Response to reviewer #2

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

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 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 will substantially revise the manuscript to address the points raised by the reviewer. Most importantly, we will improve the validation analysis by two major changes: 1) instead of using percentiles, we validate monthly and annual maxima; and 2) instead of comparing the validation for the full 1950-2024 dataset vs. the original 1979-2018 dataset, we will 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 will be restructured, Table 1, Figures 2, 3, 4 and 6 will be 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 will revise the manuscript to clarify any ambiguities that were identified by the reviewer.

We believe these revisions will 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 propose to change this sentence as follows:

"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 proposed new text is as follows:

"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 will be rephrased as:

“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 will be clarified as follows:

“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 will be as follows:

“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:

"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 will indeed be changed 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 will be rephrased as follows:

"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.

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 will be clarified in the text as follows:

"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.

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 25% in 1950-2024). The dataset was also filtered for duplicated records, which has reduced the selection further to 107 stations. Subsequently, we validate the two periods separately. The text will be modified accordingly:

L144: "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*

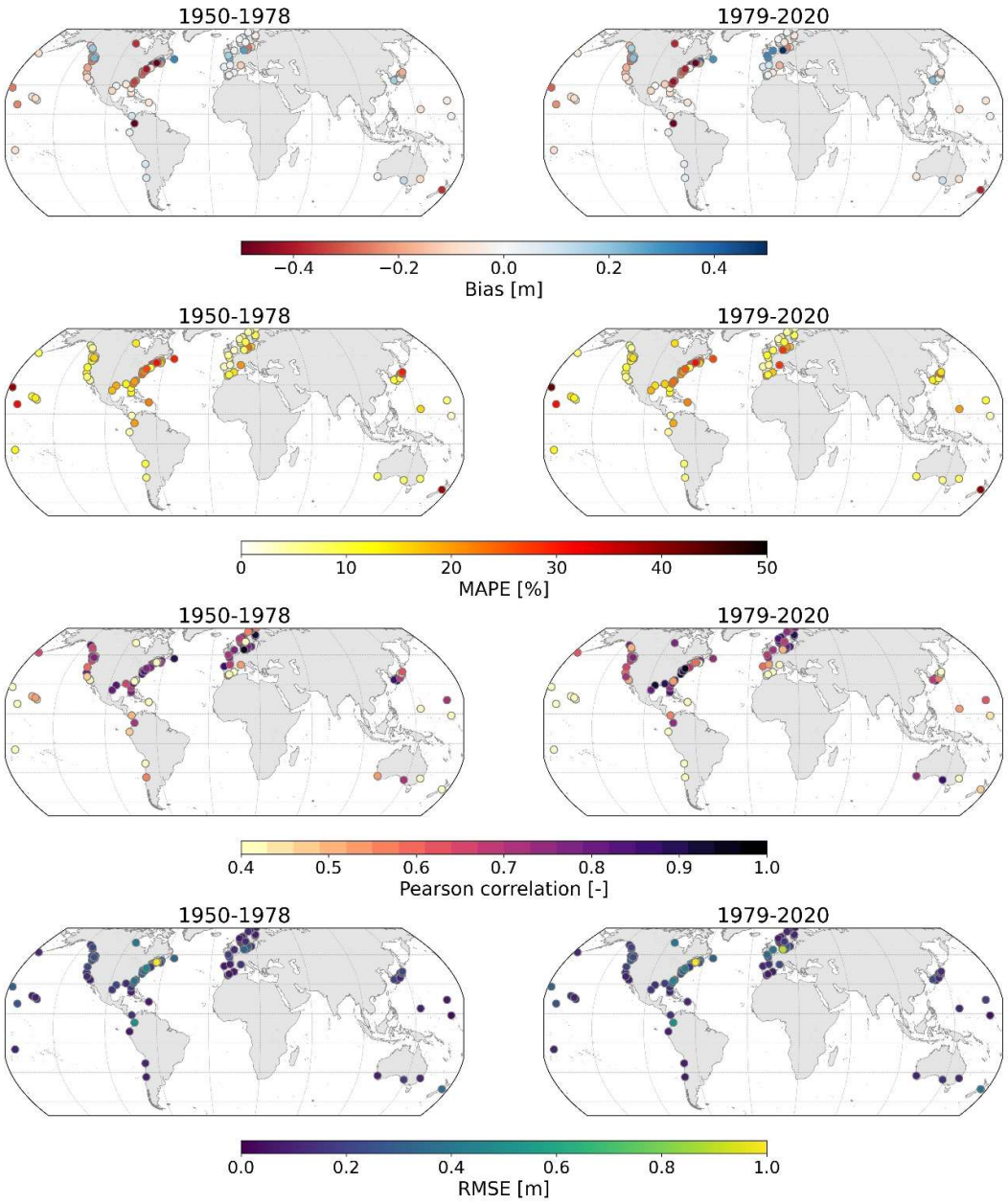
for extreme value analysis, where each station has a counterpart location in the GTSM-ERA5-E dataset (located less than 10km away). The data coverage in the resulting observational dataset used for data validation is 96.2% for 1950-1978, and 97.5% for 1979-2020 (on average across stations).”

