Review summary:

The article presents a new method to deal with historical information in classical extreme value analysis (EVA) of sea levels. The authors show that their method outperforms the Bayesian Markov Chain Monte Carlo (MCMC) approach, used in several papers to tackle the issue of partial historical information in EVA, in terms of estimating extreme sea levels (ESLs) at Travemünde, Germany. In addition, authors show that the estimation of 200-year return water level (HW200, from which is based the design height for coastal defenses at Travemünde) is larger with their method than the current official value, determined using only tide gauge records (aka systematic data), highlighting a possible underestimation of coastal flooding risk at Travemünde.

Overall, I think this is a good paper. It is well written and the objective, method and results are presented clearly. However, the article would benefit from some complementary information to better assess the novelty and relevance of the results.

I detail my review below, separating major from minor comments.

Major comments:

2.1 Extreme value analysis

- L118-120, equation(2): strictly speaking, the Generalized Pareto distribution is a 2-parameter distribution. The location parameter μ described by the authors is in fact a threshold fixed by the user. When it comes to fitting the GPd to some data, only 2 parameters are estimated in the process. In contrast, the Generalized Extreme Value distribution (GEV) has 3 parameters. There may be a confusion here, I suggest to make it clearer. Furthermore, the support of the distribution is $x > \mu$ (strictly) when $\xi > 0$ (strictly) and $\mu <$ (strictly) $x <= \mu-\sigma/\xi$ when $\xi < 0$. $\xi = 0$ corresponds to the particular case of exponential distribution.

2.2 Increasing available sea level information

- L146-173: Three methods for the incorporation of historical extremes with systematic observations are presented. I regret that the main reference here is in German (DWA, 2012) thus I was not able to check and better understand methods 1 and 3. Actually, as the only method further used in the article is method 2, I suggest not mentioning the other 2. If authors still want to keep this description, then I strongly recommend to test and compare these additional 2 methods in the study.

  Equation(4): I think there is a mistake here as the presented formula is based on a GEV distribution and not a GPd. The likelihood formulation for a GPd is more complex, see for example Bulteau et al. 2015 https://doi.org/10.5194/nhess-15-1135-2015. In a GPd framework we deal with peak event probabilities, while in a GEV distribution framework we deal with annual exceedance probabilities (in case of annual maxima). Authors should clarify this point which could lead to strong impacts on results.

3.1 Extreme sea levels at Travemünde, Germany

- L184: Authors write “The official HW200 value is 224cm above NHN”. I have a concern about datum references in the text. I believe that all values in the remaining article are given in meters (or centimeters) above mean sea level (MSL) - see section 3.2 Methods. If the authors’ statement is correct, I suggest converting the HW200 value in cm above MSL so that we can

compare it more easily with values in Table 1 for instance. However, it seems that the value 224cm is used in section 4.2 (see for instance L416) working with the GPd whose parameters are estimated based on systematic and historical data expressed as values above MSL. So I am a bit confused: is the statement at L184 wrong or is there a mixing of values expressed with respect to different datum references?

- L190-191 and L197-198 : Authors indicate that two measurements from nearby Lübeck are included in the historical dataset in order to make it larger. This is not discussed in the discussion section whereas that could generate more uncertainty in ESL estimates. Have authors tried to perform the analysis without these two measurements? I suspect that the first one (3.10-3.20m above MSL observed in 1320) has a huge impact on results. A sensitivity analysis could be performed. At least, this point should be discussed. Also, is there an argument or a study that would give credit to this merging of historical observations from 2 different sites into a single historical dataset? Are sea levels comparable between the two sites?

3.2 Method

- L255-262: I do not understand clearly this paragraph. I think another figure or table showing the problem of large variance between high-end ESLs as mentioned by the authors would help. Moreover, the authors write "this variance can be reduced if we further assume that no higher water levels occurred between any two consecutive observations". Do they mean "between two any consecutive *historical* observations"? If so, this should be clarified.
  In the process, authors replaced "any artificial ESLs which exceed [the moving threshold] with a randomly generated value sampled from the most current intermediate distribution": what if the new value still exceeds the threshold? Clarification is also needed about the moving threshold: for example let's say we have ESL1 larger than ESL2. Between these two consecutive events, is the perception threshold equal to ESL1 or ESL2?
  Authors indicate that in the classic Bayesian MCMC-MLA approach, a perception threshold is constrained to a single value for the entire historical record. This is true but it is also known that what matters most in an EVA combining systematic and historical data is not the number of historical events but the length of the historical period and the fact that we have an exhaustive dataset above the perception threshold (see for instance Payrastre et al., 2011 https://doi.org/10.1029/2010WR009812). Actually, it might lead to better results to set a higher perception threshold and ensure exhaustivity even if the historical events dataset must be reduced accordingly. Sensitivity tests would be interesting to compare authors' method with MLA with a perception threshold equal to 2.5m/MSL (1304 event) or higher. That way, the exhaustivity requirement for MLA would be better fulfilled and the comparison would be fairer.

3.3 Comparison to Maximum Likelihood approach

- Whereas the beginning of that section belongs well to global section 3.Data and Methods, I suggest to move the part from L312 to L345 to a new Results section.
- L311: in Equation(6), k is the number of estimated parameters. Once again, there are only 2 parameters in a GPd, not 3.
- It is not clear to me what is the value of the perception threshold used in the MLA. Clarification is needed. As mentioned above, tests with different perception thresholds would be interesting to conduct.

