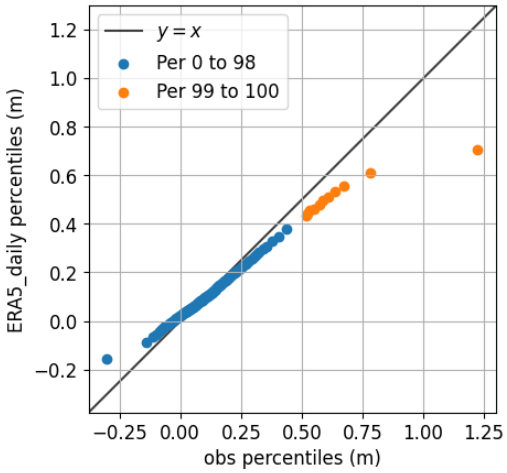
This manuscript presents a statistical reconstruction of storm surge records using 17 climate model projections along the European coastlines. It is the result of a significant computational effort, combining a (relatively) small set of dynamical numerical simulations and a data-driven model based on multiple linear regression. The methods are well described and sound. All the details are provided for the calibration and choices of the statistical model parameters. However, I have some concerns on the application of the model, and the presentation and interpretation of the results. I think the manuscript requires a major revision before being suitable for publication.

One major concern is the focus on extreme storm surges. The results of the statistical model applied to ERA5 forcing fields are daily records of storm surges that, when compared with the benchmark results of the dynamic simulations, provide satisfactory performance in terms of averaged storm surges. This is shown in Figures 3 and 4 that show the capabilities of the statistical method in terms of correlation and RMSE. The only metric focusing partly on extremes is the bias of the 99th percentile, but this is not representative of the ability of the statistical records to capture extreme events and reproduce return levels, which are the results that are analysed later on. In fact, the comparisons in figure 4 show that discrepancies can be quite large for individual events. This is something well-known for the statistical model based on Tadesse et al (2020). The approach is good in simulating the mean storm surge climate but displays limited accuracy with the extremes, and this is a major shortcoming that must be reflected in the manuscript. I attach below an example for a tide gauge in Brest that I produced some time ago for an assessment of data-driven models. Although the differences with the dynamic simulation are expected to be smaller than in this example, as the extremes will be also underestimated (shown e.g. in Figure 1), it is clear that the statistical model is not particularly well suited for extreme storm surges. I am not suggesting that the authors should change the statistical approach, I believe it has its value. I think, though, that the ability of the method needs to be better described, particularly concerning extremes. To do so, I suggest using qq-plots instead of time series to evaluate model performance. Likewise, mapping the differences in maxima or yearly maxima and/or return levels between statistical and dynamical approaches forced by ERA5 would provide the required information to the reader.



A second major concern is related to the discrepancies between dynamical and statistical simulations in climate models for some particular regions. As shown in Figure 5, regions as the Mediterranean (CNRM-CM6-1-HR) and the Baltic (MPI-ESM1-2-HR) indicate opposite changes in projected storm surges using dynamical and statistical models. To a lesser extent, also the western of the British Isles and the southern North Sea display different patterns. This needs to be clearly described. I do not think that the patterns are similar and only the magnitudes change, as claimed the lines 289-290. I agree, though, that using the hindcast instead of historical simulations for the training is mostly fine.

These discrepancies hinder the interpretability of the projected storm surges presented in figure 7. The regions where the statistical and dynamical models clearly differ should not be discussed or even mapped. This includes the Baltic and the Eastern Mediterranean Seas (the western Mediterranean seems consistent in the validation). In addition, the uncertainty range is very high for the 100-year return levels, ranging from negative to positive values. This indicates that there is no confidence in the multimodel ensemble means (panels 7i and 7m). I suggest focussing only in the 10-year return level. This is also consistent with the fact that the statistical model is less reliable for the most extreme events.

Another issue that I think requires some attention is the discussion about long-term variability in the atmospheric patterns and its impact on the performacen of the statistical model (lines 430-449). I do not think that the long-term modulation of low-frequency climate modes affects the results of the statistical model. Storm surges are caused by synoptic systems. These can be altered in frequency and magnitude by large-scale climate modes. However, the synoptic systems are still the same. In other words, changes in large-scale atmospheric conditions, like more blocking patterns, shifts of NAO, etc, will modulate the frequency and the intensity of the systems that generate the storm surges, but will not change the

process and the type of system, nor the response of the storm surge. There are some statements in this paragraph in this line that I do not think are correct (lines 434-435, 437-438, 443-444). The only exception I can think of is the arrival of tropical-like cyclones in the future climates to the European coasts. However, these would not be well captured by the coarse resolution models anyway. I think this part needs to be reconsidered.

On a personal note, I find the reading more difficult with the use of so many acronyms. My preference would be to avoid the use of at least some of them. For example: SS as storm surges, or even SDM and DDM could be referred to as, simply, statistical model and dynamical model.

Other comments:

- Line 92: I am unsure what this means. Perhaps that one in every N(?) coastal points are analysed? If so, what is the averaged distance among coastal points? Please, clarify.
- Figure 1 has a wrong caption. Reference to Fig 1c in line 105 is unclear.
- Lines 99-104: please, provide references here. This pattern is shown multiple times in the literature.
- Table S1: homogenise units.
- Figure S2: Please, increase the size of the figure and the font size. It is not readable.
- Line 150: SS (I guess storm surge) has not been defined. Please, limit the use of acronyms.
- Line 156: I do not see the reason to include both the gradient of SLP and the winds. At 1deg resolution, they are likely the same fields, and this would be overfitting the model. Please, discuss.
- Figure 4: units are missing in the legends.
- Lines 253-254: by errors, do you mean the uncertainties in the maximum likelihood adjustment of GEV? You also show 100-year return levels in the projections.
- Line 271: the underprediction of high storm surges is larger when trained with the hindcast. This can be due to the hindcast having smaller storm surges than the historical simulations. It would be worth checking if this is the case. That would mean that the model extrapolation is biased low. In addition, the climate models have been bias-corrected, adjusting means and variances to those in ERA5. It would also be good to check how the extremes are affected by this bias correction (probably less than the mean storm surges and this would explain these differences).
- Lines 279-280: Is this delta method necessary when the climate models are bias-corrected? I guess no for the mean characteristics of the storm surges,

but extremes could still behave differently (also relates to my previous point above).

- Figure 6: some scales seem saturated. If this is the case, it should be explained in the text (line 289 states that changes are +/-20%).
- Line 302: what does overprediction mean here?
- Lines 326-326: I most of the Mediterranean Sea the statistical approach does not provide reliable results (see my second major comment above), which means that even if the models are consistent, the result is not robust.
- Lines 355-361: The fitting of a GEV using maximum likelihood comes with its uncertainties, that are related to the sample size and its empirical statistical distribution. The increase of uncertainties in the return levels for low-probability events is inherent to the approach, so it cannot be blamed for the decrease in the confidence of the results. The high uncertainties come from the use of a relatively short record (20 years, i.e. 20 maxima). Even with a high goodness of fit of the shape parameter, the uncertainties would increase. Therefore, please, reconsider this text.