

We thank the referee, Ms. Villafani, for carefully reading our manuscript and for her constructive comments and 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.

RC1:

This work aiming presents FloodDEsT, an XGBoost-based tool for rapid damage estimation from pluvial floods. The training data is based on household survey data from Germany and validated on independent datasets from Amsterdam and Karlsruhe. It addresses an important innovation for under-modelled hazard pathway (pluvial floods). Yet, major revision and clarification is needed. Please see my comments and suggestions below.

Introduction

1. The motivation that pluvial floods are under-represented in damage modelling is clearly stated. However, why impacts are not systematically documented (line 21) required additional information that further supports this as the gap.

R: For a more direct and clear explanation to this gap we will extend the sentence addressing the gap more explicitly to pluvial floods. The new paragraph will read:

“Because their impacts are not thoroughly or systematically documented (Molinari et al., 2017), numerical modelling is required to improve our understanding of the phenomenon, to fill in data gaps, and to estimate impacts under defined scenarios for a thorough risk analysis (Gerl et al., 2016; Meyer et al., 2013; Molinari et al., 2020). *As pluvial floods occur away from water bodies, thus away from typical water level gauges, data on pluvial flood events is even more scarce than for riverine floods (Porter et al., 2023). At the same time, the rapid onset of pluvial floods makes it being caught by remote sensing imagery more challenging, or rather impossible. This scarcity of data in turn makes validation of loss models also very challenging. Additionally, there is a general perception that pluvial floods are less intense and less damaging (Thieken et al., 2022), leading to lesser focused studies and data collections.*”

References:

Thieken, A.H., G. Samproгна Mohor, H. Kreibich, M. Müller (2022): Compound inland flood events: different pathways, different impacts and different coping options. – Nat. Hazards Earth Syst. Sci. 22, 165–185, <https://doi.org/10.5194/nhess-22-165-2022>

2. Consider reviewing the damage models and present them in a more structured manner. For example, distinguishing them between type of model, and summarising key limitations and constraints before proposing multivariable models. Additionally, it is not clear how multivariable modelling improve damage estimations. In Line 50 you introduce this as a solution to the deficiencies in other models. Provide some references to this.

R: A structured review of damage models is not the focus of the manuscript and is therefore kept short. The main message of the referred paragraph is that most models do not even model and estimate building-level monetary (or relative) damage, but rather count of insurance claims. The advantage of multivariable models is addressed elsewhere, also conceptually, but is especially referred in the paragraph to Wagenaar et al., 2017, which has shown the improvement with multivariable models very nicely. In light of comments by other reviewers, we will shorten and rephrase the paragraph to make our points clearer.

“Consequently, only a few damage models have been developed for SWF (Gradeci et al., 2019), many of which **do not meet our goal of** modelling ~~the occurrence of insurance claims and not the damage~~

magnitude at the object-level, ~~while others do model object-level damage but employ variables that are unfeasible to query, such as building material or socio-economic variables like education level (Gradeci et al., 2019). Existing flood damage models (for SWF) vary in both their predictor variable sets as well as the form of the damage output.~~ For example, most approaches are at an aggregated level (e.g., county level [Spekkers et al., 2014]), instead of building level. Gradeci et al. (2019) notes that many models try to estimate the count of insurance claims (e.g., Sørensen and Mobini, 2017; Spekkers et al., 2013, 2015), which is in turn not a measure of damage itself. Most such approaches use rainfall as predictor (e.g. Cortès et al., 2018; Spekkers et al., 2013); although being the driver of the event, **rainfall alone** is not enough to explain damage magnitude, **even though such studies showed its usefulness in estimating** the occurrence of claims. Lastly, some models that do model building level damage require data that is not or hardly available (e.g. Rözer et al., 2019; Van Ootegem et al., 2018) in real-life cases or are too simplistic adopting a stage-damage function (e.g. Porter et al., 2023); **whilst Wagenaar et al. (2017) have shown that multivariable modelling improves damage estimation.** “

3. While the four improvements are clear, consider clarifying how FlooDEsT differs from earlier work by Bronstert et al., Lindenlaub et al., and Dobkowitz et al. beyond the extended dataset and improved computational performance. Specify the novelty of the work. Consider explicitly stating this in the text as well, as it would help to articulate the main scientific contribution.

