Replies to Reviewer #3's comments on "Partitioning the uncertainty contributions of dependent offshore forcing conditions in the probabilistic assessment of future coastal flooding at a macrotidal site ". (nhess-2021-271)

Reviewer #3:

General comments:

I would like to thank the authors for revising the manuscript and taking into account my comments. Unquestionably, this is a quite improved version relative to the previous one. However, there are still some points that I would like to be clarified before accepting this manuscript for publication.

In general, I am happy with how the authors responded to my comments on the dependency structure and the duration of the events. With respect to the structure of the paper I think there is still room for improvement and Figure 5 could be a bit clearer.

The main point that is still not clear to me is the way the Monte-Carlo simulation is performed to obtain the events used for the flooding probability simulations. Bellow I have more specific comments on this. Additionally, I would still like to see a different way of calculating the error statistics of the meta-model.

We would like to thank Reviewer #3 for the valuable and constructive comments. We agree with the suggestions regarding the Monte-Carlo-based approach as well as regarding the validation of the meta-model. Therefore, we have modified the manuscript to take on board the comments. We recall here the reviews and we reply to each of the comments in turn.

Specific comments:

Figure 3:

The blue colour for the selected events might be a poor choice with respect to visibility, against the black colour of the observed events.

My main observation though is on the simulated events, which are still puzzling me. That is why I asked to see the histograms in my previous comment.

You mention:

- *Line 223: "Step (1) Fitting of the marginals of 'amplitude' variables through the combination of the empirical distribution, below a suitable high threshold u, and of the Generalised Pareto distribution (GPD) above the selected threshold u (Coles and Tawn, 1991) using the method of moments."*
- *Line 347: "Following Step (1) described in Sect. 3.3, the extracted data are used to fit the marginals of the 'amplitude' variables using the GPD distribution with the selected threshold value uHs=6.2m, uSkew Surge=0.48m, and uU=18.9m/s corresponding to ~2 extreme events / year. The marginal distributions are provided in Supplementary Materials B."*
- *Line 355: "Note that some delineations (on the bottom left hand corner) can be noticed, which results from the threshold-based procedure to model the probabilistic distributions (see Sect. 3.3)."*

First of all, I would expect that the thresholds values you provided would match the delineations seen in the figure, which from what I can see in Figure 3, are around 4 m, 13 m/s and 2.3 m for Hs, U and SWL respectively.
The simulated events seem to be quite dense above these thresholds (i.e., extremes events), while on the opposite side there are not that many events bellow the threshold. In essence, you are simulating more extreme than mild events, which I would not expect to be the case. Please explain if I am missing something here. This would mean that your probabilistic analysis for the flood volume is affected as well.
I have mentioned this issue in my previous review as well, but I did not recieve any response from the authors on this.
We thank Reviewer 3 for these new comments. In the light of this new analysis, we now have a better understanding of the issue raised by Reviewer 3 in the first round of review. We apologise for only partially addressing this issue. We have rechecked the different steps of the Monte Carlo procedure and have come to the same conclusion as Reviewer 3, namely that the random samples below the dependence threshold misrepresent the bulk of the multivariate distribution.

