The authors would like to thank the reviewer for their time in reading through our manuscript. We are happy to hear that the study was well received and have made some changes as per your suggestions. Here we list those changes.

[Specific comments]

Figure 1: Define in Figure 1 where “Kattegat”, “Schleswig-Holstein” and “Mecklenburg-Western-Pomerania” are located.

We have updated the figure to include the names of the states, and included a description of their border in the figure caption:

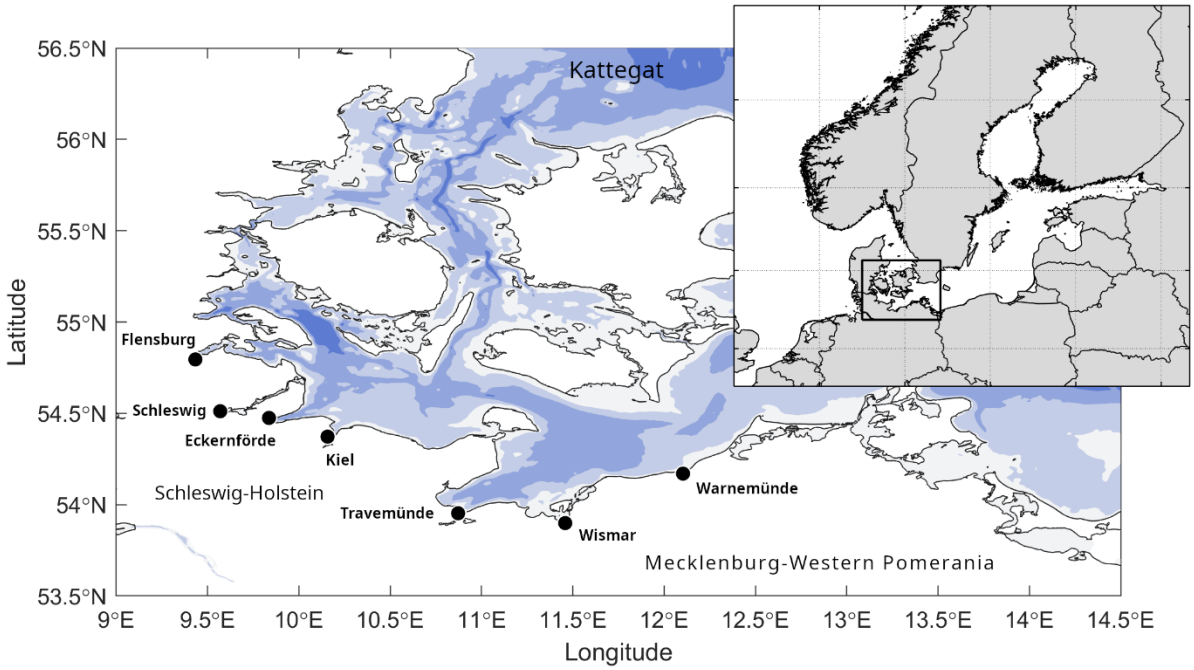
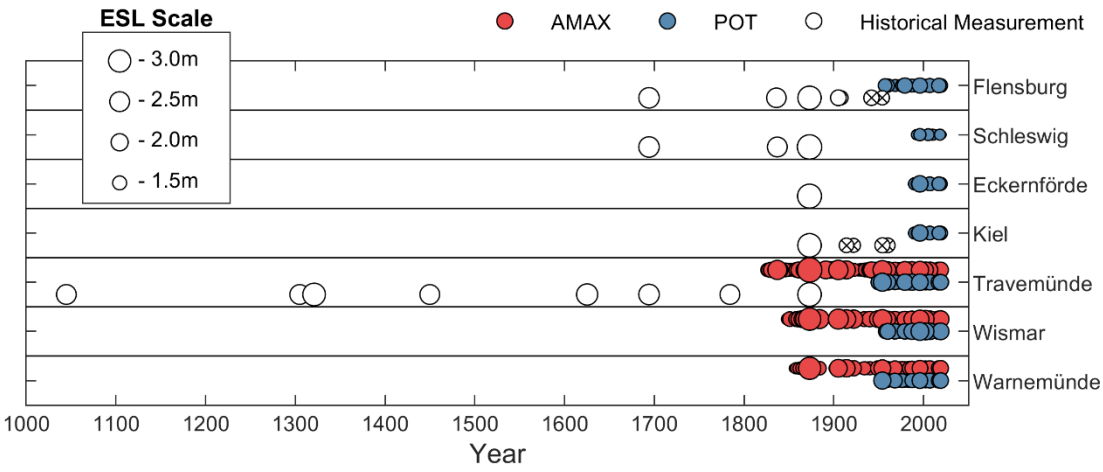