R: An earlier version of FlooDEsT was used in the three cited papers. However, the model itself was NOT fully presented in those earlier works, but only shortly addressed with minimal information. The current manuscript is the first time (besides a poster presentation at an EGU general assembly) where the whole modelling development chain, from data processing, hyperparameter tuning, calibration to validation is presented. Because there are a few differences in the newest version of FlooDEsT compared to the one used in the earlier works, we felt compelled to give this notice.

Data and methods

4. Have you considered using other tree-based methods, such as random forest or lightGBM? How is XGBoost particularly suitable for your method?

R: Yes, we have considered other tree-based methods. Random Forest usually led to “middle-ground” estimates, not performing well at the extreme (high or low) damage estimates. Additionally, Random Forest often performs well with a high number of trees, while the XGBoost achieve great performance with only a few trees, which counts towards the rapid, near real time, assessment that was aimed at. We are less familiar with lightGBM, but some discourse around it recommends it for “large” datasets, which is not our case. Therefore, we adopted XGBoost for its being relatively well known, documented, and suitable for our requirements. We will add this observation to the model choice as follows.

“Other regression tree methods were considered during the model development. For example, the popular ‘Random Forest’ often achieves good results, but with a larger number of trees and has produced overly centred estimations with our data (results not shown); other tree boosting methods have been praised, however, for advantages in applications involving much larger data sets than ours. We hence opted for the XGBoost for its suitability, its ability to implement monotonic effects, and the predictive performance it has achieved (as shown in Section 3) with a small number of boosting trees, which contributes to a fast execution.”

5. People tend to forget and/or underestimate the level of impacts for diverse reasons and adding a temporal aspect could increase this effect. You have validation points from 10 years after the first points, and 14 years

after the training set. Are you accounting for this or handling this in any way? If so, how? What happens when households do not repair/replace in that span of time?

R: There is a misunderstanding of the data collection procedure. Each event had its own data collection campaign, and as explicitly stated in line 96, the surveys took place at latest 21 months after each respective event. All monetary values are referred to one reference year to account for inflation. This later point will be added to the manuscript.

6. The training and test sets split is described, but details on the splitting procedure are missing. Please specify whether the split was random, stratified, or followed another method.

R: Please accept our apologies that we have not stated in detail the splitting method. We will add the following explanation:

“The data were sorted by relative damage and divided into [training, test, training] triplets, ensuring that both training and test sets encompassed the complete range of relative damage values.”

7. In your methods you are working with data exclusively from Germany, and have external data points from Amsterdam for the validation. This could result in more localised/regional differences inherent to the affected households after pluvial flooding events. Have you considered this as a limitation for the work? This already led to removing entries (see specific comments below).

R: We resorted to data from the Netherlands due to both the overall scarcity of data and to test the transferability of the model, even if by limited means. We will reflect this point in the discussion. We can also add that Spekkers et al. 2017 have compared the Amsterdam and the Münster (2014) datasets, and noted that “Although there are differences in building types between cities, we believe that differences in resistance are minor or slightly in favour of Münster, given the high uptake of emergency and precautionary measures”. This reinforces that the validation of the model in setting beyond the ones of the training set.

“We resorted to data from the Netherlands due to both the overall scarcity of damage data of pluvial floods and to test the transferability of the model, even if by limited means. The validation set still only comprises cases from western Europe and represent similar building standards. A broader transferability of the model to different building types is yet to be tested.”

8. The predictor set points need further clarification:

- There are many other contamination types in the water originating from sewage and industrial waste. These are not included in the predictor set. Why? Could these be indicators of higher damage levels?

R: Previous analysis of the German dataset (Thieken et al., 2005) showed that heating oil is among the most severe contaminations. As explained in line 110, we have tested other predictor variables as well. They are not presented in the main manuscript for brevity, but are included in the Supplementary Material. We have tested a different contamination score which considers other contaminants, but the single contamination by oil was deemed more relevant.