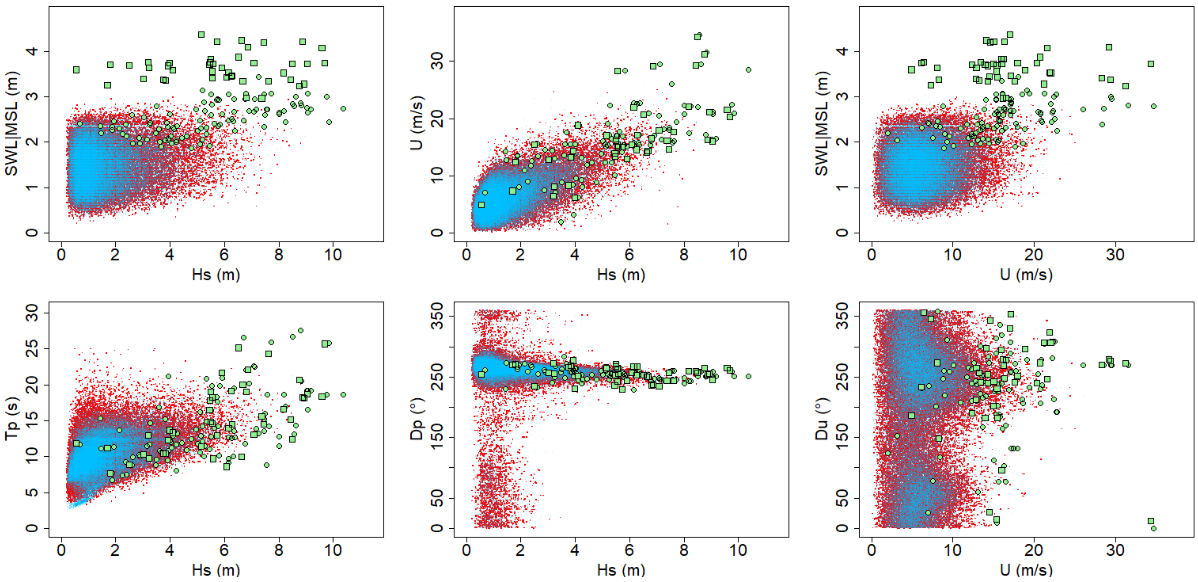
Three remedies to this problem have been applied:

1. To generate the random samples following Heffernan and Tawn (2004)'s approach, we now use the R package *texmex* (https://cran.r-project.org/web/packages/texmex/index.html) that is more stable than our in-house implementation. A clear improvement of stability was noticed when using Laplace instead of the Gumbel margins (see Keef et al., 2013);

2. We also notice that the problem particularly arises when the POT threshold is chosen too high. We propose to minimize the subjectivity in this selection by using the automatic threshold selection procedure developed by Northrop et al. (2017). This yielded lower values than the ones in the original version of the study, namely u_{Hs}=3.59m, u_{SWL}=2.37m, and u_{U}=9.51m/s;

3. In the original version of the study, we distinguished the skew surges and the tide. Given the long time series (~100 years of hindcast data), the difference with the direct approach (i.e. directly using the total water level) was reasonably small and we focus on this latter approach. The advantage is now to avoid a rescaling of the generated samples i.e. to avoid the delineations on the *SWL-Hs/U* plot, which were confusing.

New figure 3 shows the new randomly generated samples (with modifications of the colors to better indicate the different datasets as suggested by Reviewer #3).



New Figure 3: Overview of the N=50,000 randomly generated samples of offshore conditions (red dots). Blue dots correspond to the hindcast conditions used to fit the statistical methods described in Sect. 3.3. Green dots and squares correspond to the n=144 training data used to set-up the GP metamodel (the selection approach is detailed in Sect. 3.2). The squares correspond to cases that are deliberately selected outside the range of the red dots to cover a broader range of situations.

We added in Supplementary Material B diagnostic plots to confirm the goodness of the GPD fit. In addition, we provide in Supplementary Materials B the quantile-quantile plots to confirm that the generated samples adequately reproduce the marginals (see below Fig. S2-S5). We also updated the dependence measures' analysis (Supplementary Materials A).

On this basis, the whole global sensitivity analysis was re-conducted.

- As expected, the major change was a reduction in the flooding probability value (new Fig. 7) which is clearly consistent with Reviewer #3's analysis;
- The new sensitivity analysis (new Fig. 8) reveals that our previous conclusions remain quasi-similar with differences mainly corresponding to Shapley effects' changes of approximately of +/- 10% for the major sources of uncertainty (*SLR* and *SWL*) and a steeper time evolution;
- The new analysis of robustness (Sect. 5) still outlines the key role of the DEM but differs in the very long term (by 2200) with a lower influence of the RCP scenario. Supplementary materials E was also updated with these new results.

It should be mentioned that whatever efforts are made to tune the extreme value analysis, residual uncertainties will remain, mainly due to the sensitivity to the threshold selection (POT and dependence). This problem is clearly underlined in the discussion section (Sect. 5.2), and we believe that incorporating such uncertainties is a valuable line for future research.