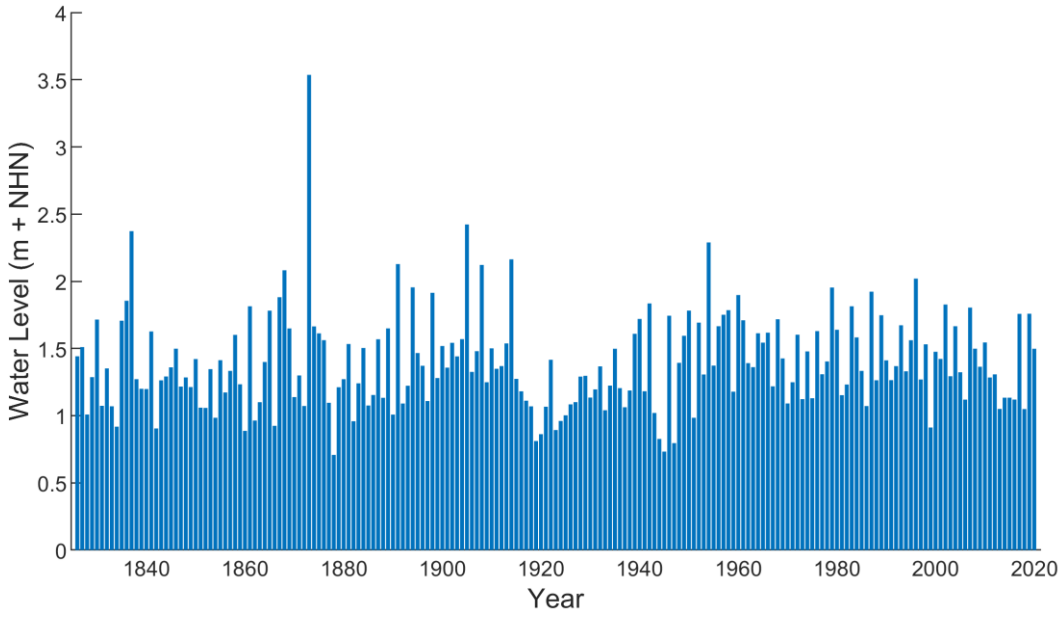
Figure 2: Although Figure 2 exemplifies the data series, it does not contain the water level. No scale illustrates the events’ magnitude, so the reader cannot visualize the water level. Including the information related to the 1872 event. How large was it? I recommend the authors plot the data series for each site containing the thresholds presented in lines 191 and 192 and the POT and AMAX selected events.

Thankyou for this comment. We decided to keep the current figure as it is used more to depict the extent of data rather than show precisely the water levels. We have added a scale to the figure so that some idea regarding the magnitude of the events can be interpreted. When we plot the data at all seven sites, it is difficult to discern the differences in record lengths, which is the intent of this figure. We have also mentioned the height of the 1872 ESL event in Section 2.1:

“The resulting ESLs along the German coast remain the highest on record, registering approximately 3.4 m above mean sea level (MSL) at Travemünde.”



In addition to these changes, we have included an extra figure which shows the AMAX record at Travemünde, highlighting the exceptional nature of the 1872 event:



Line 191: The authors justify why they had to select different thresholds for each site but do not give major details. In this sense, explain the criteria for choosing each threshold in more detail.

