Manuscript title: ***Global dataset of storm surges and extreme sea levels for 1950–2024 based on the ERA5 climate reanalysis***

Authors: N. Aleksandrova, J. Veenstra, S. Muis

We would like to thank the editor and the reviewer for taking the time to review and process the manuscript. We address each point raised by the reviewers below (the text by the reviewer is in blue). The line numbers and figure numbers in the responses refer to those in the revised manuscript.

Response to reviewer #1

The article presents a 75-year global dataset /1950-2024) of coastal water levels and storm surges, generated using ERA5 reanalysis and consistent with the earlier 1979–2018 product. The extended record enables improved long-term statistics, more robust extreme-value estimation, and detailed reconstruction of historical storm events. Validation using GESLA observations and three past severe storms demonstrates generally reliable performance but also highlights systematic limitations. ERA5 underestimates the intensity of extreme storms especially tropical cyclones leading to underestimated surge heights. The global GTSM model cannot fully resolve complex coastlines or local bathymetry, contributing to discrepancies between modelled and observed extremes. As a result, the dataset's extreme value statistics are best suited for first-order global assessments rather than local engineering applications. Downscaling and mean sea-level reconstructions can further improve regional accuracy.

The provided dataset complements the previous product provided by the same authors in 2020 with a significant extension of the period covered giving better estimations of long term extreme-value statistics. Limitations of the product are well identified and well indicated in the concluding remarks of the manuscript. This is a good example of a paper suitable to be published in ESSD.

Thank you for your positive evaluation of our manuscript. We appreciate the reviewer's assessment and are pleased that the manuscript meets expectations.

Response to reviewer #2

The paper presents an extension of a still water level hindcast from 40 to 75 years, made possible by the recent backward extension of the ERA5 atmospheric reanalysis. Its main objective is to assess how this longer dataset affects the estimation of extremes, which are often poorly constrained when derived from shorter records, particularly for long return periods. The extended hindcast represents a valuable resource for the community working on extreme sea level estimation, and the dataset itself constitutes a contribution worthy of publication. However, the manuscript requires substantial revisions to better validate the extended period and to more clearly demonstrate the implications for extreme value estimation. In particular, the results in Section 4 should be reorganized so that all validations against observations—including comparisons for individual events—are presented first, followed by a section dedicated to assessing the impact of the extended record on extreme value estimation (e.g., temporal variability and sampling uncertainty). Moreover, the validation against observations should be strengthened by including metrics specifically targeting the backward-extended period (1950–1978) and by focusing on selected long-record tide gauges, both within and outside tropical cyclone regions, in order to evaluate the added value of the extended dataset for better constraining return levels.

We thank the reviewer for their thorough and constructive assessment of our manuscript and for recognizing the value of the extended dataset.

In response, we have substantially revised the manuscript to address the points raised by the reviewer. Most importantly, we have improved the validation analysis by two major changes: 1) instead of using percentiles, we validate on monthly and annual maxima; and 2) instead of comparing the validation for the full 1950-2024 dataset vs. the original 1979-2018 dataset, we present validation for the extension-only (1950-1978) and compare it to a period without the backward extension (1979-2020). As result of this, sections 4.2 and 4.3 are restructured, Table 1, Figures 2, 3, 4 and 6 are updated, to more clearly: (1) validate the quality of the pre-1979 GTSM-ERA5 data against observations, and (2) highlight the impact of using longer time series on extreme value estimation. In addition, we revised the manuscript to clarify any ambiguities that were identified by the reviewer.

We believe these revisions significantly improve the manuscript and directly address the reviewer's comments.

Detailed commentary

L20: remove 'that'

Agree, it is removed.

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

In this sentence, we refer to datasets of water levels derived using simplified parametrizations, as opposed to dynamically resolving hydrodynamics using the a numerical model. Muis et al. (2017) provides a comparison between GTSR and one such dataset, DCESL (Vafeidis et al., 2008), where storm surge and tidal components are estimated separately, with storm surge estimated on the assumption of uniform beach slope and equilibrium conditions with a constant wind speed. To clarify this, we changed this sentence as follows:

L26: "The hydrodynamic approach has a higher accuracy than previous global return periods of extreme sea levels that used simplified parametric models to estimate storm surges (Muis et al., 2017; Hinkel et al., 2014). "

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…").