The comparison shown in the new Table 1 (below) 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.

Table 1: : **Model performance of the GTSM-ERA5 in the extended period (1950-1978) and the period without the backward extension (1979-2020), comparing still water levels to tide gauge observations (selection of GESLA-3 long tide gauge records). The values are presented as mean across stations with standard deviation (in brackets).**

Data	Comparison metric	GTSM-ERA5 1979-2020	GTSM-ERA5 extension 1950-1978
Monthly maxima	Mean bias [m]	-0.03 (0.24)	-0.03 (0.26)
	MAE [m]	0.18 (0.18)	0.19 (0.19)
	MAPE [%]	33.8 (79.0)	20.5 (14.8)
	RMSE (m)	0.21 (0.18)	0.22 (0.18)
	Pearson corr. coef.	0.74 (0.17)	0.73 (0.17)
Annual maxima	Mean bias [m]	-0.11 (0.25)	-0.10 (0.26)
	MAE [m]	0.20 (0.19)	0.21 (0.19)
	MAPE [%]	14.2 (8.9)	14.5 (8.9)
	RMSE (m)	0.22 (0.19)	0.23 (0.19)
	Pearson corr. coef.	0.60 (0.24)	0.60 (0.27)

New Figure 2 – Annual maxima validation, comparison between two periods:



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 (99th and 95th) 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 would like to restructure 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.

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.

Figure will be enlarged in the text, with (a,b,c..) labels added to each panel. This figure will be changed to provide a comparison for annual maxima validation, instead of the percentiles (see the new Figure 2 provided above in response to an earlier comment).

Regarding the Sao Paolo station, there is a dot on the first plot, but it is very light yellow, so it is barely visible. We are adjusting 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) in the validation plots, to provide more information on relative values. A plot for relative 100-year return values between the extended period and original period of GTSM-ERA5 will be added to the manuscript.

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 review made us realize that for the purpose of validating the extended dataset we need to make some changes to presented analysis. We decided to omit the percentiles and focus on monthly and annual maxima, which will reflect better how the GTSM-ERA5 datasets represent storm surge extremes. In the revised manuscript, we will update Table 1 and Figure 2 (the new table and figure are shown above in response to an earlier comment), and change the accompanying text accordingly.

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

In the revised manuscript we will show validation metrics for two periods: 1950-1978 and 1979-2020 (in line with other responses). For comparison of surge height statistics between the different 30-year periods, we will change the latter period to 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).

In the revised manuscript we will not be 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 (a revised Figure 6), which will now also include EVA of observations. This includes several tide gauge locations in TC-affected regions, to demonstrate the underestimation of TC-related extremes by GTSM-ERA5.

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 will mention more explicitly the possible effect of climate variability, when comparing statistics of surges between 30-year periods, with a reference to Lobeto and Menendez (2025).

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 Figure 4. We will contextualize this comparison in more detail, please see the extended response to the next comment below.

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 will add a more clear description of the comparison of surge statistics between the two 30-year periods, highlighting the possible effect of climate variability. Based on the other comments of the reviewer, we would like to only show this for surge, and not for total water levels, because the 95th or 99th percentiles of total water levels are significantly affected by tidal statistics, and surge percentiles are more informative. The plots with the difference in confidence intervals of 100-year return values will be added as a second 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

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 will be rephrased to:

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 is too small and robust estimates of return period for those regions require thousands of years of tropical cyclone activity"*.