We have included a more thorough description of the threshold selection process:

> *"As mentioned in Section 3.1, it is a necessary condition that the available historical information is exhaustive above a perception threshold. That is, the only events which have exceeded the perception threshold for the duration of historical observation exist within the historical record. Therefore, the perception threshold should be set high enough to ensure this assumption is true. At first, a systematic approach to setting a perception threshold was attempted based on the systematic data and the period of historical information. Here, ESLs were estimated using systematic data only for return periods dependent on the number of historical events available and the length of the historical record. For example, the perception threshold might be set to a height equivalent to a 1-in-100 year event, where a 200 year long historical record is available*

which contains 2 events. However, due to large differences in the magnitude of systematic and historical observations, relying on the systematic data alone was not sufficient, and no one method could be applied at all sites. Instead, perception thresholds were chosen on a site-by-site basis, using all available data for each case.

Given the lack of a clear physical threshold at any of the tested locations (e.g. a sea wall where all exceeding events are recorded), a threshold selection process was conducted based simply on the author's intuitive reasoning. Factors that influenced the selection process include the magnitude and occurrence of ESLs in both the systematic and historical records and the length of the historical record in question. Keeping in mind the assumption that the historical record is exhaustive, and due to the subjective nature of this method, final perception thresholds were set conservatively high at 2.3 m at Flensburg, 2 m at Schleswig, 2.25 m at Eckernförde, 2.25m at Kiel, 2.6 m at Travemünde, 2.25 m at Wismar and 2 m at Warnemünde. Historical ESLs that do not exceed the perception threshold cannot be used in the analysis, and are thus disregarded. These events are highlighted in Figure 2."

Line 199: "…AMAX samples are used in lieu of historical measurements for analysis 2,…" What is analysis 2? Also, the authors say "lack of historical information," but the AMAX samples are longer than the historical measurements in Eckernförde and Kiel, for example. Clarify these sentences.

We describe four analyses in the previous sentence: "Depending on data availability at each site, we perform four separate analyses using: 1) POT samples only, 2) POT samples with historical measurements, 3) AMAX samples only, and 4) AMAX samples with historical measurements." Analysis 2 refers to the 'POT samples with historical measurements' analysis. The statement "lack of historical information" applies only to the sites at Wismar and Warnemünde, not to Eckernförde or Kiel. Because there are no historical records of ESL at Wismar and Warnemünde, and given we are only considering POT samples from the high-resolution tide-gauge data, we use events from the AMAX record as historical events. We have rephrased the sentence to make this more clear.

"At Wismar and Warnemünde, as no historical records are available, ESLs from the AMAX records are used in lieu of historical measurements for the second analysis."

Line 205: If you refer to what can be seen in Figure 2, I cannot see it properly. Figure 2 should be reformulated to better visualization of the events.

Thankyou for this comment. Indeed, it is difficult to see any reduction in ESL values in the AMAX record from Figure 2. We have included a new figure (Figure 3: see above) which provides the AMAX values at Travemünde.

Line 228: Why a 70-year moving window?

A 70 year moving window was chosen based on the length of the high-resolution tide-gauge record, which is 71 years in length. This should be stated in the manuscript, thanks for highlighting this omission. We have added the following into Section 3.4:

"A window size of 70 years was chosen to match the length of the high-resolution tide-gauge record at Travemünde (71 years in length)."

Line 240: You haven't mentioned HW1000 before. Please include it in the methodology.

Like HW200, this refers to an extreme sea level with a return period of 1000 years. We have added in brackets the meaning of this term to clarify this point.

"HW1000 (1-in-1000 year ESL event)"

Line 245: The changes in the maximum likelihood estimation being negligible for AMAX makes me wonder what the results would be for the other sites that you used POT only. Have you tried to apply AMAX on the other sites with the data you have? Could this be related to the method applied (POT x AMAX)?

Thank you for this comment, this is an interesting point which we did not consider. Addressing the comments of another reviewer we performed the Bayesian MCMC analysis at each site using only the high-resolution tide-gauge records. Therefore, we could compare the performance of the POT and AMAX sampling using the same length of data. This was added to the manuscript as an Appendix:

