We would like to thank the referees for carefully reading our manuscript and for their constructive comments and suggestions. Please find below our point-by-point responses. If a change in the manuscript will be made, we explicitly say so and present the new excerpts below in red.

Reviewer 2:

> "This paper describes the development of a Bayesian multilevel model for flood damage estimation. These two step models first group observations by event, flood type or region and then build separate models. The study showed that grouping by flood type is most useful for developing transferable flood damage models. The study seems to be carried out well and the writing is generally good."

We thank the referee for taking the time to comment on our manuscript and offering constructive suggestions. Please find below our answers to each point raised.

R2-C1 One of the main conclusions seems to be that when developing transferable flood damage models it works best to select models by flood type rather than by event or region. This observation is very interesting but this is based on a dataset of just German data. I can imagine that in a more international setting the regional difference might become more important than the flood type differences. I think this needs to be emphasized in the conclusions and I think the paper should therefore more promote the method than the finding (which I expect to be specific to this dataset of a relatively homogenous region).

A: We agree that our results are informed by the detailed data we have about flood losses in Germany. We emphasised this in our revision (please see below). Yet we point out that the regional variation that our data cover are quite heterogenous. Since urban and land-use planning follows defined administrative and legal guidelines, buildings codes, for example, are constructed differently in different parts of Germany, partly also because of historic reasons. Wagenaar et al. (2018) developed two flood loss models for different countries (Germany and the Netherlands) and tested how well these models could be swapped between countries. They found that the number of flood events in the data was more important than only the number of datapoints from a single event. We expanded on this approach by training models on data from different flood-event years, different flood types, and different regions, thus allowing for a broad range of environmental, administrative, and socio-economic conditions that we treat explicitly as grouping levels in our analysis. The topic of transferability was also addressed by the first Referee. Therefore, we have added the following paragraph at the end of the Discussion:

"When addressing transferability, we seek models that can generalize well and go beyond local or case-specific data. Wagenaar et al. (2018) trained two flood loss models using data from two different countries (Germany and the Netherlands) and tested how well each model could predict losses in the other country. They found that the number of flood events in

the data was more important than simply the number of reported flood loss cases. Although we trained our models with data from a single country, the data used by Wagenaar et al. (2018) for Germany, comprises six event years across twelve federal states, four river basins (Danube, Rhine, Elbe, and Weser) and four flood types. We expanded on this approach by training models on data from different flood-event years, different flood types, and different regions, thus allowing for a broad range of environmental, administrative, and socio-economic conditions (representing at least Central Europe) that we treat explicitly as grouping levels in our analysis. We argue that exploring these model variants provides more clarity about whether we should use simple average models or more specific multi-level models to be able to transfer predicted loss estimates to new regions, flood types or other structures in the data."

R2-C2 Can you maybe explain better why you go for a multilevel approach rather than just adding variables like flood type, region and event to the dataset? You can then use variable importance to see how much these variables add. In other words can you clarify the added value of this approach better compared to this obvious/simpler alternative approach?

A: In a previous study by Mohor et al., (2020), we explored with simpler statistical flood-loss models the differences across flood types. We found that slopes and intercepts differed across flood types, while a complete pooling (or average) model had varying intercepts. However, both these approaches overlooked potentially informative structure in the data, for example, the role of flood types, timing, regional characteristics of building codes, or measures of flood preparation. With the multilevel modelling under a Bayesian framework, we trained regression models with varying intersects and varying slopes that duly and explicitly recognise these differing characteristics. One major added value is that the multilevel approach expresses these differing characteristics as individual model components and how they deviate from the average model trained on all the data. The multilevel approach allows us to analyse all data in one model while honouring structure or nominal groups in the data. Thus, the training of the group-specific parameters occurs at the same time so that model parameters can inform each other by means of specified (hyper-)prior distributions. This approach warrants more training data than running stand-alone models on subsets of our data, which in turn are more prone to over- and underfitting and overestimates of the regression coefficients. Given we do have an identifiable structure in our dataset, we see these advantages as welcoming, if not necessary. We extend our presentation of the method explicating these advantages and justifying our method choice, by adding the following to Line 93:

"Bayesian multilevel models weigh the likelihood of observing the given data under the specified model parameters by prior knowledge. Bayesian models thus express the uncertainty in both the prior parameter knowledge and the posterior parameter estimates. The multilevel approach allows us to analyse all data in one model while honouring structure or nominal

groups in the data. Thus, the training of the group-specific parameters occurs at the same time so that model parameters can inform each other by means of specified (hyper-)prior distributions. This approach warrants more training data than running stand-alone models on subsets of our data, which in turn are more prone to over- and underfitting and overestimates of the regression coefficients, ~~. A multilevel structure allows for partial pooling such that each level or group can learn from the others by shrinking the posterior regression coefficients towards the pooled mean,~~ while reducing effects of collinearity, and offering a natural form of penalised regression (McElreath, 2016)."

R2-C3 In the first sentence of the abstract you note that preparedness is typically ignored. I agree with this statement but its not really what this paper is about and by adding it to the first sentence of the abstract you confuse the reader. So I advice moving this statement.

A: Thank you for this observation. We agree and changed the abstract accordingly:

"Models for the predictions of monetary losses from floods mainly blend data deemed to represent a single flood type and region. Moreover, these approaches largely ignore indicators of preparedness and how predictors may vary between regions and events, challenging the transferability of flood loss models. We use a flood loss database of [...] "

R2-C4 Maybe also mention synthetic models in the introduction.