We agree that the paragraph needs to be rephrased for more clarity. The cited studies illustrate the applications of the GTSM-derived reanalysis water levels, both the older GTSR (based on ERA-Interim) and the more recent GTSM-ERA5. The new text is as follows:

L38: "GTSM timeseries of still water levels generated using climate reanalysis data, presented in the GTSR (ERA-Interim) and GTSM-ERA5 datasets, have been widely used in coastal hazards research. For example in the analysis of individual historical events both in terms of the height of water levels (Dullaart et al., 2020) and their impact (Koks et al., 2023). They have also been used to investigate the event footprint and spatial dependencies (Enriquez et al., 2020; Li et al., 2023), as well as the influence of climate variability on surge levels (Muis et al., 2018) and in comparisons with GTSM driven by climate models (Muis et al., 2023a)."

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?

We agree that the main point of referring to this citation is to highlight that application of the dataset. This sentence is rephrased as:

L43: "GTSR data was used to study the potential links between storm surges and shoreline changes globally (Ghanavati et al., 2023)."

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

This paragraph is clarified as follows:

L54: "Different methods, such as annual maxima and peaks-over-threshold have been applied in different studies: the GTSR dataset was analysed using annual maxima (Muis et al., 2016), while peaks-over-threshold method was applied to GTSM-ERA5 (Muis et al., 2023). While determining the appropriate threshold can be challenging at global-scale, peaks-over-threshold method can extract multiple peaks per year and makes more efficient use of the available data. Extreme value analysis have large uncertainties when applied in broad-scale studies (Wahl et al., 2017), but the resulting return periods are nevertheless useful for first-order large-scale assessments of coastal flood hazard and risk (Tiggelhoven et al., 2020; Lincke and Hinkel, 2018; Brown et al., 2018), including infrastructure and cultural heritage sites (Reimann et al., 2018; Verschuur et al., 2023). "

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

A reference to Muis et al. (2023) is added. Please note that even though this paper mainly focused on climate projections, it also describes the latest GTSM reanalysis dataset that was published on CDS and was extended in the work described by our manuscript.

The new sentence is as follows:

L63: "The latest GTSM-ERA5 reanalysis dataset covers the period 1979 to 2018 (Muis et al., 2023)."

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

Thank you for pointing this out. We will update the reference since we agree that the suggested reference fits better. The new sentence will be as follows:

L63: "The length of 40 years is relatively short considering the large decadal variability (Lobeto and Menendez, 2025)".

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?).

The reference should indeed have been to Muis et al. (2023). We use the same SLR dataset for the backward extension, in order to maintain consistency with the original dataset. The difference between CMIP5 and CMIP6 sea-level signal over the historical period 1950-2024 will be relatively small (Jevrejeva et al., 2021, https://doi.org/10.1088/1748-9326/abceea) since the difference mostly arise for the future projections, and the continuity with the previous dataset was considered to be more important. The sentence is rephrased as follows:

L103: "The detailed methodology behind the sea level rise dataset is provided in the supporting information for the previous publication where this dataset was utilized (Muis et al., 2023), the same sea level rise dataset is used for the extension in order to maintain continuity."

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

We are including the sea level rise fields in the GTSM simulations in order to generate the best representative water level data that includes sea level rise. It is also useful to include SLR in order to account for non-linear interactions between tides, surges and sea level rise.

When performing the extreme value analysis, we would like to have a stationary (detrended) water level dataset, which is why we remove the sea level rise from the timeseries before applying EVA.

In the validation of monthly and annual maxima timeseries, we remove the SLR from both the model and tide gauge timeseries in order to compare tides and surges without the SLR signal. For some tide gauge stations, there might be discrepancies between the measured sea level rise and the sea level rise included in CMIP5. We would like to exclude this effect from our validation statistics, to focus purely on tides and surges.

This is clarified in the following sentence in the revised manuscript:

L141: "*Then, we detrend the timeseries by removing the annual means from the still water level dataset to achieve a stationary dataset without sea level rise.*"

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

The confidence intervals are computed using the Monte Carlo method. This is implemented in the *pyextremes* Python package code, which we are using to compute the return periods. This is now clarified in the text as follows:

L145: "We use the Maximum Likelihood Estimation (MLE) to fit the GPD parameters, with the Monte Carlo method used to define confidence intervals. This analysis is realized using the pyextremes package (Bocharov, 2023)."