"Appendix A: ESL sampling

In this study, we employ two different approaches for sampling ESLs. The first technique, POT is generally preferred in literature for reasons explained in Section 2.2. Wahl et al. (2017) also note that AMAX sampling may result in larger uncertainties at the tails of the distribution when sea level records are short. While we are constrained by the records sourced from MLUV (2012), which are only available as AMAX samples, this is not the case for the high-resolution tide-gauge data. Thus, we decided to employ POT sampling for these records due to their preference in literature, and so that we may demonstrate the use of the Bayesian MCMC EVA method for both POT-GP and AMAX-GEV approaches. To directly compare the two approaches, we also performed an AMAX-GEV analysis using the high-resolution tide-gauge data. Figure A1 shows ESL estimates including 95% credibility intervals estimated at each site using the high-resolution tide-gauge data only. At all sites, the AMAX-GEV method results in larger estimates of ESLs at high return periods. Also of note are the larger uncertainties at the tails of the distribution at all sites except for Warnemünde, which supports the findings of Wahl et al. (2017). In general, the POT-GP appears to produce more reliable results given the same record duration based purely on the reduced uncertainties.

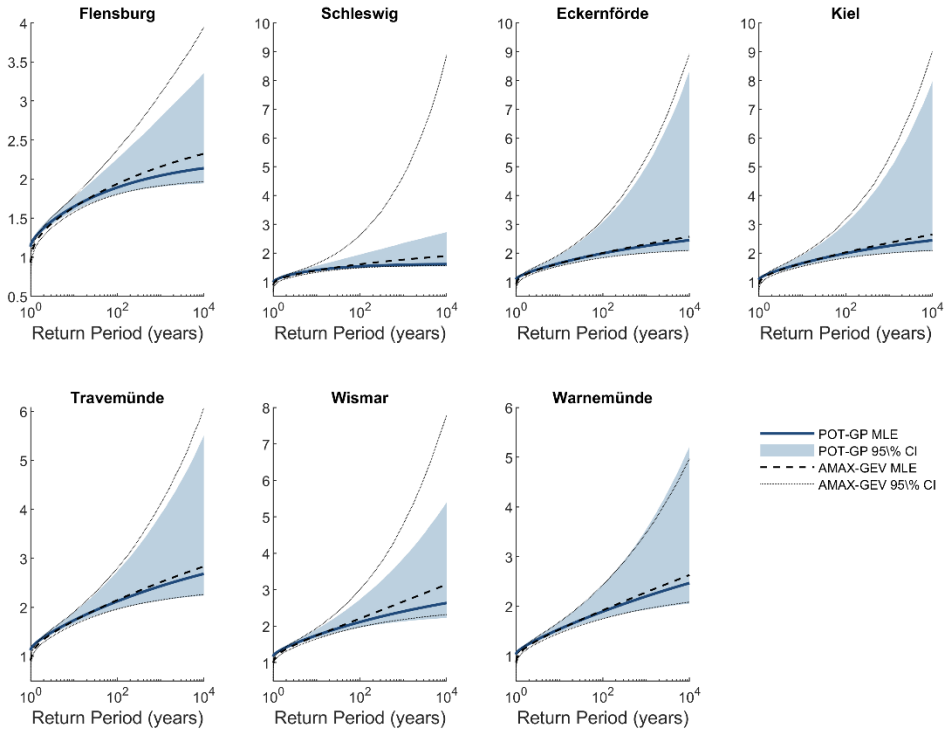


Figure A1. Comparison of POT-GP and AMAX-GEV approaches to the Bayesian MCMC method for estimating ESLs. At each site, estimates of ESLs are made using high-resolution tide-gauge data only. In general, the AMAX-GEV approach results in higher estimates of ESLs at high return periods and larger uncertainties at the tails of the distribution.

As per your suggestion, we also repeated this analysis including historical information and found that in general, the AMAX-GEV analysis outperforms the POT-GP analysis:

> *"Next, we considered how these estimates are affected by the addition of historical information. We performed the analysis again, but included historical information with measurement uncertainties given by Jensen et al (2022). Results are shown in Figure A2. As with the first analysis, both the POT and AMAX samples are taken from the high-resolution tide-gauge record. For both POT-GP and AMAX-GEV analysis, the introduction of historical information is beneficial in terms of reduced estimate uncertainties. Interestingly, we find that the AMAX-GEV approach performs better in terms of reduced uncertainties at all sites except Schleswig. Differences in the maximum likelihood estimates between the two analyses are much reduced."*

