

General comment:

The authors have presented a manuscript on an interesting and important topic within the scope of NHESS and certainly has a lot of interest for the international audience. The authors apply storm surge residuals to assess annual maxima return periods and trends in 11 different cities in the northeast Indian Ocean. The work is of appropriate length in terms of written and visual content. The work leans on a freely available data products and applies four different statistical models to cover larger uncertainty range, as opposed to the usual one model approach. The main outcomes from this manuscript are results and discussions on the changes in storm surge residuals for two different periods (1950s and 2000s) covering a relatively large spatial extent. The study region is data sparse (limited observations, no observations or no public access to observations) and evidence on the impact of how different statistical models can lead to differing results are explored. Considering that the work relies mostly on data analysis with known statistical models, the potential value of the work is definitely there. However, there are several aspects of the work that are unclear and needs substantial considerations to be addressed.

Major comments:

mac1: My largest concern is related to data (which ironically is the reasoning behind a work like this). Namely the quality of the reanalysis data used for producing the estimates. The authors cited few of the benchmarks from Aleksandrova et al. (2025), but none of the validation points were in the Indian Ocean, not to mention in BoB or its vicinity. There are also number of studies regarding the quality of ERA5 wind for instance and presumably this quality is not as good in global regions where there is less observational data (especially historically) for data assimilation. I understand this is what the authors try to do, shed some light into these no, to sparse data regions. But it would be highly beneficial to discuss and quantify the shortcomings and conclusions based on these limitations. In some ways applying this work first to some other similar regions with some continuous data measurements might be more beneficial to do your own data validation and prove the multi model ensemble point. Also there is quite a lot of complex coastal regions in most of the regions used in this study - model resolution of 2.5 km is potentially too rough in some of these to capture the surge dynamics with satisfactory accuracy + the accuracy regarding wind and pressure fields, locally highly contrasting disparities are quite likely. It would thus also raise questions, how good is the overall assessment of ERA5 atmospheric data in the NIO region? Is it possible to quantify the data related uncertainty? More substantial work here, would greatly strengthen the underlying proposal of this study.

mac2: HighResMIP experiments are excellent products, but I am not sure I fully understand the authors motivation to use the historical GCM model runs in this study? For evaluation against ERA5, the main usual reasons, among others, would be to e.g. 1) assess the GCM performance and find/assess model biases, 2) evaluate and quantify climatological accuracy (extremes as well) to discuss performance and limitations, and use same model data for future projections (which allows to discuss uncertainty in future projection results, e.g. for SSP5-8.5 scenario etc.). If these GCM analysis have a valid use and role in this manuscript narrative, please make it very clear. Since this work leans on data analysis, I would recommend the authors to make use of the dataset in more full and potentially include future projections as well.

mac3: Regarding trends and return periods. How sensitive is this analysis to outliers and data quality? I am wondering this because the data used here, does have varying limitations (which should be discussed more), and the approach to use a criteria based city virtual stations might not always be the best or correct choice – it might be technically reasonable, but actually can bring in unwanted skewed data that propagates to the whole analysis. When the authors use word “robust”, a more concrete evidence should be provided to claim robustness with higher levels of confidence.

There are several comments here that tie into this point. But interested to hear what the authors think about this – especially if addressing the existing data options differently might increase the confidence in results? Since at the end of the day, dealing with secondary, locally unverified data – which is fantastic that we have at all globally and at this resolution, but does raise various questions how to best go about using it. So ultimately this comment is more about the discussion around the topic raised and interested to hear authors take on it.

mac4: Discussions on results, specifically on model differences. Currently most of the manuscript is presented bit more like report in terms of results and discussions around them. However, discussions around model choices, their use, their differences, limitations and why they are likely performing they way they do, are not presented. As this study also situates itself among the first to consider ensemble approach, this seems like a crucial point. Also, introductions sections does not really establish these models and methods within the larger context. Please consider extending this, especially to emphasize for the readers the new it aims to provide.

Minor comments:

mic1: Lines 85-89. Roughly follow what the authors try to say here, but the logic is a bit lost on me. First, the sentence regarding estimates validation is confusing, please explain in more detail. Second, how do these studies or data products “justify” reanalysis as practical reference for ESL estimation? Currently, I do not see the evidence for such statements. Same with last sentence, yes model ensemble is a brief but solid argument, but the whole sentence is slightly confusing in its meaning. Please check and explain more.

mic2: Lines 91+. I would like to see more contextual information here for the reasons behind uncertainty in single vs ensemble models. For future projections uncertainty is large and so is often model spread. More contextual logic behind the model (and even data) uncertainties would be needed to bolster this line of thought.

mic3: L137. Please check the observation station nr and stats. In their 2025 paper, they were different.

mic4: Please include the model names used in the abstract.

mic4: Please check the language again for consistent and correct use. E.g. L53, “coast” to “coastal”, L130 placement of “the”, L245, L273, L311 etc.