Finally, although the extreme value analysis is an ingredient of the study (mainly based on previously published papers), it is the Shapley effect combined with Gaussian process metamodels that is new. This is the main message of our manuscript (see the abstract). We believe that the performance analysis (see below) as well as the physical significance of the sensitivity analysis (Sect. 4.3, Sect. 5.1 and Supplementary Materials E) give confidence in the value of this approach from an operational viewpoint.

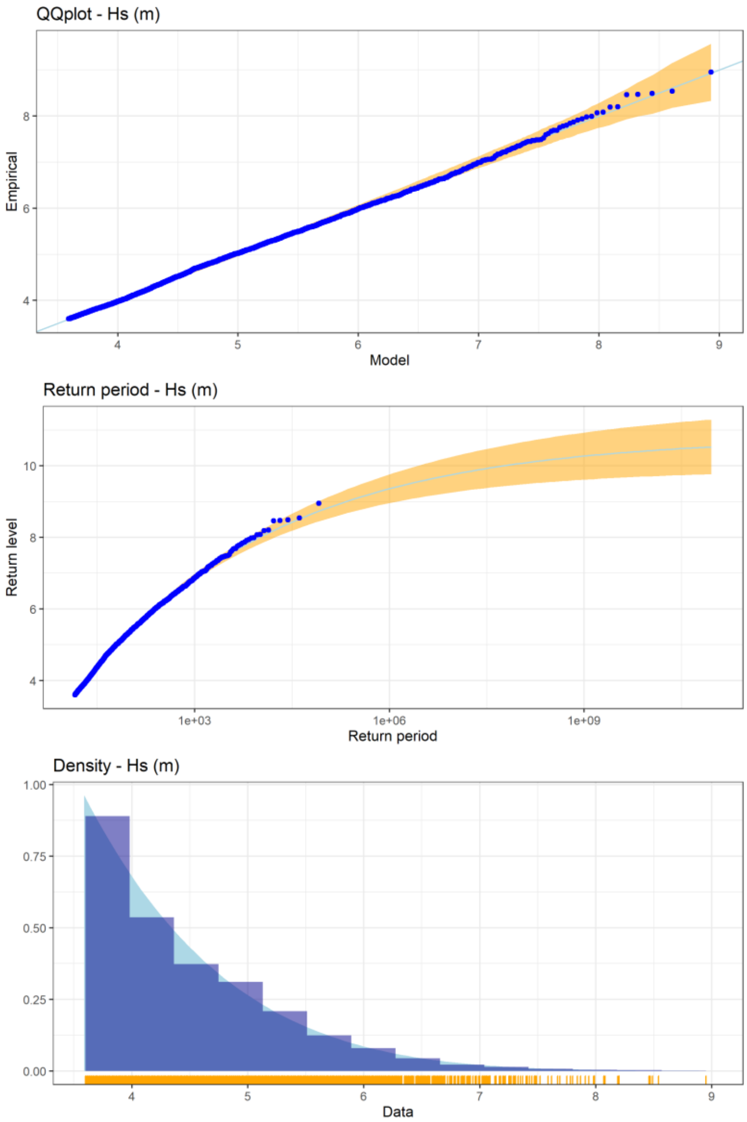


Figure S2: Diagnostic plots for the GPD fitted to *H*s. (a) Quantile-Quantile plot (the band is the 95% confidence envelope); (b) Return level plot (the return period is expressed in number of events); (c) Comparison between the empirical and the theoretical density distribution.