4 Results and discussion

- This global section should be renamed “Discussion” and all content related to the results should go to a new Results section. The Results section should contain: 1) results of comparison between MLA and the proposed method, 2) estimations of HW200 using the proposed method and comparison with current official values.
- L361-363: This statement about non-stationarity in EVA incorporating historical data seems to be in disagreement with section 4.3 Outlook. From the one hand authors say methods are not yet capable of mixing non-stationarity and historical information, and on the other hand, they claim that theoretically there is no obstacle. Clarification is needed.
- It is important to highlight that the proposed method cannot apply to uncertain data. Only historical data with known values can be dealt with as the estimation of GPd parameters is performed the classical way (besides it is put forward by the authors in the conclusion as an advantage). However, historical information is often partial and uncertain. In many other places in the world, one can only access to lower bounds of historical ESLs or ranges of values (see for example the 1320 event in Table 1), and exact values are rare. Only a Bayesian framework is able to deal with this issue and to properly incorporate uncertainty in the EVA. This could be a limitation of the study and the proposed method as it may not apply everywhere. This point should be discussed.

Minor comments:

- I suggest the title of the article to be changed. It does not reflect the novelty claimed by the authors in the text. Moreover, it is not shown nor proved in the article that ESLs estimates are *improved* while incorporating historical information: authors themselves say that their method lead to *larger* estimates than current official values underlining the crucial assumption of stationarity to end up with this result, assumption which cannot be confirmed. In a risk prevention perspective, that might be laudable, but it cannot be said that ESL estimates are improved as we do not know that for sure.
- L15: Authors write “This paper introduces a new method for the incorporation of historical information in extreme value analysis which outperforms other commonly used approache**s**.” In fact, only the Bayesian MCMC - or MLA approach is compared with the authors’ method. This sentence should be revised accordingly.
- L17: Authors use terms such as “posterior” and “prior” distributions. These are commonly used under a Bayesian probabilistic framework. As this is not the case here, I suggest rephrasing to avoid confusion.
- L50: dash is missing (100-year return period)
- L110: there is a mistake in the definition of i : i is the rank of the event ranging from largest to smallest, or in other words in descending order, not the opposite.
- L115: “(…) due to the Pickands-Balkema-de Haan Theorem.” A reference should be added here.
- L121-129: Authors discuss the possibility of increasing available data to perform an EVA by using sea levels outputs of numerical models in addition to tide gauge data. They expose some issues such as long run times and forcing factors that must be sound to get relevant results. In addition, authors could mention the issue of combining heterogeneous data in EVA: tide gauge data and model outputs do not carry the same uncertainty, which can lead to errors hard to quantify in ESL estimates. One possibility to deal with this, is to explicitly incorporate uncertainty on modeled values in the EVA through a Bayesian framework (see for example Nicolae Lerma et al., 2018 https://doi.org/10.5194/nhess-18-207-2018).

- Table 1: column 2 should be relabeled as follows “Level (m above MSL)”. The legend should be modified as follows: “A list of historical extreme sea levels measured at Travemünde and Lübeck, Germany) (refs).” Also there are two events occurring in 1867, and this is not in accordance with Figure 1 (one event in 1867, one event in 1868).
- Figure1 & Figure3 & Table1: There is a 15th historical event in Figures 1 and 3 (occurring in 1304) which is absent from Table1.
- L213: Authors may add here that all water levels expressed in the following are expressed in meters with respect to MSL. That would clarify and avoid confusion when datum reference is not mentioned.
- L216: “generalized” should be written with a capital letter “Generalized”
- L222: word should be replaced by the one in bold “(…) the total number of historical observations is taken **as** 1,487, assuming(…)”
- L223: a word is missing “(…) is the same as **in** the systematic period (…)”
- Figure2: y-axis should be relabeled “Water Level (m above MSL)”
- Figure3: how dates are attributed to artificial events? Is the Poisson process modeled in the stochastic generation?
- L291: Here is the first time authors mention the bias between historical and systematic records. Definition should be given here.
- Figure 6: explanation of boxplots is missing in the legend (cf. legend of Figure 7).
- L337: Authors claim that MLA is capable of providing more accurate estimates of HW200 in 26% of simulations, or that their method provides better estimates in 74% of simulations (L332). I do not see where those figures come from.
- Figure 7: Adding a horizontal line at 0% of error on each graph would make it easier to read.
- L367: word in bold must be deleted “(…) high-end estimates rather than **the** those in the (…)”
- L381: a personal communication is used as a reference for the future climate-surcharge value of 100cm. Does this refer to a design constraint specific to Travemünde or German coasts? If so, when will this new value be implemented?
- L406: rephrasing suggestion: “In contrast, no less than 4 events exceeding this height occurred during the **previous 100 years.**”
- L408: rephrasing suggestion: “(…) of the largest ESL event **in the systematic period**”.
- L417: Authors write “Over a 100-year period, the likelihood that no events exceeding this level would occur is approximately 61%”. Is this result obtained using systematic data only ?
- L419: The sentence is not clear. I suggest rephrasing: “Thus, the likelihood that no event exceeding this level would occur within 100 years is reduced to ~21%”
- L423: “a two-fold increase **in** the likelihood”

References:

- Numerous journal names are missing in references (for examples: L475, L477, L482, L489, L500, L504, L510, L513, L515, L518, L523...)
- Sometimes journal name is written entirely sometimes only with abbreviations. Authors must follow nhess standards.