Ghanavati et al., 2023)*". This phrase seems out of place here, as this study doesn't focus on shoreline changes. I guess the idea is to highlight an application of the dataset. It should just highlight that the dataset has been used also for studying the potential link to shoreline changes, or something like that?

L51 "*Different methods, such as annual maxima and peaks-over-threshold have been applied*". You mean different methods have been applied across the different GTSM-based reanalyses and associated publications? You could be more explicit about the logic behind choosing one or the other for the different datasets

L57 "*The current GTSM-ERA5…*" the latest? Please add the reference.

L57-58: You cite *Calafat et al. 2022* but this study looks at long term trends not decadal variability? If you explicitly want to highlight internal variability of extremes, you should cite studies such as *Lobeto and Menendez (2025)* https://doi.org/10.3390/rs16081355

L96: Sea level rise: The study Muis et al. (2020) is cited, but I believe it should be Muis et al. (2023)? In the former, no SLR is applied for historical periods I believe. The information of the fields based on Le Bars (2018) appears on the SM of the 2023 publication.

Regarding SLR, in Muis et al. (2023) it was argued that the CMIP5-based estimates were used because unavailability for CMIP6 at the time of the simulations. I guess this was not the case anymore this time round. I know this only concerns the last years of the reanalysis but wondering if we can expect a big difference with CMIP6, in any case the choice should be justified (e.g. for continuity with the previous reanalysis dataset?).

L134: What is the rationale behind imposing a yearly SLR field if this is removed in the post-processing? To capture non-linear interactions? This should be justified and clarified.

L139: what method do you use to derive the confidence intervals?

L188: in average across stations. In this paragraph, you should also highlight the comparison in terms of standard deviations.

Section 4.1:

I think the authors should also show the quality for the backwards extended period alone, despite the lower density of observations available. In particular when comparing to GESLA, we need to understand how much the backwards extended period is weighting into the statistics. I think this is crucial to understand the quality of the extended dataset, whose biggest asset is precisely the backwards extension, and to justify claims such as that in line 192 "*This proves that the backward extension of the ERA5 dataset for the period 1950-1978 is well suited for use in the modelling of global water levels*".

Another general comment for this section is that, especially for validation, I think providing the metrics for storm surges is crucial, since this is the part of the waterlevel that will be influenced by the atmospheric forcing, and hence will reflect the quality of the ERA5 dataset. Tides are deterministic and we should therefore have the same quality for the extended period. In places with substantial tides, tides will be dominating many of the metrics shown (correlation, RMSE, even high percentiles such as the 95th percentile, as highlighted in the text for some estuarine locations).

Figure 2: Please increase the resolution. **Please assign a label to each subpanel,** and refer to this in the text, the results description is otherwise difficult to follow. Why do we see points in the first figure (upper left) that don't correspond with the point in the figure right below? E.g we see a blue dot around Sao Paulo in the second plot, but there's no dot on the first plot. **Please remove also the last row of plots (RMSE and correlation of monthly maxima) as these are not cited.**

Section 4.2:

Figure 3: It would be useful to have the relative differences, at least in supplementary materials, as 0.2m difference is very small in places with large tidal range but can be very large in other lower sea level range locations.

It is also not clear the logic of the comparisons, and it's not clearly described in the text. First, the 95th percentile of 2 periods of equal length are compared (first vs last 20 years). why? What is the objective? It is stated at the beginning of the section (L224) '*to understand the impact and potential benefits of the longer timeseries'* but the link to the comparison between periods is not explicit. This should be explained. Then, for the 100-yr return level, different periods are compared (original vs extended period, more in line with the stated objective), but it is shown in the same plot as the 95th percentile comparison. Please explain the logic of the different comparisons, and consider using separate plots if periods being compared are not the same, or clearly state in figure titles.

Regarding the 2nd 30-year period (1990-2019), why not use the very last 30 years (1995-2024)?

Regarding the 100-yr event differences, why do we see quite large differences in tropical regions but statistics of the 100-yr event error relative to observations average to the same value between original and extended datasets? (Table 1).

L242: You compare the 95th percentile in Figure 3, not the 99th percentile as stated in the text.

L245-247: Please have a look at Lobeto and Menendez (2025) (link shared in previous comments) to understand if climate modes could be playing a role.

Figure 3: again, what is the objective of this comparison? Why not also look at the 100-yr point estimate differences between original and extended datasets, as for waterlevels before?

Perhaps results would appear more clear if all plots showing the 95th percentile changes are shown in one figure (e.g. waterlevels on the left, surges on the right), and describing them in the same paragraph which should start with the objective of the comparison between the 30-yr timeslices (e.g. highlight temporal variability in extremes that is 'hidden' when typical 30-40 year periods are used?). And then put together the 100-yr plots too, and corresponding descriptive paragraph (with its own different objective, e.g. understanding the impact of sampling uncertainty which is reduced for longer periods?). You could add the difference in confidence intervals as a 3rd row to this plot.

L256-258: "*On the other hand, an increase in the width of the confidence intervals is observed at clusters of locations in tropical cyclone regions. This is expected, considering that the sample size for those regions is too small, even with the ERA5-E dataset.*". Why do we expect an increase relative to shorter periods? Please elaborate

L270 "*higher long-term extremes*" what do you mean by this? Clarify

In relation to the analysis in Figure 5, it would be interesting to see in tide-gauges with long and trustworthy records how the extended dataset may help reduce (or not) the error in the calculation of return levels. The question is: is the longer dataset really more reliable, or longer but of poorer quality in the extended part such that return levels are negatively impacted? Are the additional events in the dataset well captured relative to observations? In the validation presented in section 4.1, as highlighted before, no specific validation for the extended period 1950-1979 is provided, and it is also not shown how the error in the 100-yr event changes between original vs extended dataset, the authors only provide the global average return levels in Table 1 and only the error of the extended dataset is shown in Fig 2. Given results in Fig 3, we can expect a relatively important impact in some tide-gauge locations. I therefore suggest showing the impact on the return level curves for a few locations with long records, together with earlier validation analyses.

This is partly covered in section 4.3, where the performance for historical events in the extended periods is analyzed. However, I find that the analysis focuses too strongly on tropical cyclone events. ERA5 already does a poor job in representing these in the original, more recent period, so it's a systematic skill limitation of ERA5 and not a specific performance issue for the extended period (e.g. from the reduced assimilation of observations). The quality for the extended period needs to be more thoroughly analyzed and presented. Finally I suggest to move the analysis of these specific events to the section 4.1, or to a subsection of that, and keep current section 4.2 as the last section, which focuses on highlighting the different possible impacts of the longer dataset on the estimation of extremes.