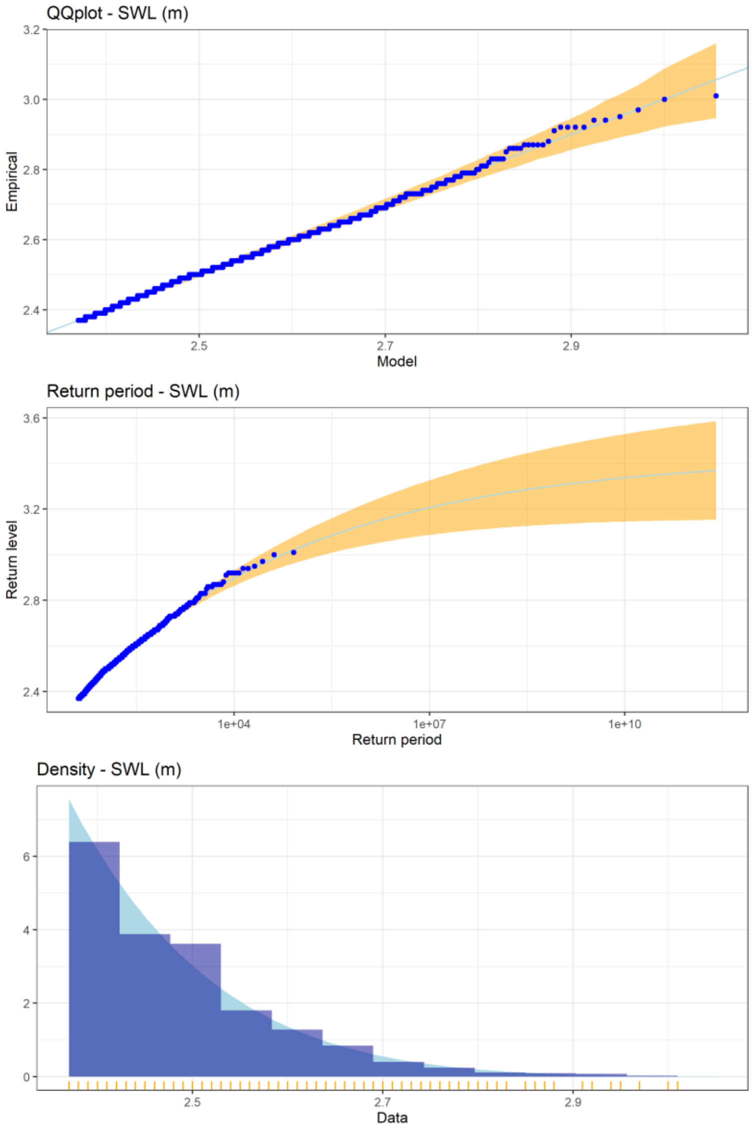


Figure S3: Diagnostic plots for the GPD fitted to *SWL*. (a) Quantile-Quantile plot (the band is the 95% confidence envelope); (b) Return level plot (the return period is expressed in number of events); (c) Comparison between the empirical and the theoretical density distribution.

