The manuscript ‘Multivariate statistical modelling of extreme coastal water levels and the effect of climate variability: a case study in the Netherlands’ assesses an impact function that can reproduce inland water levels in a human controlled system by event sampling and conditioning the drivers. By modelling the dependence structure between the different drivers to generate paired synthetic events, the authors are able to assess compounding effects of surge and precipitation on inland water levels. Overall, this study is an interesting read and I commend the authors for their nice work. It uses well-established methods and builds on previous assessments. Furthermore, it provides new insights in modelling compounding effects of surge and precipitation, and an interesting analysis of climate variability and using a short subset of the data. The manuscript also provides interesting and detailed information and discussion on underlying processes of the predictor selection and interpretation of the compounding effects of the two drivers. However, in its current form this study has a number of limitations that I would like to see addressed. For instance, the contextualization of using a case study in an area with a high degree of human management is lacking, steps undertaken in the methods section need more clarification, and decisions undertaken in the results need more clarification and transparency. Therefore, I propose to reconsider this manuscript for publication upon revision of the following issues.

Specific comments:

1. The title of this manuscript is framed as a case study that provides a statistical framework for assessing extreme coastal water levels and climate variability that can be used for other case studies as well. By framing the title like this I would expect a discussion in the manuscript that addresses how this statistical framework (e.g. conditioning the drivers) can be used for other areas of interest or even a different region in the Netherlands. This contextualization of how a user can use this framework in other areas of interest is lacking in the manuscript’s current form.
2. The case study in the Netherlands provides an analysis on an area with a high degree of human management. As the title of the manuscript does not cover this, I would suggest to either add this information to the title or add a short discussion on how this statistical framework can be used for other areas which do not have a high degree of human management.
3. Throughout the manuscript, water levels are most often referred to as inland water levels (line 1), however sometimes the authors use solely water levels without the adjective ‘inland’ (e.g. line 99), or extreme coastal water levels as is stated in the title of the manuscript. I would suggest to stay consistent with the terminology and provide a clear description of the water levels (e.g. how much inland, coastal/inland water levels).
4. While the manuscript discusses relevant previous studies in the introduction (line 59-73), the research gap is not pointed out clearly. As a consequence, the novel aspects of this study and the research gap do not come across strongly. Therefore, I suggest adding more detail to the this section in the introduction.
5. In order to improve readability, I would suggest to rephrase line 116 by using frequencies, i.e. more frequent in original data or less frequent in shuffled data.
6. In section 2, data and study area, please provide more background information how the predictors total surge and precipitation were derived. For instance, information how the surge and tide are added (van der Hurk et al., 2015).
7. In line 137-138, you mention that the performance of the impact function is highly sensitive to the selection of the predictors, yet no sensitivity analysis or the degree of sensitivity is reported or shown. Please provide more information and details on how sensitive it is.
8. Please contextualize, if possible, why the annual maxima surge events are at least 3 days (5^{th} percentile), see Figure 1b and lines 174-176.

9. In lines 176-178, please provide more contextualization about the tradeoff and why the minimum total surge is selected.
10. In Table1, the selected predictors of the two cases are reported, taking into account the three aspects mentioned in lines 141-145. Additionally, information is provided for the selection of predictors in lines 159-161. Please provide more information about the optimization technique used. Why was the maximum (next to the minimum and mean) for the conditioning not included? Which approach was used for this during conditioning of the drivers (MLR, MLRbin, ANN, etc.)? Is the performance of the predictor selections evaluated on the metrics used throughout this study, or the tradeoff between the metrics and visual inspection of the events that exceed the flood warning level as in line 222-223? What is the time step of additional hours prior to the event used for this selection? Were all possible combinations of the selected time steps and statistics evaluated or was an optimization technique used for this (e.g. random search)?
11. In line 191-193, please provide short details on which architecture and hyper parameters are used for the machine learning approaches.
12. Like equation 2, are the predictors in equation 3 for the 3D case also standardized?
13. In lines 250-253, please provide more information on what terms the 3D case generally does not outperform the 2D case. To me it seems that the 3D case performs better on the reported metrics. Above the flood warning level, the differences looks only marginal (confidence interval of the 1000 bootstrap runs not reported). If the focus of this manuscript is on the distribution of extreme cases above the flood warning level, then it should be clearly stated in the manuscript. Additionally, lines 366-367 report that adding complexities does not necessarily improve performance. However, the reported metrics show an improvement. In lines 367-368 the authors report that the performance between the two cases differ slightly for higher return periods. Why did you choose to not report metrics (e.g. MAE) of those extremes of extreme events? Moreover, lines 440-441 report that the 3D case did not lead to an overall improvement. Pleas provide more information to the respective section why those decisions are taken and on what basis (e.g. define overall in overall improvement).
14. Line 291-292, please provide more information or give possible examples on the underlying physical processes
15. In line 297-298, you mention that separating the analysis in seasonal clusters did not lead to an improvement, but do not report to what extent. Please provide more information to the respective section. Additionally, in line 324 you mention that separating the statistical analysis in tidal clusters did not lead to an improvement. Please specify to what it did not lead to an improvement.
16. The section about seasonal variability evaluates the dependence structure of the predictors and reports the Kendall's rank correlation for the respective seasons. This is a very interesting read and discussion, however the authors report in line 298-299 that the spread of annual maxima events is uneven and that for some months few events occur. Have the authors considered restructuring the inland water levels maxima in seasonal maxima, resulting in 800 maxima inland water levels per season?
17. Please provide contextualization on the results reported in lines 327-330.

Technical corrections:

- In line 124-132, it would improve readability to also refer to the respective sections in the methods for the different steps of the conceptual model.
- In line 200-202, extreme water levels exceeding 0 meter is used to describe the higher end of the water levels, however it would be more sensible for the reader at this stage to refer to in percentiles or flood warning level as indicated in the sup.

- In line 260, do you mean ‘inland’ water level?
- Line 338 misses a word: ‘in the following …’.
- Line 366 Fig S14 should be Fig S12
- Line 372 now reads as if empirical analysis consists of 100,000 events. Please rephrase.
- In the caption of Fig 6, transparent should be added to ‘Green illustrates the uncertainty …’
- In the caption of Figs 6 and 7 c, d, and e don’t match up.
- Table 4 subpanel e, to my understanding the copula of 800-year ensemble should be marked and not the total surge of 800-year ensemble.
- Please clarify the sentence in line 414-415 starting from ‘hence’.
- In the supplementary information line 45 contains a duplicate of ‘a’.
- In the supplementary information lines 51-52 reports difference between shuffled dataset and dependent dataset while using the same symbols. As a suggestion the authors can use for example $A_{C_{max,shuffled}}$ in order to increase readability.
- The caption of Fig S7 contains a duplicate of ‘that’.