General comments

Pianosi et al. examined dominant input uncertainties on flood-induced annual loss, at Rhine River basin in Europe and Queensland state in Australia. They used a JBA Risk Management's flood loss model to conduct a global sensitivity analysis. Overall, this study represents a contribution to large-scale flood model community and has the potential to provide the model developers to improve model performance by showing the key sources of uncertainty.

Reply: Thanks for appreciation of our paper and for contributing to the discussion!

I believe that the research methods and the obtained results are described clearly. One of my concern lies in the characterization of uncertainty presented in Table 1. First, I do not fully understand the rationale for assigning a ±50% range to the nonlinear value of the return period. Furthermore, the return periods used in hazard maps and loss calculations are generally predetermined based on river improvement standards as well as the economic and geographical conditions of the site. Therefore, I find it difficult to grasp the reasoning behind discussing uncertainty by varying this value itself.
With regard to the flood event set, I can accept the idea of examining the uncertainty range of the event generation model. However, in determining the ranges of parametric and structural uncertainty in the flood inundation model, I believe that some evidence-based justification should be provided. For example, one might reasonably consider the typical error ranges of global models in precipitation data, evapotranspiration processes, or river discharge calculations.

Reply: the rational for sampling the return periods of hazard maps is described in Sec. 4.2. Essentially, this is an efficient procedure to produce a different flood hazard map for each model execution, while maintaining spatial consistency over the simulated domain. This is not the only possible way to produce spatially consistent maps, and other approaches are described in the same Section. Obviously, sampling maps through sampling of their return periods does not allow capture of any structural uncertainty in the flood inundation modelling that produced those maps (hence the comment in Table 1 under "Aspects not captured by the characterisation"). As for the choice of a plus/minus 50% range, this range comes from considering uncertainties reported in previous flood frequency analyses for the Rhine rivers (at various gauging station). The related literature review is reported in detail in the PhD thesis referred in Sec. 5.1 (Sarailidis 2023) but, based on this and other comments from other Reviewers, will be summarised and included in the revised manuscript.

Still, we agree that the definition of these bounds remains highly uncertain, and although we used some evidence (previous flood frequency studies) to define those bounds, we would not claim that the bounds adopted here are definitive – given the current state of the field. We discuss this problem in Sec. 4.3 ("Characterising poorly bounded input uncertainties") and we highlight its relevance in our Outlook and Conclusions (lines 455-465) – one of our key conclusions is precisely that the selection and characterisation of input uncertainties is the most critical step in setting-up the analysis.

Based on this and other comments and reviews we will consider how to further reinforce the point in a revised manuscript.

Since variations in the vulnerability curve directly influence the loss calculation, I believe that the method for determining the range of uncertainty should also be explained with due care. As for the damage ratio, I assume it is determined based on structural types, for instance, distinguishing between above-floor and below-floor inundation. Regarding the lower bound of the depth range, is it assumed to correspond to the height of the first floor? In Table1, DR_min can be set to zero, but does this indicate that, even at the same inundation depth, cases where no damage occurred are taken as the basis for assigning zero damage? Providing explanations for these points would enable a more informed assessment of whether the results are valid.

Reply: Yes, vulnerability functions vary by line of business, occupancy type, coverage type and peril. And yes, the initial step in the vulnerability function is meant to represent the height at which water can enter a building (the threshold height of the property) and therefore the point at which damage can occur. This may be the height of the lowest doorstep of external doors, or the height of features such as airbricks, or at which pipes enter the building. In the revised manuscript we will clarify these points.

Finally, the most novel aspect of this study appears to be the simultaneous evaluation of multiple sources of uncertainty. However, because the adopted approach counts only the most influential factor in each comparison, the current method of defining uncertainty ranges reveals that the vulnerability curve contributes the largest share, followed by the flood event set. Yet, the description is insufficient regarding how these uncertainties ultimately affect AAL or LE. For example, if the vulnerability curve were increased by 30% (on the conservative side) and AAL recalculated, by how much would the AAL value change? Would it be by 10%, or by a factor of three? We can see part of the answer from Figure 7 regarding to the AAL in the case of Queensland, but explicitly addressing the magnitude of such differences in the target outcomes would, in my view, greatly strengthen the analysis.