mic5: Fig.1. How did the authors choose the numbers of stations for the cities? And did the authors check whether the nr of stations for any city location skews the results significantly? E.g. if S24 Parganas has 9 stations (with complex and spread out placement) and a place like Chennai has 3 along the seemingly open coast, the near stations setting and nr is something that should in most cases influence any analysis. Would be interesting to hear more about this and get authors insight into choices made and their potential impacts on analysis. Bit is explained in the paper and shown in Figure S2.

mic6: Please check the references for formatting issues. E.g. the Aleksandrova et al (2025) reference has some artefacts.

mic4: Please explain a bit more about the data used in the statistical models. For example, is the annual surge maxima a single time series maxima from a 10 minute data output? Why annual? Higher temporal resolution assessment would be open doors to quite a few more interesting analysis, specifically regarding shifts in seasonal surge activities.

mic7: For surge maximas at 11 NIO cities, are there no sources that can be used to assess the validity of some of the results? Since ERA5 is reanalysis, the extremes should be reflected in the surge model, and some evidence for e.g. TC-based coastal surge should exist (even if surge proxies are used, as often done post impactful storm surge events). Some level of corroboration should exist by the authors or by other studies using the data in the area.

mic8: L192. Could the authors explain here, why 4000, why discard the first half? What is the procedural logic behind these choices.

mic9: L252. Please explain more why are the models showing such disparities?

mic10: Fig2. There is quite a lot of negative space and difficult to see any of the lines.

mic11: Fig2. Please add description for Tr, and explain more why the dashed line is (seemingly) the same between models in t1 and t2.

mic12: It would be overall good to have a better overview of the surge residual data statistics itself. The authors have only presented their return period analysis, but full data description (related also to previous question about choosing annual max vs higher resolution spread) would be of interest for many readers I believe and sets a better stage for subsequent analysis.

mic13: L275. Understand what is meant by “farther inland” but it sounds like on dry land. Better to use different description for this. Also could the lower surge values at that virtual station be related to global model limitations? The point is deeper in the river estuary, not sure about the locations, but would not assume flat out smaller surge component relative to other points.

mic14: Fig3. And discussions. This brings back to extremes within the data itself. Can the authors explain more about how they really represented in there? For example the 2008 TC Nargis caused a substantial surges in Yangon region (estimated over 2 m, depending on source). How are events like this represented in this study?

mic15: L323-324. As it currently reads, this sentence I would say adds no value, and actually might cause confusion. ERA5 either way cannot be used for validation of future projections, it can however and is very well suited for assessing historical CMIP experiment assessments (whether GCM models can capture climate dynamics). Maybe that is what the authors mean? Please check and clarify. Same for L334-335.

mic16: GCM usage. I think this is confusing the readers (myself included). If authors use a single model, better to stay focused on that and not give much focus on the 5 ensemble talk.

mic17: L335. What is “it”?

mic18: L336-339. I do not follow this conclusion here. Please explain what the authors mean here in relation to results from CMIP6 model?

mic19: Fig.5. In isolation this analysis is not as strong, I would suggest adding here ERA5 results as well for direct comparison, as that is the closest reference for it.

mic20: L362-370. This likely comes from earlier confusions about the GCM use in this study, therefore hard to follow the logic here. Why are MSL trends suddenly added here and why from GCM? What is the 6b ensemble then based on, GCM? And 6a ERA5? Please check and explain in more detail.

mic21: Fig.6a. What is the spread based on, stations in each city? If so, that adds a whole range of potential discussion points and warrants a closer look in interpretations.

mic22: Conclusions and discussions section. In here the content is largely not contained in focused topics of the study done. Please consider adding topic focused subsections. In this type of study it would good to see focused discussions with concise conclusions as separate section.

mic23: ~L380-383. Since the HighResMIP and ERA5 are used by the same people who created the global surge data, it would be good to explain more in the manuscript what are the inherent differences between the two data characteristics. For a GCM to reproduce actual climatology with relatively good accuracy is very good, but depending on variables, this can be very difficult and is a challenge. For example wind, and especially tropical cyclones are one of these things, and more so in the extreme end of things. HighResMIP is something that tries to tackle that gap. So when we talk about its ability to reproduce the extreme end of weather (which directly relevant for surge maxima), it should be discussed and analysed in more detail. If not possible by the authors themselves, then stronger discussion points based on literature.

mic24: L384. Just to confirm, based on what are for example at Chittagong the hazards the greatest, geographically it certainly looks like with high surge hazard potential, but other than current study results, is there more evidence in here? And the bias spread is likely because of the complex coastlines and nr of virtual stations used (e.g. vs. the Colombo, Chennai and Mueang Phuket)? Bit more discussions around this and similar instances would be good to make the results and some of the statements more clear/evident.

mic25: To increase the potential reach of the work for general and outside the field audiences, I would recommend the author to check the overall writing and try to simplify the lingo where possible or add extra nuances to break concepts down. Especially regarding the methodology and conclusions-discussions of the study. This can be either through extending the meaning of concepts or adding additional figure for methodology process.