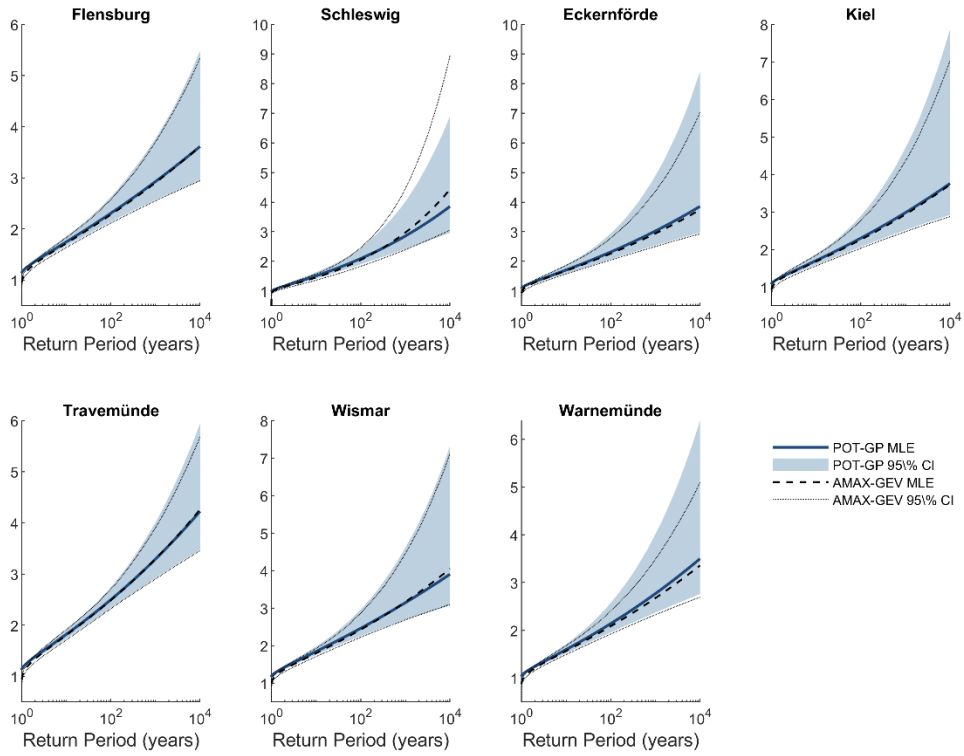


Figure A2. Comparison of POT-GP and AMAX-GEV approaches to the Bayesian MCMC method for estimating ESLs including historical information. At each site, estimates of ESLs are made using the same high-resolution tide-gauge record in combination with historical information. In contrast to results where historical information is omitted, the AMAX-GEV approach performs somewhat better than the POT-GP approach in terms of reduced uncertainties at the distribution tails. Differences in the maximum likelihood estimates between the two methods are much reduced.

We have also included a paragraph dedicated to this in the Discussion:

"Large differences exist in the estimates of ESLs made using either the POT-GP or AMAX-GEV approaches. While the incorporation of historical information reduces these differences, it does not provide any insight into which method performs best. Indeed, the POT-GP approach is generally preferred in the literature (Arns et al. 2013, Wahl et al. 2017}, but this does not necessarily apply to the case of the German Baltic Sea coast. We find that when both methods are constrained to the same record length (see Appendix A), the POT-GP method generally performs better with lower uncertainties at the distribution tails. At all sites, the AMAX-GEV provides larger ESL estimates at high return periods. Interestingly, including historical information in the analysis produces a different result, with the AMAX-GEV analysis providing lower uncertainties at high return periods. One possible explanation for this involves the sampling threshold for the POT method. We assume that this threshold is constant for the full duration of historical and systematic observation, following the study by Bulteau et al. (2018), but this may not be the case. Indeed, large differences in results due to the inclusion of historical information suggests this assumption may be false. Thus, an advantage of the AMAX-GEV approach is that no sampling threshold is required. Given a single sea level record with no historical information, we would recommend the POT-GP approach over the AMAX-GEV due to the reduced estimate uncertainties. However, the AMAX-GEV approach may provide more

precise results when historical information is available. Where a longer AMAX record is available, such as in this study, the AMAX-GEV approach provides clearly better results due to the increased data."