Reference:

Thieken, A. H., Müller, M., Kreibich, H. & Merz, B. (2005). Flood damage and influencing factors: New insights from the August 2002 flood in Germany. *Water Resources Research*, 41 (12), 314. <https://doi.org/10.1029/2005WR004177>

- Floor space is used as a continuous variable from 50—12,000 m². The distribution is likely highly skewed. Was any transformation considered or tested? If so, why was the raw scale retained?

R: Due to the structure of tree-based methods, a monotonic transformation, such as log-scale, is expected to change the splits only minimally. Preliminary tests did not indicate improvements in the model performance, therefore we retained the raw scale in the final model.

- The PLPM score is a composite. A short note on its construction and components could be informative.

R: That is advisable. We will extend the explanation in Table 2 to:

“Score of implemented precautionary measures at the property, from 0 up to 48, *based on a set of measures weighted by pre-determined qualitative damage-reducing effectiveness* (Laudan et al., 2020; Thieken et al., 2008) [-]”

- The variable "cellar" has very low variability. Please add in your discussion whether the variable is non-informative and whether it should be retained in future versions of the model.

R: As explained in line 110, the fine tuning of the model considered inclusion or exclusion of predictor variables. The fine tuning resulted in the inclusion of “cellar” among the predictor variables. Even though it has the lowest feature importance (as shown in Fig. 6), based on previous literature (e.g. Kreibich et al., 2005; Schwarz & Maiwald, 2007), the presence of a cellar is a relevant information, supporting its inclusion. This will be reinforced in the discussions as well.

“The absence of a cellar shows lower relative losses, but this result must be seen with some reservation, given the low percentage (3%) of buildings without a cellar in the dataset. *Even though it has the lowest feature importance, based on previous literature (e.g. Kreibich et al., 2005; Schwarz & Maiwald, 2007), the presence of a cellar is a relevant information, supporting its inclusion.*”

References:

Kreibich, H., Thieken, A. H., Petrow, T., Müller, M. & Merz, B. (2005). Flood loss reduction of private households due to building precautionary measures – lessons learned from the Elbe flood in August 2002. *Natural Hazards and Earth System Sciences*, 5 (1), 117–126. <https://doi.org/10.5194/nhess-5-117-2005>

Schwarz, J., & Maiwald, H. (2007). Prognose der Bauwerksschädigung unter Hochwassereinwirkung. *Bautechnik*, 84(7), 450–464. <https://doi.org/10.1002/bate.200710039>

9. The filtering of Amsterdam data by removing entries with water entry via windows, roofs or downpipes should be described more explicitly by clarifying how many points were removed. Is this introducing any selection bias?

R: No, this procedure does not introduce a bias. In the survey conducted in Germany, a filter question explicitly requests for the flood pathway of interest, which implies the dismissal of such cases where water entered the house via windows, roofs or downpipes. The survey in Amsterdam included other pathways as explicit information, allowing for an easy filtering, exclusion and harmonization of the dataset to the type of flooding we are in fact studying.

10. The estimation of building value with linear mixed-effect models is a critical step to obtain absolute losses. Its performance is not reported. Please provide performance metrics for the building value model. Discuss how uncertainty propagates to relative loss and damage class assignment.

R: Regarding the building value estimation itself, the modelled values showed a very high R^2 of 0.986. Using this modelled building value, the re-estimated “observed” relative losses also showed a very high R^2 of 0.995. This information will be added to the respective Section 2.2.

11. In Line 139 you mention a case was dismissed because the building value was "much lower than reported losses". Could you explicitly provide information as to which case that was? Please specify the used criterion. Are there other points close to the threshold? Was this done before or after fitting the model? Additional explanation is needed for the Amsterdam case. Building values there would be based and modelled on the German insurance market standards.

R: Data from Amsterdam did not include building value or relative loss. The building value were estimated using the described methodology, but could not be checked. The values were adjusted to the year and by purchasing power parity from Germany to Netherlands. Only one case across calibration and validation datasets was dismissed due to estimated building value being lower than reported loss, namely 20% lower. Because the difference was so large, we preferred dismissing the case instead of simply rounding it to a 100% damage. In all other cases, the estimated building value was higher than the reported damage and therefore not problematic.