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

We agree with this comment, and the standard deviations across stations will be added in the text. In short, the standard deviation shows that there is considerable variability in performance across stations.

Standard deviations are now explicitly named in the text discussing the validation results:

L188: "*In general, there is a good agreement of the validation metrics between the two periods, indicated by very similar values for Pearson correlation of 0.6 for annual maxima (with standard deviation across stations being 0.24 and 0.27 for the recent period and backward extension period respectively). Also the RMSE shows very similar values for both periods. For both periods there is a similar degree of underestimating annual maxima, by 0.10-0.11 m on average (standard deviation across stations of 0.25-0.26 m).*"

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".

Thanks for the valuable comment, we agree that it is important to demonstrate the performance for 1950-1978 separately. To do this, we have decided to revise the validation analysis. The selection of GESLA stations was narrowed down to a set of stations where no more than 25% of data is missing in both the 1950-1978 and 1979-2020 periods (instead of the original requirement of no more than 25% missing in 1950-2024). After also filtering the dataset for duplicated records, we obtain a selection of 107 stations. Based on these selected records, we validate the two periods separately. The text is modified accordingly:

L150: "To validate the dataset, we compare the modelled still water levels with high-frequency observations of sea levels from the tide gauge stations in the Global Extreme Sea Level Analysis (GESLA) dataset, version 3 (Haigh et al., 2022). The validation is performed for two periods: 1979-2020 (recent period, where GESLA data is available) and 1950-1978 (backward extension period). We only use the tide gauge stations where there is less than 25% of data missing in either of the two periods, and remove duplicated stations, keeping the longest records among the duplicates. *Additionally, the data records are filtered according to the data quality flag value to only use records where no data quality issues are indicated. This selection results in 107 stations that can be used for extreme value analysis, where each station has a counterpart location in the GTSM-ERA5-E dataset (within 10km distance). The data coverage in the resulting observational* *dataset used for data validation is 96.2% for 1950-1978, and 97.5% for 1979-2020 (on average across stations)."*

The comparison shown in the new Table 1 shows that the validation of monthly and annual maxima between the 1979-2020 and 1950-1978 periods shows very similar results. On average across stations there is a similar underestimation of annual maxima. The values for RMSE and Pearson correlation are also very comparable. The text is adjusted as follows:

L160: 'Each observation record is resampled to hourly data. In order to match the vertical reference of the GTSM-ERA5 dataset, the mean water level in 1986-2005 is subtracted from each record, and daily, monthly and annual maxima are calculated. The model performance of GTSM-ERA5-E is evaluated by comparing observed and modelled values for daily, monthly and annual maxima. For each station, the model data is filtered to only include the periods where observation data is available, and resampled to daily, monthly and annual maxima timeseries. We use the mean bias (MB), the mean absolute error (MAE), and the mean absolute percentage error (MAPE) to compare the modelled and observed data. Next to that, we calculate the root mean square error (RMSE) and Pearson correlation coefficient across timeseries for each station. Mean and standard deviation of these parameters across stations provide an indication of the model performance as compared to observations. Comparing these metrics between the two validation periods (1950-1978 and 1979-2020) provides insight into the differences in performance between the original period and the backward extension. "

Figure 2 and Table 1 have been updated accordingly, please refer to the revised manuscript.

We have also added the following note to the Conclusions section, to reflect on the sparse observations available for validation and ERA5 performance:

L343: "The validation of the GTSM-ERA5-E dataset for 1950-1978 relies on long-term tide gauge records, which are geographically sparse and are mainly restricted to Europe, North America and Japan, and lack data for e.g. Africa, South America and South Asia. At the same time, the ERA5 backward extension also relies on data assimilation of long-term weather records (in the pre-satellite era, before 1979), which are similarly unevenly distributed. It is plausible that the quality of ERA5 data in terms of capturing extreme events is higher in the regions with the highest density of observations. Based on the validation we conclude that the backward extension can be used to increase the length of time series for used in extreme value analysis, although caution is advised when using this dataset to estimate long-term trends since the quality of ERA5 could improve over time and affect the trend analysis."

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).

