Response to Referee 1

We would like to thank the referee for the time and effort put into reviewing the manuscript. This response carefully addresses all the comments. Where applicable, changes are proposed to the manuscript accordingly. Following the guidelines of the NHESS Editorial Board, the revised manuscript was not prepared at this point.

The paper addresses an interesting research topic, evaluating the utility of using ensembles of multiple flood damage models to improve loss estimations and quantify the related uncertainties. The work is well structured and generally well presented. The analyses carried out provide good evidence of the benefits of using multi-model ensembles compared to the application of single damage models. I believe that the manuscript is worth of publication in NHESS, provided that Authors address a few issues.

Main points

- although the paper reads well, I believe that a better structure could improve its usefulness for the reader. In particular, most subsections of Section 3 include first a description of the work carried out, and then the results (e.g. Section 3.1 begins with the description of the methods used to build and evaluate the ensemble of models, followed by presentation and discussion of results). I would suggest to separate the method descriptions from the presentation and discussion of results, putting them in different sections; this would improve the readability of the paper and make easier the consultation.

We agree with the Reviewer that the structure of the manuscript can be improved. After careful consideration, we propose the following changes:

- Section 2 will be renamed to “Setup of validation exercise”. A new subsection called “Evaluation methods” will be added. This will briefly describe and provide references on the evaluation methods adopted in the study, specifically: RMSE, MBE, CRPS and the Rank Histogram. These changes will aid overall clarity, as well as improve readability of Section 3.4.

- We believe that apart from the evaluation methods, the remainder of Section 3 follows a line of thought where the justification for each step of our study is made clearer after the previous step has been presented, applied, and discussed. In order to facilitate consultation and improve readability, we instead propose to restructure this section as follows:

3. Ensemble construction and evaluation
    3.1. Model rating
        3.1.1. Method
        3.1.2. Application
    3.2. Ensemble-mean performance
        3.2.1. Based on model rating
        3.2.2. Based on simulated non-informativeness
    3.3. Probabilistic application
        3.3.1. Skill and reliability
        3.3.2. Loss estimation

- How did the Authors select the models for their work? The paper by Gerl et al (2016) reviews a larger number of models, so the Authors need to explain the criteria applied for their selection.

We agree that this is a relevant question, which requires addressing in the manuscript. From the paper by Gerl *et al.* (2016), we first pre-selected all deterministic flood vulnerability models describing loss to the asset type this work focuses on (residential buildings), and then excluded models based on following criteria:

- The documentation is insufficient for model implementation;
- The model uses explanatory variables that are not available in most practical applications;
- The model has a functional form that is judged inappropriate in the light of the state of the art in flood loss modelling (e.g. too simplistic or discretised);
- The model is based on the same underlying dataset of another model deemed more appropriate for the application settings (this is to ensure model independence and avoid potential biases in the ensembles).

We propose to add this explanation in Section 2.1.

- An important point that I miss regards models availability. As a matter of fact, several flood damage models are hardly usable in practice, either because not accessible (e.g. commercial models), or because the publicly available information is incomplete and does not allow application (e.g. some research models). Are the models selected by the Authors freely accessible? This would be a major point to foster the use of multimodel ensembles as recommended by the Authors.

The documentation of all models implemented in this study are openly accessible. References are provided in Table I, such that readers may consult the specific formulations of the models. We agree that this is an important point, which we will emphasize in the revised manuscript.

Minor points

- title of Section 3.3 is not much informative for the reader, please change it.

This will be addressed according to the reply to the first question.

- Please indicate the measure unit in Tables 3 and 5.

This will be added (Table 3: million €; Table 5: €).

- Table should be numbered according to the order of citation in the text.

This will be corrected.

- Descriptions at page 11, lines 12-30 are not completely clear to me (e.g. I did not understand what measure is used to build the rank histogram), could you please add some more details?

In the manuscript, we included this brief explanation to aid clarity, as readers from the flood loss modelling community may not be familiar with the Rank Histogram. However, because this is a widely used and well-documented method to assess the reliability of ensemble members, we think that describing this particular method in additional detail would be outside the scope of the manuscript. Two key references where additional information can be consulted are included.