12. To improve clarity and robustness of this section, some gaps need to be filled.

- Describe more explicitly how distributions were fitted for each variable type. You can also provide this information in your supplementary material, alongside graphs for the fits.

R: 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 could also be explored per missing variable, given that the predictor variables have shown different importance (see Fig. 6) to the model. “

- Clarify whether reported times in computational performance include data or only the computation.

R: As stated in lines 284-286, the time is considered “without counting the pre-processing of the basis data, which is fixed for all runs afterwards”.

- Clarify whether the current performance already meets or approaches the target of what is sought in real-time. Is there a threshold of forecast availability time that could be used as a point of comparison?

R: The first information in the modelling chain are the radar data, which have a time-resolution of 5 minutes. Nevertheless, an update of expected damage information every 5 minutes would anyway be excessive. A cycle of 15 minutes was envisioned within the project for the damage estimation. Rainfall and runoff estimations showed in a pilot a runtime shorter than 2 minutes (not published).

13. In your sensitivity analysis, have you considered testing scenarios with multiple missing variables simultaneously? This could provide a more realistic setting in operationalising realistic data gaps. Report not only RMSE but changes in damage class under these scenarios.

R: The simultaneous missing of water depth and flood intensity was tested, as shown in Figure 8. Assuming the absence of many or all predictor variables, testing the “worst performance” of the model, seems counter-productive and not informative to us. Reporting changes in damage class for each missing variable across multiple buildings is also cumbersome because class agreement index with skewed group sizes, despite existing methodologies, is not straightforward in our view; RMSE was chosen for the elegance of an aggregated value, allowing for an easy visualization of the model performance and easy comparison of each situation (of a missing variable).

Results and Discussion

14. The independent validation datasets are valuable, but with a small total sample size. The manuscript acknowledges this limitation, but the implications for the robustness are not fully explored. Discuss how sensitive RMSE and the damage classes are to individual points, given the small sample size. Confidence intervals or other uncertainty estimates could be a valuable addition.

R: The paragraph between lines 254-259 provides the robustness assessment of the methodology. Replicating the same for the validation dataset, given that it is of small size and does not represent the whole sample space, would be less informative in our view. Therefore, we would prefer to not adding this.

15. The model fails to predict "very high" damage cases in the validation set. This is a critical limitation for a tool intended for flood damage estimation, where extreme losses are key.

- How do you define the damage classes? Are they stable? Please add to the discussions.

R: The classes were defined based on Jenks natural breaks of the relative losses of the training data with 5 classes (0-3%, 3-9%, 9-16%, 16-37%, above 37%); the resulting two highest classes however created very small samples, thus we kept four classes only, with "above 16%" as the highest damage class. Losses from pluvial floods are in general lower than losses from riverine floods, therefore it is relevant to have a good estimate of the many low cases. At the same time, the class defined as “very high” for pluvial floods would be much more frequent in other more aggressive types of floods such as riverine or flash floods.

- Discuss possible reasons, including possible associations with building types, contamination cases, or other factors not captured by the current predictors.

R: The seven “Very High” cases 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.

- Consider exploring or proposing strategies to improve performance for extreme losses.

R: Overall, adding more variables could improve such estimates, as losses are caused by different factors (e.g., building material or building age). This is however always attached to having enough data with enough variability for the model to learn something from it.

16. You provide references mostly for the results in the calibration performance. Please provide additional references in the following subsections of your results. These provide a more in-depth understanding and discussion of your approach, strengths and weaknesses.

R: Several questions posed by the Reviewers have resulted in additions to the Results and Discussion section, including additional references. We appreciate all the constructive comments from all reviewers.

17. Are there other ways to optimise the procedure? How can reproducibility be ensured? Please consider including a short paragraph on the main text.

R: Local reproduction is ensured in which the scripts were set to a fixed seed-number (fixed randomness). Most of the data are accessible via the database HOWAS21 (see <https://howas21.gfz.de/>), but given the privacy issues with the building-level (household-level) data however, a reproduction by third parties is limited.