After further consideration, we agree that the percentiles (99^{th} and 95^{th}) are still heavily influenced by tides and are not the best metric to validate the full water levels. Unfortunately, a surge-only version of GESLA-3 (or comparable dataset) spanning the full validation period is not available, and generating such a dataset of sufficient quality is beyond the scope of this study. However, based on the review, we have restructured the validation to place more focus on the annual maxima of still water levels (instead of the percentiles). This would help to assess the performance for storm surges, when comparing the results for the two validation periods.

We also added tide gauge measurements data to the three storm-specific examples in the new Figure 3 (L236, discussed at L221-235), and to the extreme value analysis presented in Figure 7 (L309, discussed at L279-L306). These examples demonstrate how the extended dataset reflects extreme events at multiple tide gauge locations where long-term observations are available.

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.

This figure has been changed to provide a comparison for annual maxima validation, instead of the percentiles (see the new Figure 2 in the revised manuscript at L218). The resolution of the figure has been increased. The figure is linked to Table 1, showing the comparison metrics for annual maxima between the two assessment periods.

Regarding the Sao Paolo station in the original figure, there was a dot on the first plot, but it is very light yellow, so it is barely visible. We have adjusted the plots to improve visibility for all stations by adding visible edges to all data points.

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.

In the revised manuscript, we are adding a metric for mean absolute percentage error (MAPE) of annual maxima in Table 1 and Figure 2, to provide more information on relative values.

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.

We thank the reviewer for this valuable comment. Reflecting on this and other related comments in the review made us realize that for the purpose of validating the extended dataset we need to make significant changes to the presented analysis. We decided to omit the percentiles and focus on monthly and annual maxima, which reflects better how the GTSM-ERA5 datasets represent storm surge extremes. In the revised manuscript, we updated Table 1 and Figure 2, and change the accompanying text accordingly. The structure of Sections 4.1-4.2 has also been revised, to first present the validation of monthly and annual maxima based on the available tide-gauge records in Section 4.1, and then to compare surge statistics between periods in Section 4.2 globally, followed by the analysis of how the extreme values and confidence intervals are affected by using the extended dataset. The latter is additionally illustrated by Figure 7 containing extreme value analysis for several tide gauge locations, where observational data is added to verify the performance of the dataset in terms of extreme water levels.

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

For comparison of surge height statistics in Figure 4 (L251) we now compare two 30-year periods in terms of mean annual maxima (instead of percentiles), as this is more illustrative. This comparison is made between the backward extension period (1950-1979) and the most recent period of 1995-2024, following the suggestion by the reviewer.

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).

In the revised manuscript we are no longer including global statistics of GTSM-derived and observations-derived extreme values, only annual maxima comparison, which is a more straightforward and transparent comparison metric. To illustrate the comparison of return periods and extreme events from the model and observations, we include EVA plots for 8 global stations (Figure 7 in the revised manuscript, L309), which now includes EVA of tide gauge observations.

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

Thank you for pointing out this typo. As described in response to earlier comments, we decided to replace the percentiles-based validation with monthly and annual maxima, as a more transparent metric which also gives us more insight into storm surges and extreme events.

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.

We have now mentioned more explicitly the possible effect of climate variability, when comparing statistics of surges between 30-year periods, with a reference to Lobeto and Menendez (2024). For full text accompanying Figure 4, please see the response to the next comment below.

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?

This comment might be referring to the original Figure 4. The figure comparing the surge statistics (Figure 4, L251) has been replaced with a comparison of mean annual maxima, with accompanying text as follows:

"To quantify the difference in surges between the backward extension and the more recent period, we compare the surge height statistics by calculating the mean annual maxima values based the 30-year timeseries for 1950-1979 and 1995-2024. *The difference is shown in **Figure 4**, where it can be seen that in by far the most areas the difference is minor (within 0.1 m), indicating that the surge statistics are reasonably consistent between the extended and the original periods. There are, however, several areas where larger differences are visible. In general, there are negative differences in high-latitude regions, such as Antarctica and Northern Europe, and positive differences difference for most of the rest of the globe. It is not clear what drives these differences in surge statistics and further research could investigate the source, it is possible that it is caused by climate variability (Lobeto and Menendez, 2024).* "

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.

Thank you for this suggestion. We have added a more clear description of the comparison of surge statistics between the two 30-year periods (see previous comment), highlighting the possible effect of climate variability. We do not use percentiles in the revised manuscript, as it is more informative to focus on lower-frequency values such as the annual and monthly maxima (see responses to earlier comments). The comparison of mean annual maxima surge statistics now precedes the analysis of extreme value statistics.