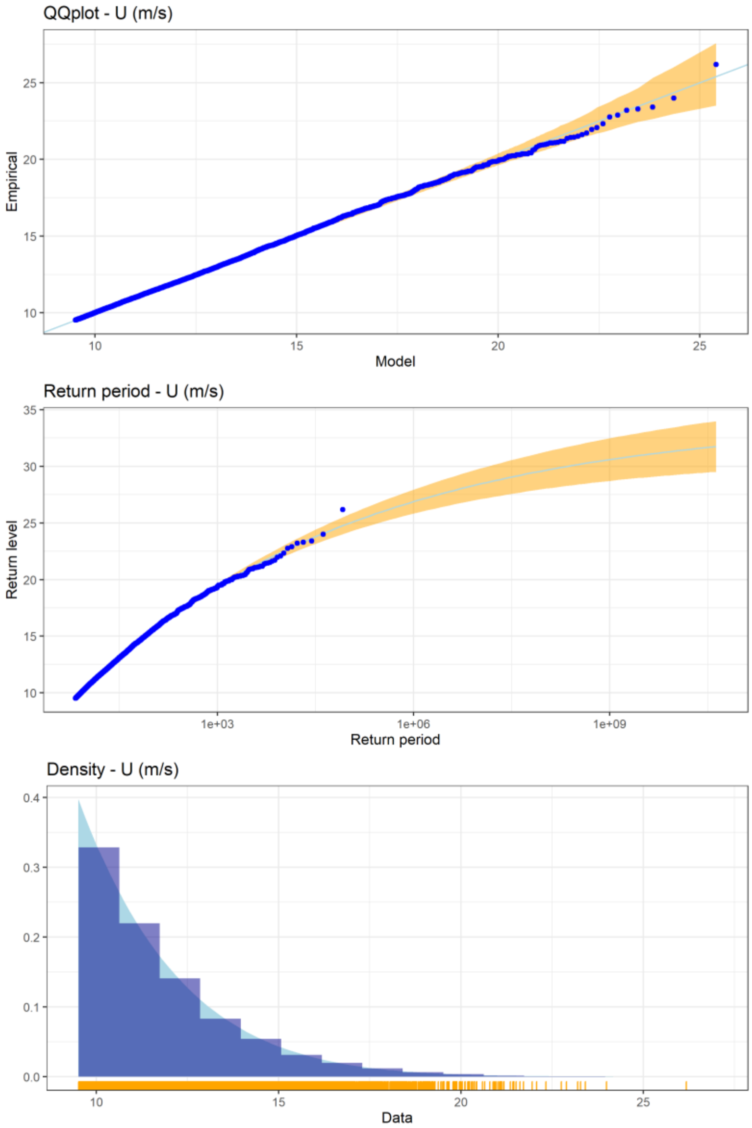


Figure S4: Diagnostic plots for the GPD fitted to *U*. (a) Quantile-Quantile plot (the band is the 95% confidence envelope); (b) Return level plot (the return period is expressed in number of events); (c) Comparison between the empirical and the theoretical density distribution.

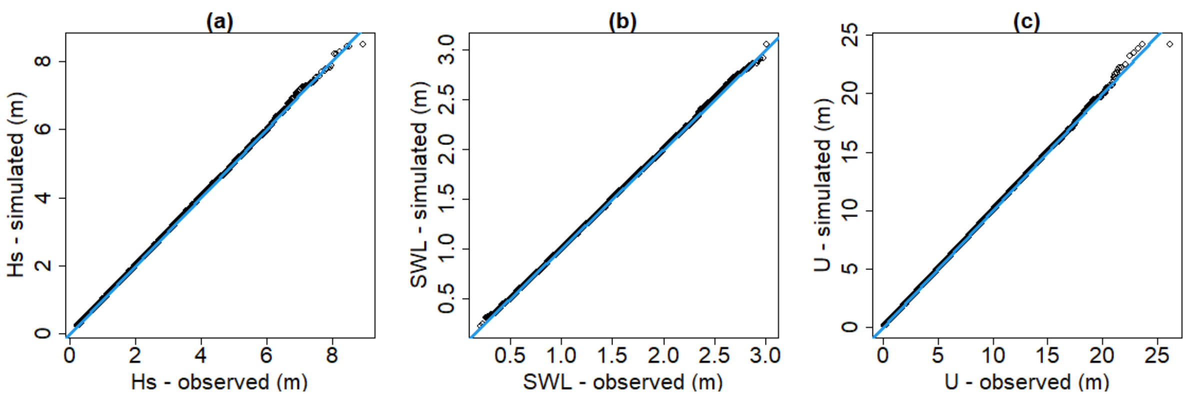


Figure S5: Quantile-Quantile plot considering the observations (hindcast) and the N=50,000 randomly simulated samples for *Hs* (a), *SWL* (b), and *U* (c).0

References

Keef, C., Papastathopoulos, I., and Tawn, J.A.: Estimation of the conditional distribution of a multivariate variable given that one of its components is large: Additional constraints for the Heffernan and Tawn model, Journal of Multivariate Analysis, 115, 396-404, 2013.

Northrop, P. J., Attalides, N., & Jonathan, P.: Cross-validatory extreme value threshold selection and uncertainty with application to ocean storm severity. Journal of the Royal Statistical Society: Series C (Applied Statistics), 66(1), 93-120, 2017.

