The manuscript 'Global sensitivity analysis of large-scale flood loss models' applies the GSA method to two case studies. As the main author has published about GSA in the past, the main contribution are the case studies. However, I see several issues with how those are described and evaluated, as already also mentioned by other commenters here. It reduces the value of an otherwise interesting and potentially useful publication, but I believe the authors will be able to address all issues in the revision.

Reply: We wish to thank the reviewer for their constructive criticisms, which we hope will help improve the clarity of the paper. Detailed responses on how we intend to do this are given below.

I would particularly stress the importance of comments by the other reviewer, Francesco Dottori. I concur that the lack of description of the JBA flood model is an issue to be addressed. No mention of flood protection levels is also troubling, given that it is often found to be the most sensitive parameter (see e.g. https://doi.org/10.5194/nhess-18-2127-2018).

Reply: We agree with both Reviewers that the description of the flood inundation model needs reinforcing, and we will do this in the revised manuscript. In particular, in Section 2 we will add a description of how flood defences are represented in the model and in Section 5 we will clarify that not varying the associated protection standard means we are likely underestimating the uncertainty in the hazard component. We will also include the reference suggested by the Reviewer, although that paper focuses on coastal floods – investigating the relative importance of standard of protection uncertainty across different types of floods would indeed be a very interesting topic for future research.

Also, the inconsistency of uncertainty ranges in Table 1 is problematic: vulnerability curves are assumed to have wide margins, much more than the flood event set and hazard maps (which are given arbitrary +/-50% range) or exposure (which has about +/-25% range). This rather unsurprisingly leads to conclusion that vulnerability is the most sensitive component for the Rhine. Better explanation of the choice of uncertainty ranges is needed.

Reply: In the revised paper, we will provide more details on how the uncertainty ranges where defined and provide a summary of the literature review fully described in Sarailidis (2023, Chapter 3) that led to defining such ranges.

I would highlight that the Rhine catchment was subject to a GSA study before, which is not mentioned by the authors (https://doi.org/10.5194/nhess-18-3089-2018). There, a major sensitivity of impacts to changes in flood protection, land use and reservoirs is highlighted. The authors of that analysis also used well-designed uncertainty ranges with a strong rationale from literature and empirical data. It would be important to compare and discuss those results in context of the results presented here.

Reply: Thanks for pointing us to the paper by Metin et al 2018, which is very relevant for our work, and we will include it in the revised manuscript.

The paper resonates with our own work in two key points: it is motivated by the "*lack of knowledge of how and to what extent changes in influencing factors propagate through the chain and finally affect flood risk*" because most of previous research has focused on uncertainties in flood hazard assessment, instead of risk; it concludes that "*components that have not received much attention, such as changes in dike systems or in vulnerability, may outweigh changes in often investigated components, such as climate*". This is very much in line with our own findings that (1) GSA literature has focused on drivers of hazard rather than risk (lines 90-100); (2) uncertainty in vulnerability and exposure are found to be more important than uncertainty in hazard in our case studies (lines 447-450).

On the other hand, there are also several differences between our work and Metin et al 2018. Some are technical (for example, they use a different sampling strategy, which we can add in Sec. 4.2 "Generating physically-plausible input samples" – along with a discussion of its pros and limitations). Some other differences are more substantial:
1) Different conceptual framings. Metin et al 2018 use GSA as a tool to learn about drivers of changes in risk in the real system, under the (implicit) assumption that the model is a valid representation of that system. Instead, in our work we use GSA primarily as a tool to learn about the model, to understand the key drivers of uncertainty in model predictions and thus inform model validation and future improvements. Both are legitimate uses of GSA but in revising the manuscript we shall make sure that we highlight these different ways of using GSA.
2) Different scale of investigation. Metin et al 2018 focus on a sub-basin of the Rhine, the Mulde catchment, covering an area of about 7,000 km2. This is much smaller than we investigate in our paper (185,000 km2 in the Rhine and 1,700,000 km2 in the Queensland application). Hence, we would not put their work under the umbrella of GSA application to "large-scale" flood risk modelling, which is our specific focus as increasing scale raises specific challenges that we aim to discuss in our paper (see lines 93-103).

In the revised manuscript, we will include the suggested reference and use it to put our results into context of previous studies but also highlight conceptual and methodological differences.

The case studies are implemented with very different methods and presented also completely differently. With the selection of model components and the choice of uncertainty distributions being quite arbitrary, the paper shows that there is a wide possibility of pre-selecting an outcome of any "validation test" (mentioned in the introduction), hence going against the authors' final sentence; "We hope this paper will provide motivation as well as practical ideas to foster the application of GSA to flood risk models and contribute to increasing their transparency and legitimacy." The challenge remains to fairly evaluate different models components with realistic uncertainty ranges and indeed identify the elements that matter: not only flood protection, reservoirs, land

use, vulnerability beyond depth-damage functions, but also inundation modelling method (e.g. https://doi.org/10.1029/2024EF005164, model resolution (e.g. https://doi.org/10.1029/2023WR035100 or hydrodynamic parameters (e.g. https://doi.org/10.1016/j.coastaleng.2024.104541). An improved discussion on this would be beneficial, given that the title, abstract and introduction suggests a more comprehensive, overarching study.

Reply: We agree with the reviewer that GSA results depend on the selection of which input factors to vary and how. This is indeed a very important point that we stress multiple times in our paper while also giving a range of practical suggestions on how to tackle it (e.g. lines 245-260, lines 455-465).

However, we do not believe that "*there is a wide possibility of pre-selecting an outcome of any validation test*", as given the complexity of the models investigated here, reverse-engineering the problem to select input distributions that lead to a desired GSA outcome sounds implausible (if a modeller could do that, they would probably know the model behaviour so well that they would not need to embark in a GSA in the first place!).

This said, we reiterate that a degree of subjectivity in the definition of what qualifies as "realistic uncertainty range" is unavoidable, and GSA results should always be interpreted as answers to the *relative* question "what are the controls of model outputs *if these inputs are varied within these ranges?*". This point is also made in the paper suggested by the Reviewer (Metin et al 2018) who acknowledge: "*The presented results are subject to limitations related to the flood risk chain model and the subjective assumptions for the reasonable change scenarios*".

In the revised paper, we will further highlight this point. We will also stress again that the key contribution of our paper is not in providing insights about any of the specific analysed case studies but rather in discussing generic methodological issues and possible solutions when applying GSA to large-scale models – and case studies are presented for illustration purpose rather than for giving definite answers on the model's predictions for those specific study regions.

Minor points:
Several figures include underlining of errors created by MS Office

Reply: Thanks, we will check this

CRESTA acronym is explained after one of the subsequent uses, not the first one.

Reply: Thanks, we will revise this

L339: what level of CRESTA units is used here?

Reply: high resolution CRESTA