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

The increase in the confidence intervals is caused by the fact that there are often more extreme TC-induced events included in longer timeseries. For many of these regions, tropical cyclone-induced events can be seen as ‘outliers’ in the extreme value analysis given that they are so much higher than evens induced by other types of storms. As a result of including more of these very extreme events, the confidence interval becomes larger. Reducing the uncertainties for those locations would require a much larger sample size, which can only be obtained with synthetic events.

In the revised manuscript, the sentence is changed to:

L272: *"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 is too small and robust estimates of return period for those regions require thousands of years of tropical cyclone activity (Lin and Emmanuel, 2016; Bloemendaal et al., 2020).”*

L270 “higher long-term extremes” what do you mean by this? Clarify

This has now been rephrased for clarity. This sentence is meant to highlight that for some stations, the use of longer time series results in higher extreme values for 100-year return periods. The text is rephrased as follows:

L285: *“At other locations, such as Fremantle and Hosojima, higher extreme return values are obtained when using longer timeseries – at these locations the extension of the time series allows the inclusion of more severe storms that were not part of the original 1979-2018 dataset”.*

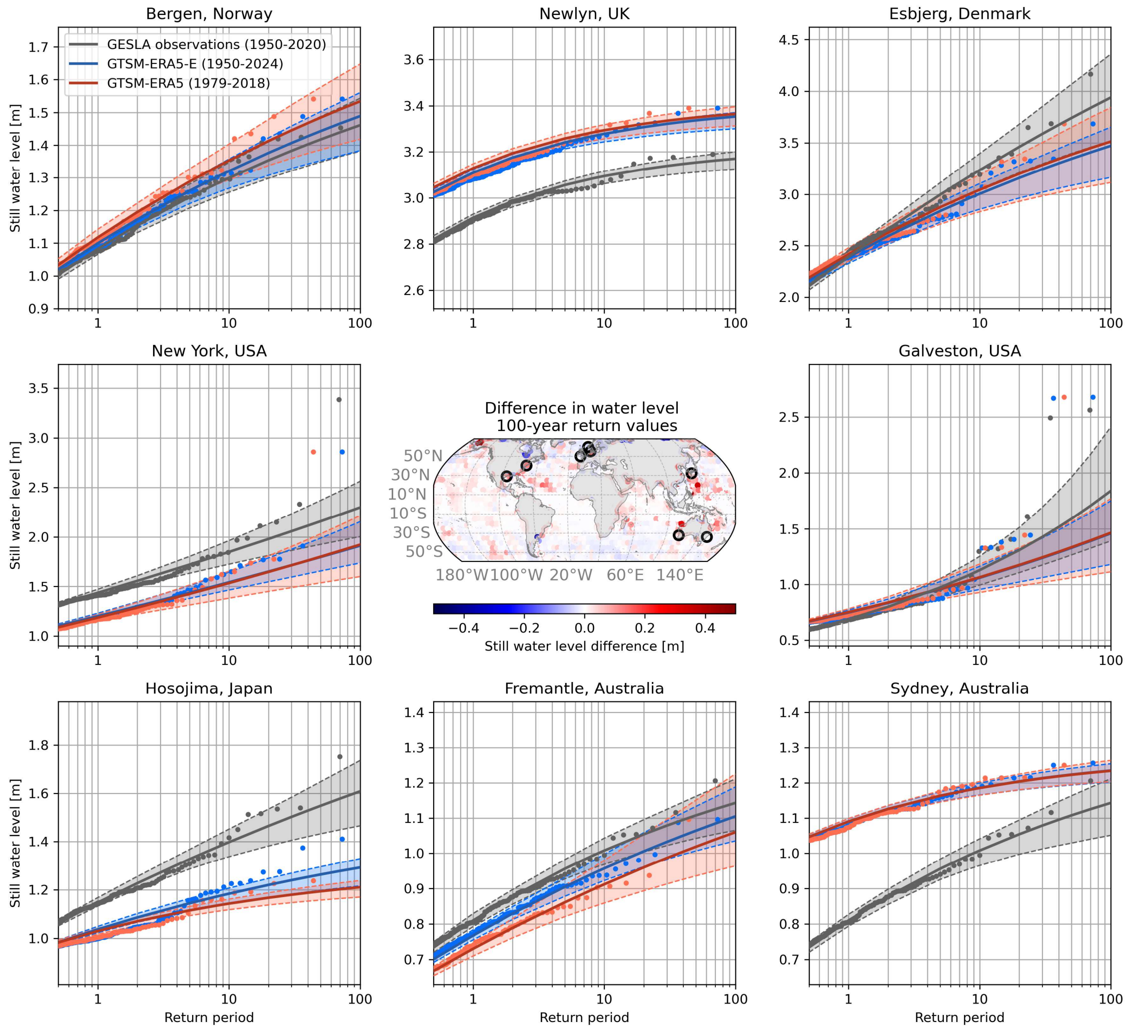
In relation to the analysis in Figure 5, it would be interesting to see in tide-gauges with long and trustworthy records how the extended dataset may help reduce (or not) the error in the calculation of return levels. The question is: is the longer dataset really more reliable, or longer but of poorer 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.