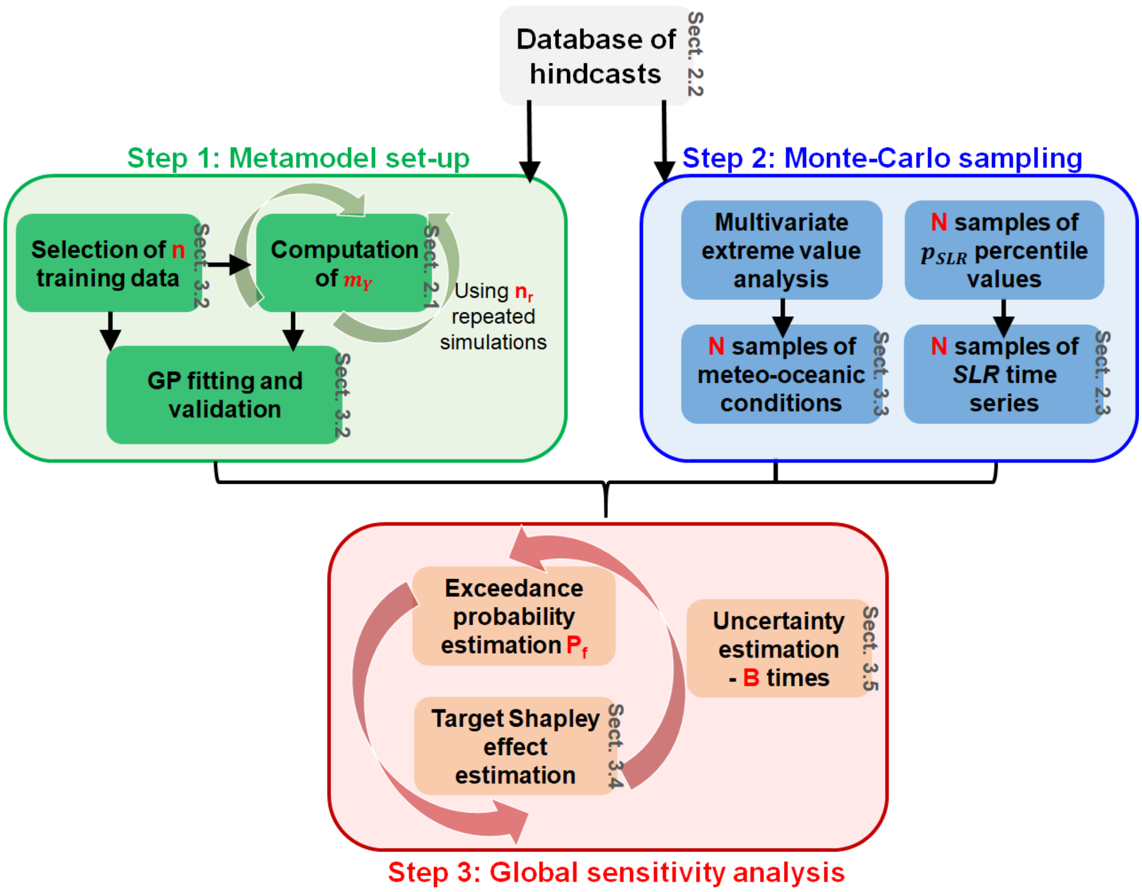
Figure 5:

It is still a bit challenging to navigate this figure. For example the n training data used in STEP 1 are subsampled by the output of STEP 2. I would expect that a change of order in STEP 1 and 2 will make this clearer.

I think it will help to add some general titles in the steps as well, like for example "Meta-model for flood volume estimation", "Monte-Carlo sampling", and "Global sensitivity analysis" for the current steps 1-3.

It is not explained (and I do not understand) why some boxes have more intense colours.

Thank you for these suggestions that will clarify the readability of Figure 5. A new Figure 5 is proposed (see below). Regarding the issue of reversing STEP 1 and 2, this was actually a recommendation during the first round of review and we choose to keep this ordering.



New Figure 5: Flowchart of the procedure. The sections describing the methods/data are indicated in grey next to the boxes.

Figure 6:

Thank you for clarifying why there is only a single value for Q2 and not a mean with an uncertainty range. Yet, the method you are using to calculate that (Hastie et al., 2009, which I noticed is missing in the reference list in the end of the manuscript) is used in Hastie et al., 2009 to calculate an estimate of the prediction error (based on a loss function). While here you are applying this to calculate an error-statistic (Q2) that is affected by the variance of the observed (modelled) data, which since you are merging all the k-folds together is quite high. Hence, I would propose to calculate Q2 for each of the k-folds, and then present the average values with an uncertainty band. I am curious to see if the mean Q2 will be as high as 99.2%.

This is an interesting point of view and we agree with Reviewer #1 that an 'analysis per fold' has many advantages. The description in Sect. 3.2 has been re-written in this sense and the presentation of the results have been reformulated by focusing on the statistics of Q^2. See new Figure 6.