A: We will reinforce this topic in the introduction. Synthetic models are a good approach to harmonize loss estimation. However, when it comes to including behaviour they are limited by their assumptions. In general, synthetic models tend to reduce (natural) variability of data and are rarely validated (Sairam et al., 2020). We added the following text to the introduction:

"In contrast to empirical models, synthetic models are developed based on expert opinion and offer a good approach to harmonize loss estimations. However, how these models rely on assumptions is problematic when preparedness and other behavioural variables are concerned. In general, synthetic models tend to reduce the variability of data and remain rarely validated (Sairam et al., 2020). Therefore, we train our Bayesian model using reported data. "

Reference: Sairam, N., Schröter, K., Carisi, F., ... & Kreibich, H.: Bayesian Data-Driven approach enhances synthetic flood loss models, Environmental Modelling & Software, 132, 104798. https://doi.org/10.1016/j.envsoft.2020.104798, 2020.

R2-C5 Line 26: The introduction frames that having a lot of detailed information automatically leads to overfitting and reasons that you therefore need multi-level models. This is not necessarily true, overfitting can be controlled in almost all data-driven methods

so its possible to produce more general models with detailed data. Multi-level models are just another way of doing this not the only way.

A: We argued for a balance between too generalized and too detailed models. We agree that multi-level modelling is not the only way. Indeed, we wrote that "multilevel or hierarchic models offer a compromise […]" (Line 30), meaning that there are of course alternatives. To clarify, we added to this section the importance of other strategies, such as feature selection to minimise overfitting by using cross-validation or regularization (the latter is something which our Bayesian approach offers by design). The revised paragraph now reads:

"In this context, multilevel or hierarchic models are one alternative and offer a compromise between a single pooled model fitted to all data and many different models fitted to subsets of the data sharing a particular attribute or group. Bayesian multilevel models use conditional probability as a basis for learning the model parameters from a weighted compromise between the likelihood of the data being generated by the model and some prior knowledge of the model parameters. These models explicitly account for uncertainty in data, low or imbalanced sample size, and variability of model parameters across different groups (Gelman et al., 2014; McElreath, 2016). There are several approaches to the bias-variance trade-off (McElreath, 2020). We conduct a variable selection through cross-validation to achieve a balance between predictive accuracy and generalization. Using priors in the Bayesian framework is using regularization by design and keeps the model from overfitting the data (McElreath, 2020)."

R2-C6 The explanation in 2.2.1 and 2.2.2 is a bit difficult to follow. Could you try improve the explanation, maybe using a figure.

A: Based on the comments of another referee (see R1-C5), we are updating the Tables in section 2.2.2 Model comparison. We added also the following outline of the model selection steps. This new presentation now clarifies our procedure. The following paragraphs replaces lines 156-169:

"On the one hand, testing all models possible without any underlying concept is far from good scientific practice and computationally inefficient; on the other hand, predictors are rarely fully independent. Hence, we fitted candidate models in three steps of model comparison outlined below. We compare the model candidates in each step via the expected log pointwise predictive density (ELPD), which is the sum of a log-probability score of the predictive accuracy for unobserved data. The distribution of these unobserved data is unknown, but we can estimate the predictive accuracy with leave-one-out cross-validation (ELPD-LOO), which is the sum of the log-probability scores for the given data except for one data point at a time (Vehtari et al., 2017; McElreath, 2016). According to Vehtari (2020), an ELPD-LOO difference

>4 may be relevant and should also be compared to the standard error of the difference. Hence, we selected models as follows:

1- We compared models with a gradually increasing number of predictors, based on the prior knowledge of predictor importance reported in a study using single-level linear regression by Mohor et al. (2020). This study considered water depth, for which data are the most widely available and adopted in flood loss models (Gerl et al., 2016) up to a maximum of twelve predictors (Table 1). For example, model 2 (named "fit2") has water depth (WD) and building area (BA) as predictors, while model 3 ("fit3") has the previous two plus contamination (Con) as predictors; model 12 ("fit12") has all twelve predictors (Table 1). The model candidate with an ELPD-LOO difference >4 compared to the previous candidate was selected for the next step.

2 – For the model selected in step 1 – "fit_s1" with predictors $X^{(s1)} = \{x_1, \dots, x_{s1}\}$, we compared models with $X^{(s1)}$ predictors plus one of the remaining predictors at a time, i.e., $\{X^{(s1)}\}$, $\{X^{(s1)}, x_{s1+1}\}$, $\{X^{(s1)}, x_{s1+2}\}$, … , $\{X^{(s1)}, x_{12}\}$. All model candidates that present an ELPD-LOO difference larger than four and with a difference larger than its standard error were selected for step 3.

3 – We compared the model candidates combining the selected candidates from step 2. If, for example, two different candidates $\{X^{(s1)}, x_{s1+a}\}$ and $\{X^{(s1)}, x_{s1+b}\}$ were selected, we compared the model candidates $\{X^{(s1)}\}$, $\{X^{(s1)}, x_{s1+a}\}$, $\{X^{(s1)}, x_{s1+b}\}$, $\{X^{(s1)}, x_{s1+a}, x_{s1+b}\}$. The model candidate with the least number of predictors and an ELPD-LOO difference >4 as well as a difference larger than the estimated standard error was selected eventually.

We compared all candidate models using leave-one-out cross-validation (LOO-CV) with the Pareto smoothed importance sampling (PSIS-LOO), which is an out-of-sample estimator of predictive model accuracy (Vehtari et al., 2017), implemented in the R package loo (Vehtari et al., 2019). "

R2-C7 I think the title of 2.2.2 should be more like model tuning rather than model comparison, because you really use the same model but with different settings.

A: We disagree with this statement. We compare models with different sets of predictors, thus different number of parameters and input data (we maintain the same number of datapoints, but use more predictor variables). Therefore, a better term would be "model selection" and we decided to use this term in the revised version of the paper.

On behalf of all co-authors,

Guilherme S. Mohor