Reply: we are not sure we fully understand the comment, and particularly the sentence “the adopted approach counts only the most influential factor in each comparison”. If “counts” means “considers”, this is incorrect because the GSA approach does not require us to make any assumptions about which factor is most influential – this is the result provided by the GSA, not its starting point! (But maybe we misunderstood the point?) Regarding the mapping of how variations of input are reflected in variations of output, we agree that this a very useful insight provided by GSA, and that is why we showcase it for the case of Queensland. We will consider whether we can add a similar picture for the Rhine River basin, although we also need to strike a balance between different visualisation tools that we want to show in a limited space.

Specific comments

Fig.1: The upper part of Fig. 1 is almost identical to the text and therefore cannot be considered an effective schematic figure. The lower illustration comparing "precise" and "accurate" is easy to understand, but I would suggest reconsidering whether this figure is truly necessary.

Reply: Our intention with this Figure was to provide a visual summary as well as a complement to the text. We take note of this comment and will see what other reviewers think of this before deciding whether the Figure is useful or not!

Fig.3, caption: I did not find the use of colored boxes particularly effective. In the figure, the hazard map and event set are shown as inputs to the "interpolation" box, but in reality, the flood depth is selected from the chosen event set. Therefore, I believe there may be a more appropriate way of illustrating this process than labeling it as interpolation.

Reply: Flood depths are not selected from the event set; return periods are. The model then looks up a depth for that return period from the hazard maps. Given that there isn't a map for every return period, the depth is taken from a log-linear interpolation between the depths in the two maps that are closest in return period order (as shown by the inset image below the interpolation box). For example, if an asset is affected by an event of 180 years return period, then its flood depth will be estimated (green arrows in the sketch below the interpolation box in Figure 3) using the depths from the 100 and 200 years maps (brown dots), using a loglinear relationship (black line).

L294: Is there any rationale for setting the value to 0.3? How would the results change if a different value, for example 0.5, were used instead?

Reply: Tricky comment! No there is no particular rationale for the value of 0.3 other than looking at some example rankings at different locations and noticing that a threshold of 0.3 would reasonably reflect the qualitative judgement we can make based on visual analysis. A way to assess the impact of this choice is obviously to repeat the analysis for different threshold values and see if results differ (a sensitivity analysis of the sensitivity analysis!).

L338-349: You have evaluated which factors influence the basin as a whole, but is it appropriate to assess locations that experienced the most severe damage—for example, areas with greater inundation depth—in the same manner as those less affected? For flood-prone areas, might we expect different results, such as certain uncertainties having a greater impact? Since Fig. 6 already organizes the analysis by different return periods, I thought it might be useful to expand the explanation in connection with that figure.

Reply: we are not sure what the reviewer means by "*You have evaluated which factors influence the basin as a whole*". One of the key points we want to showcase with the Rhine River basin example is that we can perform a spatially-distributed

GSA where sensitivities are computed (and visualised, as in Fig. 5) for each spatial unit, and not for the whole basin. Perhaps this should be better clarified in the manuscript. Of course, a Figure like Fig 5 shows differences/similarities in AAL sensitivity across space but does not show the spatial variability of AAL. We agree this second piece of information is very important too and can think of ways of including it within the same figure. Again, the problem is striking a balance between the amount of information embedded in a figure and the need to keep the figure sufficiently readable and accessible!

L411-427: Could it be that the aggregation method employed here is problematic? In particular, the fact that the residential portion changes so drastically gives the impression that there may be an issue with the approach itself. If you have any ideas or suggestions for how this could be improved, adding a brief comment on that point would, I believe, be valuable.

Reply: There are many ways to disaggregate an aggregate portfolio, all of which have pros and cons. The changes seen in the Queensland case study are a consequence of the characteristics of the portfolio used (including the variability of the total insured values), how vulnerability functions are attributed, and how large an area aggregate exposure is disaggragated to (as explained in the paper) and therefore the range of events the locations are impacted by. JBA continuously tests and improves different disaggregation methods internally and it's a difficult thing to do! We wanted to include portfolio resolution in the GSA because it's still common to receive fairly coarse portfolio information, so understanding how portfolio resolution contributes to loss uncertainty is important for risk managers and could be used to lever better data in the future.

Technical corrections
L121: Is "$L_{s,t}$" a typographical error for "$L_{s,k}$"?
L1599: Is "(FDs,t) a typographical error for "(FDs,k)"?

Reply: yes, thank you!