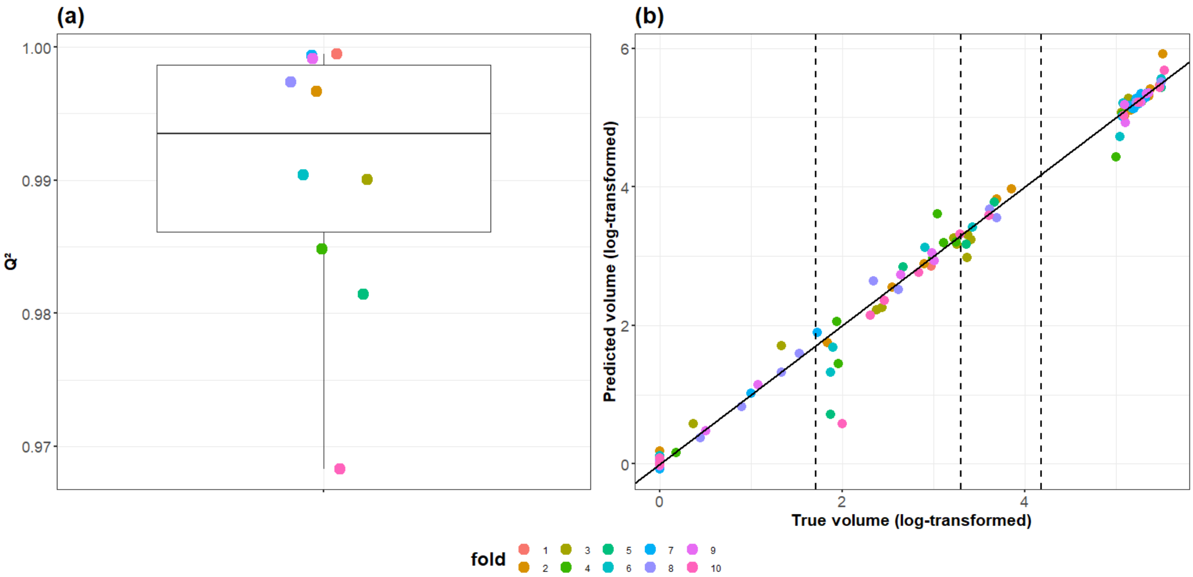


Figure 6: (a) Boxplot of performance indicator values Q^2 (for the 10 folds of the cross validation procedure). Each colour indicates the number index of the corresponding fold. (b) Comparison between the volume (log_{10}-transformed) estimated using the "true" numerical simulator and the ones predicted using the GP model for each of the 10-folds of the cross-validation procedure. The closer the dots to the first bisector, the more satisfactory the predictive capability of the trained GP model. The vertical dashed lines indicate the threshold Y_C (log_{10}-transformed) used in the study (50, 2,000 and 15,000 m^3).

Table S4 (Supplementary Materials E) has also been updated.

Table S4. Statistics of Q^2 (calculated across the folds) of the different GP metamodels calculated using a 10-fold cross validation procedure.

GP model	Median of Q^2 (%)	Interquartile of Q^2 (%)	Analysis
Median Q_{50}, DEM 2015	99.35	1.25	Sect. 4
3rd quartile Q_{75}, DEM 2015	99.22	1.12	Sect. E4
1st quartile Q_{25}, DEM 2015	99.45	1.73	Sect. E4
Median Q_{50}, DEM 2008	95.97	3.81	Sect. E3

Finally, we have added the missing reference (Hastie et al., 2009).

Minor comments:

Line 351: "...*threshold υ of Eq. (2)..." I imagine this is equation 4 now*

Thank you for noticing this mistake. It has been corrected.

Line 353: *"N=50,000 events (representative of 1,000 years)." How do you scale the number of events to number of years?*
Since we focus on the flooding probability estimate and not necessarily the return period of the forcing conditions, we have removed this aspect as well as in Sect. 4.3.

Orleans,
July 4th, 2022
J. Rohmer[1] on behalf of the co-authors

[1] BRGM, 3 av. C. Guillemin - 45060 Orléans Cedex 2 – France