https://doi.org/10.5194/egusphere-2026-3462
Preprint. Discussion started: 1 July 2026

Rapid flood damage estimation tool for urban pluvial floods with scarce data

Guilherme Samprogna Mohor[1], Sarah Lindenlaub[1], Annegret H. Thieken[1]

[1]Institute of Environmental Science and Geography, University of Potsdam, Potsdam, 14476, Germany

Correspondence to: Guilherme Samprogna Mohor (guilherme.samprogna.mohor@uni-potsdam.de)

Abstract. Reliable and rapid flood damage estimation is crucial for both disaster risk reduction and crisis management. Yet, existing models primarily focus on riverine floods, neglecting urban pluvial floods – a substantial gap, as heavy rainfall can lead to flooding virtually anywhere. Here, we present FlooDEsT, a new machine learning (ML)-based tool to rapidly estimate building-level damage from urban pluvial flooding with four key improvements, compared to traditional models. First, the model was trained specifically on damage data from urban pluvial flood events. Second, the tool utilises XGBoost, a ML technique capable of capturing complex non-linear data relationships. Third, the tool efficiently utilises geographical information only as necessary, reducing pre-processing time. Fourth, to address the common challenge of missing data, the tool uses smart random sampling techniques to impute building-level features that are representative to buildings affected by this flood pathway, reducing exposure bias. 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. FlooDEsT tackles critical gaps in damage modelling, offering valuable support for disaster preparedness and response.

1 Introduction

Floods are one of the most numerous and damaging disasters in the world (Centre for Research on the Epidemiology of Disasters (CRED), 2025). In fact, 85% of economic losses in Europe in 2024 stemmed from floods (Copernicus Climate Change Service (C3S) and World Meteorological Organization (WMO), 2025). 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).

Floods are not only caused by different drivers, such as intense, short convective or long advective rainfall (Hundecha et al., 2020; Merz and Blöschl, 2003), but develop on the ground through different pathways, such as breaching levees, overflowing river banks, or elevated groundwater levels (Mohor et al., 2020, 2021; Thieken et al., 2022). Riverine floods are the most documented flood pathway and are likewise the pathway for which most damage models have been developed. In their review, Gerl et al. (2016) listed 47 damage models, the majority of which was developed for fluvial floods, a few for coastal floods,

https://doi.org/10.5194/egusphere-2026-3462
Preprint. Discussion started: 1 July 2026

levee breaches, groundwater floods or lake floods, but none of the reviewed models addressed pluvial floods. In contrast to riverine floods, pluvial floods, also called surface water floods (SWF), can occur virtually anywhere, also afar from a water body through overland flow and ponding (see Bernet et al. (2017) for terminology). These, however, have been less documented and studied, even though they can cause high damage (807 m EUR insured in 2011 in Copenhagen; 1.86 bn USD in 2012 in Beijing; 70 m EUR to insured private households in 2014 in Münster; Spekkers et al. (2017)). The non-systematic assessment of pluvial floods probably leads to the underestimation of the overall flood risk as well as the population's perception of flood risk, as observed in the USA, where ""FEMA's Flood Insurance Rate Maps (FIRMs), which are the most widely used flood risk assessment and communication tools in the US, exclude pluvial flood risk" (Porter et al., 2023). The European Union Floods Directive (2007/60/EC) was integrated into the German Federal Water Act (WHG), requiring the preparation of flood hazard and risk maps, however only for riverine and coastal floods (Riese et al., 2019).

Consequently, only a few damage models have been developed for SWF (Gradeci et al., 2019), many of which model 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 (Sörensen 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, is not enough to explain damage magnitude, but rather 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 multivariable modelling has been shown to improve damage estimations (Wagenaar et al., 2017).