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

This will be 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 will be rephrased as follows:

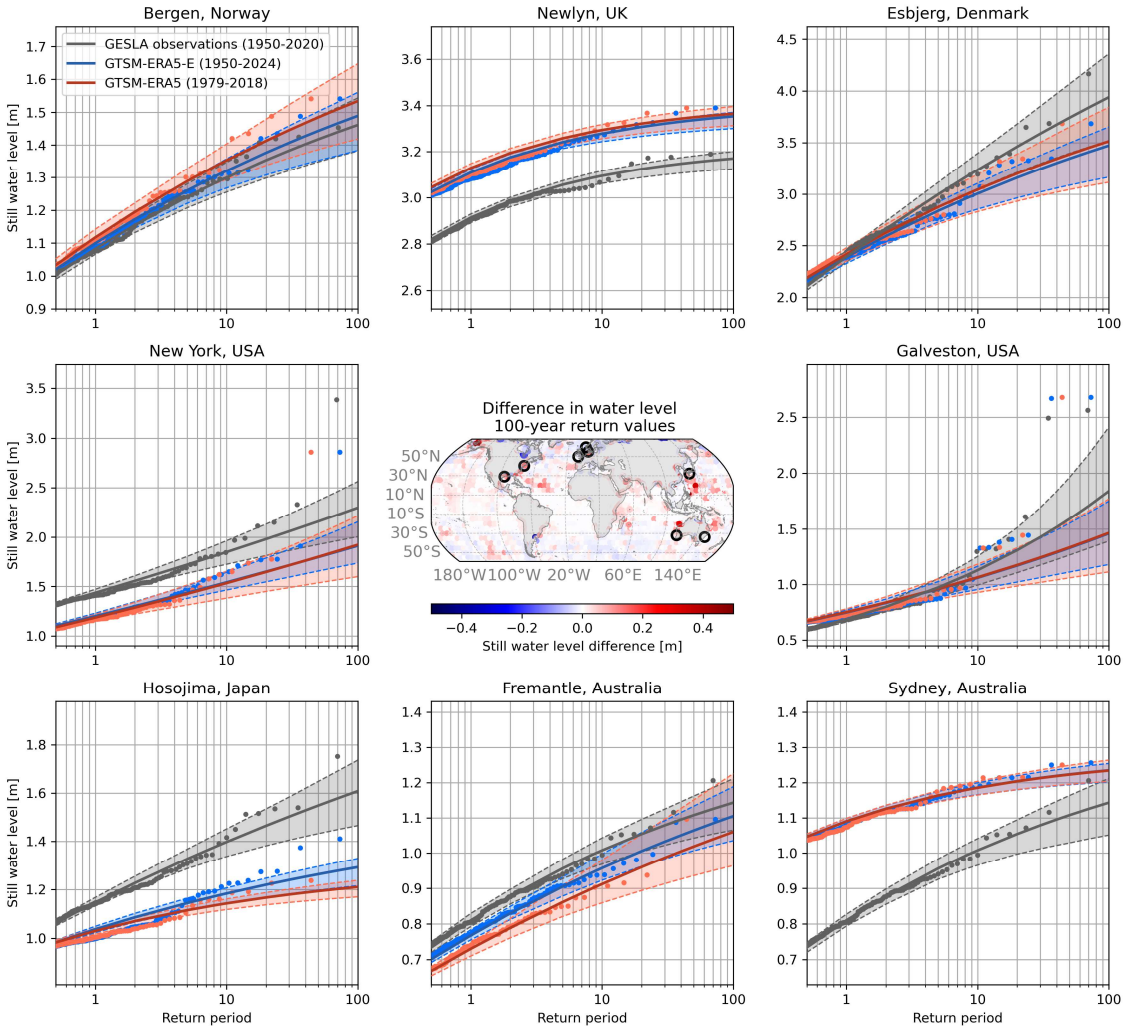
L270: *"…at these locations the use of longer time series results in higher 100-year return value estimates, when the extended time series allow the inclusion of severe storms that were not part of the original 1979-2018 dataset".*

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 Figure 6 to be able to demonstrate the model-derived and tide gauge-derived extreme values for specific stations. A new selection of stations for Figure 6 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 in a new Figure (see below).

Based on this figure, we will update the text with an interpretation of these results. The comparison shows that GTSM-ERA5 captures the extremes quite well - the extreme values from the extended GTSM-ERA5 (blue dots) match the frequencies of extreme values from tide gauges (black dots) in terms of frequency. The biases and underestimations seen in this plot 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).

New Figure 6:



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 would like to put the event-specific timeseries comparison just below Figure 6 (return periods for extremes at 8 global stations), and change the event selection to events that are also visible in the

observations. This way, the event-specific time series example serves to strengthen the validation for extreme events. 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. Good agreement with observations.
- Hurricane Carla (1961, Galveston tide gauge):
  2nd highest measured water level in 1950-2020. The water level peak is overestimated by GTSM-ERA5, but of a similar order of magnitude to the measured values.
- Typhoon Marie (1954, Hosojima tide gauge):
  The highest measured water level in 1950-2020, also the highest in the GTSM-derived time series, although the peak water level is underestimated by 0.5 m. Surge signal is clearly visible.

The new Figure 7 will be like this: