

We thank the referee for carefully reviewing our manuscript. We are very glad with the positive feedback and thank you for the constructive suggestions. Please find below our point-by-point responses. If a change in the manuscript is foreseen, we explicitly say so and present the draft excerpts below highlighted in red.

RC3:

I appreciate the opportunity to review this manuscript. The manuscript presents FlooDEsT, a machine-learning-based tool for rapid building-level flood damage estimation in urban pluvial flood events. The proposed methodology combines survey-derived damage data with an XGBoost model to estimate relative building losses while addressing common operational challenges, particularly incomplete exposure information through a probabilistic imputation framework. The model is calibrated using survey data from several pluvial flood events in Germany and evaluated using independent datasets from Germany and the Netherlands. Overall, I found the manuscript well organised and relevant to the scope of the journal. The topic is timely, as flood damage models specifically developed for urban pluvial flooding remain relatively scarce despite the increasing frequency and impact of these events. I also appreciate that the authors have focused not only on predictive performance but also on developing a methodology that is computationally efficient and suitable for operational applications.

In my opinion, the manuscript is suitable for publication after addressing a few comments that mainly concern the clarification of some methodological aspects and the interpretation of several results.

Overall comments

The manuscript presents a clear workflow, and the validation using independent datasets provides confidence in the proposed methodology. Most of my comments are intended to improve the transparency of the methodological choices and to further discuss the applicability, interpretation and limitations of the proposed model. I do not believe that additional analyses are necessary to support the conclusions; rather, I think that a few aspects would benefit from further clarification and discussion.

Specific comments

Page 1, Lines 16–18. The concluding sentence of the Abstract states that FlooDEsT "tackles critical gaps in damage modelling." While the Abstract introduces several contributions, including the development of a model specifically for pluvial flooding, the treatment of missing exposure data, and improved computational efficiency, it is not entirely clear which of these are referred to by the expression "critical gaps." I would appreciate a slightly more explicit statement of the main scientific and operational gaps addressed by the proposed methodology so that the principal contribution of the manuscript is immediately clear to the reader.

R: Given the journal's abstract length limitation, we address the advancements in the abstract without dwelling onto the gaps too much. These gaps are however common across damage modelling. Models are training to specific predictor variables which are not always available; processing thousands of buildings can take several minutes depending on the model architecture, unfit for a (near-)real-time-forecast. We will try to rephrase the concluding statement of the abstract highlighting the potential real-time application of a validated tailored loss model. Still, we will struggle with the word limit.

“The tool’s computational performance was evaluated in two German case studies, involving about 2300 and 440,000 buildings. The tool provided damage estimates in ~~respectively~~ 2.1 to 5.6 seconds per thousand buildings, representing a 2.7- to 6.6-fold improvement in speed over a baseline approach. FloodDEsT tackles critical gaps in damage modelling, offering valuable support for disaster preparedness, ~~including and response~~ **impact forecasts.**”

Page 8, Lines 155–162. I found the probabilistic imputation framework to be one of the most interesting aspects of the proposed methodology. Nevertheless, I would appreciate some additional justification regarding a few implementation choices. In particular, it would be useful to explain the rationale behind selecting Beta, Gamma and Multinomial distributions to represent the different predictor variables, the motivation for performing exactly 100 imputations, and the criterion used to determine whether the errors introduced by the imputation procedure were considered "acceptable". Although these choices may have been empirically justified during model development, a brief discussion would improve the transparency of the methodology and facilitate its application by other researchers.

R: We are happy to read that the imputation framework caught your interest. Regarding the distributions, discrete variables (classes or scores) were invariably fit to a multinomial distribution. Continuous variables (i.e. water depth and floor space) were tried to Beta, Gamma, Normal and Lognormal and the best one, based on sum of the square errors, was chosen. In the end, only Beta and Gamma distributions were selected among the predictors. The 100-time imputation was empirically decided without further sensitivity testing. The “acceptance” of the introduced error was also empirical, as presented in the manuscript. We will extend the methodology section to be clearer about our development.

“To overcome this challenge, we impute the missing data based on the distribution of each variable observed in the survey data (Figure 1), replicating the patterns observed in SWF events (Sieg et al., 2019). *The imputation through random sampling was set up empirically. Statistical distribution families were decided by best sum of the square errors across Beta, Gamma, Normal or Lognormal distributions, for continuous variables with the data at hand. Other statistical distributions could be further explored for a better fit in future developments or an expanded dataset. [...] The random sampling imputation was set empirically to a 100-sample to cover for uncertainties. A sensitivity analysis of the sampling size could also be explored per predictor variable, given that they have shown different importance (see Fig. 6) to the model.*”

Page 11, Lines 209–212. One result that particularly caught my attention is that floor space emerges as the most influential predictor, even above water depth. Since the model estimates relative building losses, I found this result somewhat difficult to interpret from a physical perspective. It would therefore be helpful if the authors could further discuss why floor space plays such an important role in the model. For example, does floor space itself influence the relative loss, or is it acting as a surrogate for other characteristics that are not explicitly represented in the dataset, such as building value, occupancy, construction type or the value of the exposed contents? Intuitively, larger buildings may experience greater absolute losses, but it is less evident why floor space alone would strongly influence

the relative loss. Clarifying this point would help readers better understand the physical interpretation of the feature importance analysis.

*R: Your interpretation is correct: The floor space is highly correlated with the building value and this explains the lower *relative* loss pattern. The larger the building (the floor space), the smaller is the relative loss, for a given absolute loss. The fact that floor space (fs) comes singularly as more important than water depth might be related to the fact that water depth (wd) and flood intensity (fi) interplay. That is, perhaps the “importance of the hazard information” ended up being split between ‘wd’ and ‘fi’. We will add these two points to the discussion.*

Page 14, Lines 255–259. The analysis shows that, despite the variability introduced by the probabilistic imputation, 80% of the estimates remain within the same damage class. I found this result encouraging. However, I wondered whether these stable classifications are evenly distributed across the different damage classes or whether they are mainly associated with particular categories. For example, are the most consistent predictions concentrated in low-damage cases, or do they also include the highest damage classes? If this information is readily available, a brief discussion of this point would help readers better understand whether the reported stability is consistent across the full damage spectrum or mainly reflects the predominance of certain classes.

R: In this sensitivity analysis we did not check the absolute damage, but the variation across each 100 predictions of each damage estimate. Each estimate (average of 100 predictions) was then classified into one of four damage classes ('Low' to 'Very high'). For estimates classified as 'Low', 'Medium', or 'High', the underlying 100 predictions showed occasional variation across all four damage classes. However, in 81%, 76%, and 71% of these cases respectively, all 100 predictions agreed with the final estimate. In contrast, for all estimates classified as 'Very high', the 100 predictions consistently fell into this class. This demonstrates that while predictions for lower damage classes sometimes vary, they vary similarly and often align with the final estimate, and predictions for 'Very high' damage showed perfect consistency.

Page 15, Lines 265–269. The manuscript states that the seven very high damage cases were not correctly predicted, possibly because they were associated with relatively moderate water depths (up to 80 cm). As a reader, I would appreciate a slightly more detailed discussion of these cases. If relatively moderate water depths alone do not fully explain the observed damage, what building characteristics or vulnerability factors do the authors believe contributed to such high relative losses? For example, were these buildings characterised by particular vulnerability or exposure conditions that are not fully represented by the current predictor set, or do these cases simply reflect rare situations within the available dataset? A brief discussion of this point would help readers better understand the limitations of the proposed methodology when predicting extreme losses.

R: The seven “Very High” cases in the validation set include 3 cases with the highest flood intensity, pre-war era buildings and low implementation of precautionary measures (a maximum of 16 in a score ranging between 0-48), both of which suggest higher damages. Additionally, they included only one-family homes and none of them are masonry, the most frequent type in the training dataset. These two variables are however not included in the model

as predictor variables, which could explain an additional bias. These two points will be added to the discussion.

Page 15, Lines 275–276. The discussion on model transferability could be clarified further. In Section 3.3, the manuscript states that the validation dataset is limited to western European building stocks with similar building standards and that broader transferability remains to be tested. However, the Conclusions (Page 17, Lines 315–322) also state that the validation demonstrates the tool's practical applicability beyond the training data regions, before acknowledging that the current dataset may limit its transferability to regions with different architectural characteristics. I would therefore appreciate a brief clarification of the intended domain of applicability of FlooDEsT. In particular, under which conditions do the authors consider that the model can currently be transferred with confidence, and when would recalibration or additional validation be advisable? A short discussion reconciling these statements would help readers better understand the current capabilities and limitations of the proposed methodology.

R: We agree that the statements are somewhat conflicting. We therefore clarify in which range of building standards / types the model was tested and is therefore transferable. Model transferability remains however limited beyond that. We will add to the conclusion what follows.

“FlooDEsT was demonstrated to work well for mostly masonry buildings (83%), winterized buildings (e.g., possessing a heating system), with built rooftop (90%; 62% with fitted attic), with basement (96%), constructed post-war (85%; 65% since 1965), and mostly for one-family homes (68%)”.

On behalf of all co-authors,

Guilherme Samproгна Mohor