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:

We, the authors, thank the anonymous reviewer for providing a thorough and comprehensive review. We acknowledge their suggestions will help us improve the quality of the manuscript and hope the suggested amendments satisfy their concerns. Below we provide our point-by-point response to their 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.

The statistical framework can be used to other areas of interest, given the availability of relatively long overlapping records of flooding drivers and impact variable. Defining appropriate impact-based predictors that optimize the performance of the impact functions depends on the hydrological characteristics and management of a given system. For systems with low to no management, we would expect a more straightforward construction of an impact function. In any case, composite (average) plots can guide this process where appropriate lags between drivers and impacts should be accounted for. We will add a paragraph in the revised manuscript to discuss this transferability as well as limitations. This study did not include wave-driven water levels (i.e. wave set up). This is a reasonable assumption in the shallow Wadden Sea (sheltered by barrier islands), where surge is the main flooding marine driver. In other locations, the wave contribution, or other drivers such as snow melt, might need to be considered as well. Additionally, we did not account for low-frequency

variations of water levels such as sea level rise, which would need to be considered if results where to be extrapolated into the future.

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.

This information will be added to the title and will also be discussed in the main text. A new tentative title is "Statistical modelling and climate variability of compound surge and precipitation events in a managed water system: a case study in the Netherlands".

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).

We agree with the reviewer about the ambiguity of our previously used terminology. We will modify the terminology throughout the paper in the following manner: coastal water level, to refer to water elevation on the coastal side; and inland water level, to refer to reservoir levels. We hope it is clearer now.

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.

The last paragraph of the Introduction mentioned the novel aspects of our study. However, to emphasize the novel aspects of our study in relation to existing gaps, we will add more details in the last paragraph of the Introduction, as suggested by the reviewer. We will also change the title including two novel aspects: 1) compound analysis in a managed water system, 2) sensitivity of such analysis to climate variability. Another novel aspect of our study, which will be also explained in the introduction, is the analysis and interpretation of the correlation coefficient of impact-based predictors.

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.

We prefer to keep the term return period rather than frequencies, as it is a well-known term that is used in many contexts: risk analysis, impact assessment, infrastructure design, etc. However, we will rephrase the paragraph to improve readability.

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).

We will provide more information about how total surge (coastal water level in the revised manuscript) and precipitation were derived in Section 2. In a nutshell, surge is calculated from wind speed via an empirical equation that was previously calibrated for the study area. The tidal time series is an artificial extension of the standard astronomical equations and was calculated using all known current tidal constituents for a complete period of 800 years. Total water level is the sum of surge and tide.

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.

We will add a few examples in the Supplementary Material to illustrate the sensitivity of the impact function performance to the selection of predictors. To do so, we will generate some plots assessing the performance of the impact function for different predictor choices (for example: Min Surge 12, 24, 36, 48, 60h, CumPrcp 5d, 12d; Max Surge 36h, Mean Surge 36h, etc.). The inland water level in this specific location is more sensitive to variations in storm surge than to variations of precipitation. Hence, the sensitivity analysis will focus on storm surge.

8. Please contextualize, if possible, why the annual maxima surge events are at least 3 days (5th percentile), see Figure 1b and lines 174-176.

It is reasonable to assume that the relevant duration of storm surge is 3 days as three days is the mean duration of cyclones over East-Central Europe (Bartoszek, 2017). Note that from Figure 1b we cannot conclude however that events linked to annual maxima surge events are at least 3 days long. The shaded area depicts values from 5^{th} to 95^{th} for each time considered. However, storm surge time series of individual events are not necessarily parallel to the mean storm surge, or lower/upper envelopes. As the lower envelope of storm surges is not necessarily linked to a single event, the associated duration below 3 days does not necessarily have a probability below 5 %.

Reference:
Bartoszek, K. (2017) The main characteristics of atmospheric circulation over East-Central Europe from 1871 to 2010. Meteorology and Atmospheric Physics, 129, 113-129.

9. In lines 176-178, please provide more contextualization about the tradeoff and why the minimum total surge is selected.

We will rephrase the text to avoid confusion. Our case study is a water-managed system. To try to prevent inland water levels from reaching extreme levels, the system is regulated by opening the gates around low tide. However, if the coastal water level at the low tide is too high, gates cannot open, water cannot be released, and inland water level might increase due to precipitation and river runoff. Therefore, the annual maximum water level better relates to the previous local minimum of coastal water level (i.e. total storm surge), rather the local maximum, as shown in the composite plot (Fig 2). This differs from a natural water system in which the maximum or mean storm surge would probably be a better predictor to describe extreme water levels. In fact, when the tide and surge are separated, we found that the mean surge (and not the minimum surge) is a better predictor. The choice of 36 h is not randomly selected as a value in between 72h and 12h (from the 3D marine predictors) but is the result of optimizing the impact function performance, for which a wide range of time lags were tested.