18. Figure 10 is a very useful way of showing visually the estimated damage classes. Consider adding a brief interpretation of patterns if known, to connect the damage outputs to the drivers. It could be connected to what point 15 mentions.

R: We appreciate the acknowledgement of Figure 10. The results are however from a scenario and cannot be truly validated.

Conclusions

19. Please consider adding recommendations for practitioners, including testing under different infrastructure, etc.

R: Thank you for this remark. One main factor addressed is data scarcity, especially at the building level, which is reflected in the small validation sample size. We included in the conclusion that more

data must be collected. Given the absence of a legal requirement in Germany for such data collection, and knowing that most countries neither have a systematic damage data collection, this works still falls often in the hands of researchers.

General remarks:

Consider reviewing the language and other minor details. Examples include:

- Line 29: Please revise the sentence. You are describing several models, not one.

R: We are unsure what is meant here. In the sentence, “the majority” [of models] agrees with “was”.

- Line 36: There is one additional quotation mark.

R: Thank you for noting that. Will be corrected.

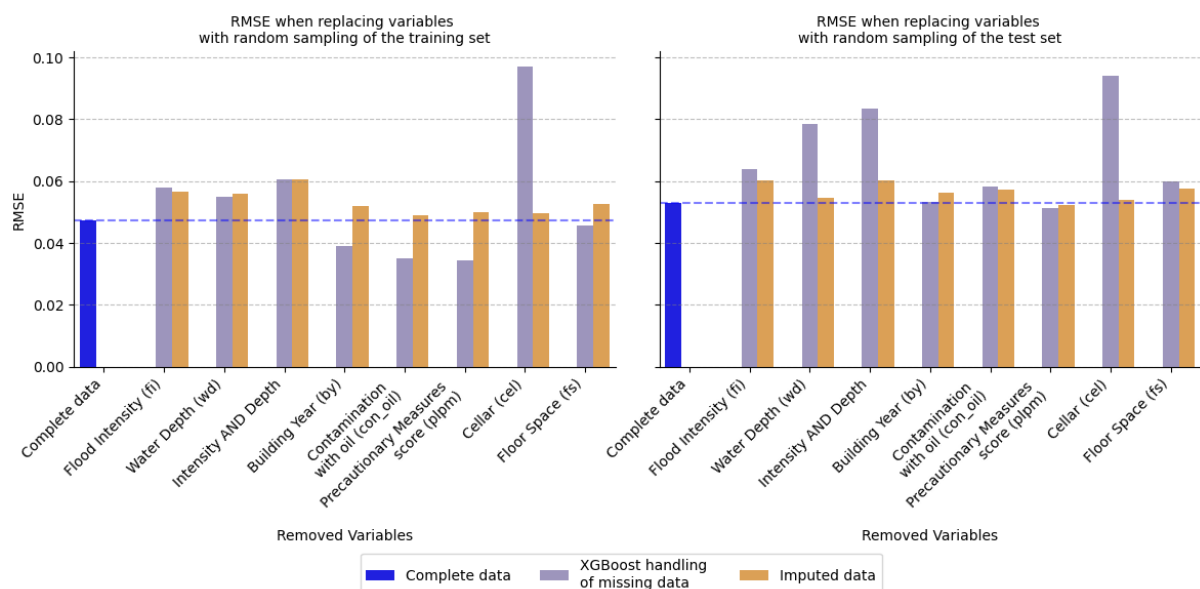
- Line 47: Please revise the sentence.
 - You are contrasting two ideas without a clear link. Make it explicit.
 - You are referring to rainfall as a driver. The object is missing.

R: Thank you for noting that. The sentence reads now:

“[...] although being the driver of the event, rainfall alone is not enough to explain damage magnitude, even though such studies showed its usefulness in estimating the occurrence of claims.”

- Figure 8. Please move the legend below the names of the two plots.

R: Sorry for that. As pointed by another reviewer, the legend label is not ideal. “Missing data” will be replaced with “XGBoost handling of missing data”.



On behalf of all co-authors,

Guilherme Samprogna Mohor