Following this comment, we have revised this figure (Figure 7 in the revised manuscript) to be able to demonstrate the model-derived and tide gauge-derived extreme values for specific stations. A new selection of stations was made, focusing on stations with long-term observations with data availability of over 95% in the 1950-2024 period. The EVA results (GPD-POT fit, individual data points and confidence intervals) are demonstrated based on GTSM-ERA5 pre-extension and including the extension, as well as based on long-term tide-gauge observations.

Based on this figure, we updated the text with an interpretation of these results:

*"The comparison between GTSM-ERA5-derived and observations-derived extreme values in **Figure 7** shows that for most stations there is a vertical bias. This offset can be seen both for the fitted extreme value distribution (blue and black solid lines) as well as the individual extremes (blue and black dots). These biases are expected in a global model, and can be resulting from inaccuracies in the tidal signal (due to low resolution of the model and complexity of some of the coastlines, as well low accuracy of bathymetric information in some regions) and underestimations of surge magnitudes (which can be a result of underestimations of wind speeds in ERA5, especially for tropical storms, as well as due to inaccurate bathymetric information in some regions). However, the observed frequency of the individual extremes matches with the GTSM-ERA5 datasets, which can be seen by comparing the blue and black dots in **Figure 7**. "*

The new figure 7:



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.

We agree with the reviewer's suggestion to move the timeseries comparison to earlier sections, and make the text less focused on tropical cyclones. We have moved the timeseries comparison to Section 4.1 (starting at L221), and selected locations where long-term tide gauge records are available, in order to be able to verify the modelled water levels with observations. The event-specific time series example thus serves to strengthen the validation for extreme events. The figure in the manuscript has been replaced (Figure 3, L236 in the revised manuscript).

The following storms were selected from the long-term measurements:

- North Sea storm (1968, Esbjerg tide gauge). 4th highest measured water level in the 1950-2020.
- Hurricane Carla (1961, Galveston tide gauge). 2nd highest measured water level in 1950-2020.
- Typhoon Marie (1954, Hosojima tide gauge). The highest measured water level in 1950-2020.

This comparison is discussed in the revised manuscript as follows:

L221: "In addition to the impact of the dataset extension on water level statistics, ***Figure 3*** *also demonstrates the GTSM-ERA5 water levels comparison to observations for several specific major storm events in the extension period (1950-1978) at three tide gauges with long-term records. This includes the storm surge caused by the extratropical storm in the North Sea in 1968, hurricane Carla (1961) that impacted the United States, and typhoon Marie (1954) that impacted Japan. The storm surge caused by the North Sea storm in 1968 at the Esbjerg tide gauge is the fourth highest water level in 1950-2024 and it is well-captured in the GTSM-ERA5-E with good agreement both in terms of timing and magnitude of the maximum water levels. For the Galveston tide gauge, water levels during storm Carla are the second highest over the 1950-2024 record (second to Hurricane Ike in 2008); the magnitude and timing of the storm surge are well-represented by the GTSM-ERA5-E, with differences in the decay of the elevated water levels after the storm surge peak – in the observations the water levels stay elevated for a longer period. In the case of typhoon Marie and the Hosojima tide gauge (highest measured water level in 1950-2024), GTSM-ERA5-E captures the timing of the maximum water levels well, but underestimates the storm surge by approximately 0.5 m. "*