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)?

We will add more details of the performance assessment and optimization. We tested a wide range of predictors, mean, max and min values, for different time lags. This selection was guided by the composite plots and physical understanding of the water system. The revised manuscript will provide some representative examples.

11. In line 191-193, please provide short details on which architecture and hyper parameters are used for the machine learning approaches.

The learning process of the artificial neural network used here is based on stochastic gradient descent, and the activation function is the sigmoid function. The architecture of the network is as follows: input layer with 2 (2D case) or 3 (3D case) neurons; 2 hidden layers with 8 neurons each, output layer with 1 neuron.

The number of trees in the random forest was set to 50, after performing a sensitivity analysis assessing the overall performance of the approach (estimated as root-mean-square error (RMSE) via k-fold validation approach) depending on the number of trees. We selected 50 because increasing the number of trees beyond that value did not lead to an increase in performance.

We refrained from using additional (more sophisticated) machine learning approaches and testing other architectures because we achieved a reasonably good performance using regression for most return periods by means of bin-sampling. We believe this approach is easier to implement, which can aid transferability. We will include this information in the manuscript.

12. Like equation 2, are the predictors in equation 3 for the 3D case also standardized?

No, equation 3 for 3D case is not standardized. The standardization was only shown in Eq. 2 to preliminary demonstrate the importance of total surge (coastal water level in the revised manuscript) to drive extreme reservoir levels, as compared to precipitation.

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).

The 3D case performs better based on the reported metrics for the impact function, but the increase in performance compared to the 2D case is minimal so we opted to showcase the simpler case. The focus of the manuscript is on the distribution of extremes but not necessarily above the warning level, as very few events are available exceeding the warning level and this leads to high sensitivity in return period estimates. We will rephrase these results to make our interpretations clearer.

In our manuscript, lines 366-367 report that adding complexities does not necessarily improve performance based on a visual inspection of Figure 5 and Figure S12, instead of reporting metrics. By doing so, we refer to performance when reproducing the return period curves from van den Hurk et al. 2015, i.e., we are not only referring to performance of impact function (where the metrics the reviewer is referring to are reported) but to the entire multivariate statistical framework. In lines 367-368 we also report differences in performance based on a visual inspection, and our decisions to report not overall improvement when using a 3D case are performed on the same basis. We chose this way of reporting findings as we expected small differences in the metrics and opted to shorten these paragraphs by doing so. We agree with the reviewer that reporting appropriate metrics will help clarify our findings. By means of RMSE, calculated by comparing empirical estimates from van den Hurk et al. 2015 and our

equivalent estimates (i.e., taken estimates of WL associated to the same return periods), we do acknowledge a small overall improvement in the 3D model: while the RMSE for the 2D case are 0.0202 and 0.0202 m (dependence and shuffled, respectively), the RMSE obtained in the 3D approach are 0.0189 and 0.0187 m. Regarding performance at the higher return periods, we feel reporting metrics could be highly deceiving, as the empirical estimates above the warning level are not enough no ensure stable statistics and, as reported in our manuscript, there is high sensitivity in the tails, which depends on the number of the available events. We will include the overall metrics and edit the paragraphs containing this information in the manuscript. We hope we addressed the Reviewer's comment properly.

14. Line 291-292, please provide more information or give possible examples on the underlying physical processes

The inter-seasonal variation of the correlation coefficient linked to annual maximum water levels results from the marginal distribution of non-conditioned precipitation and surge. For example, in winter precipitation tends to be lower, so extreme water levels are mostly surge driven. Differently, in summer the likelihood of heavy precipitation increases, which increases the chance of compound surge and precipitation leading to extreme water levels. We will rephrase this information to make it clearer in the manuscript.

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.

We separated the 800 annual events into seasons and into clusters (defined as a function of tidal range), respectively. None of these options led to a better performance of the impact function (in terms of RMSE, MAE and correlation coefficient) and return levels (visual inspection). We will add more details in the revised text, as suggested by the reviewer.

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?

We considered performing a seasonal analysis but eventually discarded this option for two reasons. We wanted to replicate the results of van den Hurk et al (2015) where annual maxima is used. Also, using seasonal maxima might lead to consider non-extreme water level events, which are not the focus of this study.

17. Please provide contextualization on the results reported in lines 327-330.

These results are intended to give an overview of the effects of climate variability in the estimation of the correlation between predictors. We divided the dataset into subsets of 50 years and assessed the correlation for each subset. We found that correlation varies significantly among 50-year subsets and shortening the dataset can often lead to not sufficient data to get statistically significant correlation estimates. We will rephrase this paragraph to make it clearer.

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 ACmax,shuffled in order to increase readability.
- The caption of Fig S7 contains a duplicate of 'that'.

We thank the reviewer for pointing out these technical corrections. We will address them in the manuscript appropriately.