Line 251: I believe you are saying 71 years of systematic data, but how many POT events are considered in these 71 years? Is the result difference caused by the amount of sampling or the different techniques? This discussion could be added in the manuscript.

The number of POT samples depends on the selected threshold. This is a very important part of the POT approach and care must be taken to select an appropriate threshold. This is described in Section 3.3. If a threshold is set too low, more samples will be allowed but this may bias the analysis by introducing non-extreme events. In our case, there are 119 ESL events within the 71 years of data at Travemünde. The difference between the results of the AMAX and POT analyses when considering systematic data only is indeed due to the large difference in record lengths. There is some differences due to the method employed, but this is a minor consideration when the records are substantially different. A comparison of the two methods using the same observational record has been conducted and included in the manuscript as Appendix A (see above comment).

Table 2: Review all numbers referred to in Table 2 in the text. It could be a rounding issue, but some of the numbers presented in the text are slightly different from the ones in the table—for example, the difference shown in line 272 is 47 instead of 48 cm.

This is indeed due to rounding. All figures have been rounded to the nearest cm in both the table and in the text. When we calculated differences between the figures, we only performed the rounding after the operation was completed to avoid double rounding. This does result in some discrepancies between the table and the text, but is also the most correct method of presenting the results. We have added a note to the caption of Table 2 that all figures have been rounded to the nearest cm.

Line 273: The authors present differences in HW200 for Wismar and Warnemünde (2 cm and 5 cm, respectively) but then say in line 244 that "the effect of incorporating historical information can only be examined at Travemünde". Where do these results come from?

The earlier statement is regarding the AMAX-GEV analysis only. As no historical records are available for Wismar and Warnemünde, it is not possible to conduct the AMAX-GEV analysis with historical information at these sites. The statement on line 273 refers to the POT-GP analysis. As mentioned earlier, when conducting the POT-GP analysis at Wismar and Warnemünde, we use AMAX samples as historical events due to the lack of true historical records. Confusion is due to the poor wording on line 269: "Looking only at the sites where long AMAX records are available (Travemünde, Wismar and Warnemünde), we see much better agreement between POT-GP and AMAX-GEV analyses when historical information is considered" We have updated this sentence to provide clarification:

"Looking only at the sites where long AMAX records are available (Travemünde, Wismar and Warnemünde), we see much better agreement between the POT-GP analysis

including historical information and the AMAX-GEV analysis of systematic data only, despite the significantly shorter tide-gauge records."

The authors say the systematic record at Travemünde started in 1949. But there has been a continuous measure of AMAX since 1826. In this case, the systematic data mentioned in Figure 5 are from 1826, right? And what about the syst.+hist one? If historical data are the white circles in Figure 2, in a moving average of 70 years, from 1900 onwards, it would include only one datum, the one from 1872. Is this one event changing that much in the results from using systematic data only when compared to systematic + historical data? Later you explore this better in the discussion, but I believe that Figure 5a needs to be better described in the results section.

Thanks for the comment. Firstly, we realise that our naming of the high-resolution tide-gauge record and the AMAX record has introduced some confusion. Although the AMAX record consists of both systematically measured ESLs and historical information, their compilation provides a systematic record of annual maxima sea levels which we can use as systematic data. From line 78, we state that "the longest systematic record was installed at Travemünde at the end of 1949". This refers to the high-resolution tide-gauge record and we have updated the manuscript to clarify this point:

"The longest high-resolution (hourly sampling) systematic record was installed at Travemünde at the end of 1949."

Yes, Figure 5 shows an analysis of the AMAX record which begins in 1826 and is considered to be a systematic record given its compilation. The Syst + H.I. plot from Figure 5a shows ESL estimates from a 70 year window of AMAX values including all available historical information (not just those values within the 70 year window). This is described in Section 3.4 (lines 227-230) but we have added the following sentence to clarify our methodology:

"For the second case, all available historical information is used, even those events which occur outside of the 70-year window."

The effect of 1872 on ESL estimates is shown in Figure 5. When considering systematic data only (70 years of AMAX ESLs), we can see a large decrease in ESL estimates at 1943 where 1872 is no longer considered (Figure 5a blue). By including historical information, 1872 and other large historical events continue to influence our estimates which remain relatively steady over the period considered (Figure 5a).