SWFs can occur virtually anywhere, are usually quick, caused by intense rainfall and have a short duration. Such characteristics make its damage modelling crucial for both scenario-based projections –complementing the flood risk assessment based solely on typical, riverine floods and supporting prevention planning– and for real-time forecast –as impact-based forecasts become more and more desired to provide locally accurate information in support of emergency operation and early warning (Merz et al., 2020; Rözer et al., 2021).

Therefore, in this article we present the newly developed Flood Damage Estimation Tool (FlooDEsT). The development focused on tackling two challenges: being specific for SWF and being operational, that is, fast and applicable in a real-life situation. In this work we show that FlooDEsT performs well in predicting observed SWF damage and that the whole application, between acquiring the hazard data (already processed and readily available) and delivering the damage estimates can be run in a real-time forecasting system.

This is an improved version over the model used by Bronstert et al. (2025), Lindenlaub et al. (2025) and Dobkowitz et al. (2025), with an extended damage data set, renewed optimization and improved computational performance.

https://doi.org/10.5194/egusphere-2026-3462
Preprint. Discussion started: 1 July 2026

2 Data and Methods

In its entirety, FlooDEsT – the Flood Damage Estimation Tool– comprises a damage function and its application, including data pre- and post-processing (Figure 1). In what follows we describe the model, the training data set and hyperparameter optimisation (2.1), the validation of the model with data unseen by the model (2.2), the random sampling procedure for model application (2.3), and its computational performance (2.4).

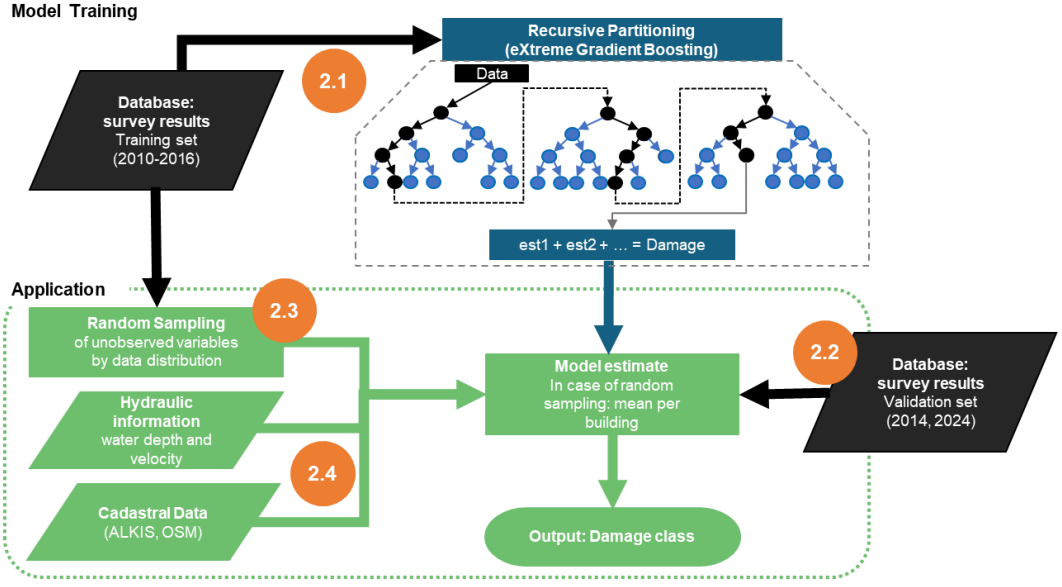


Figure 1. Flowchart of the FlooDEsT training and application. OSM: OpenStreetMap; ALKIS: Authoritative Real Estate Cadastre Information System. est: estimation

2.1 Training the damage model

Recursive partitioning is a flexible machine learning (ML) technique that iteratively splits variables at optimal points to maximize information gain or minimize impurity (Strobl et al., 2009). While various algorithms and approaches are based on this concept, they all essentially divide the n-dimensional space of the input data into regions with similar responses. As these algorithms make stepwise, dichotomous "decisions" on where to best split a given variable, they are commonly known as decision trees.

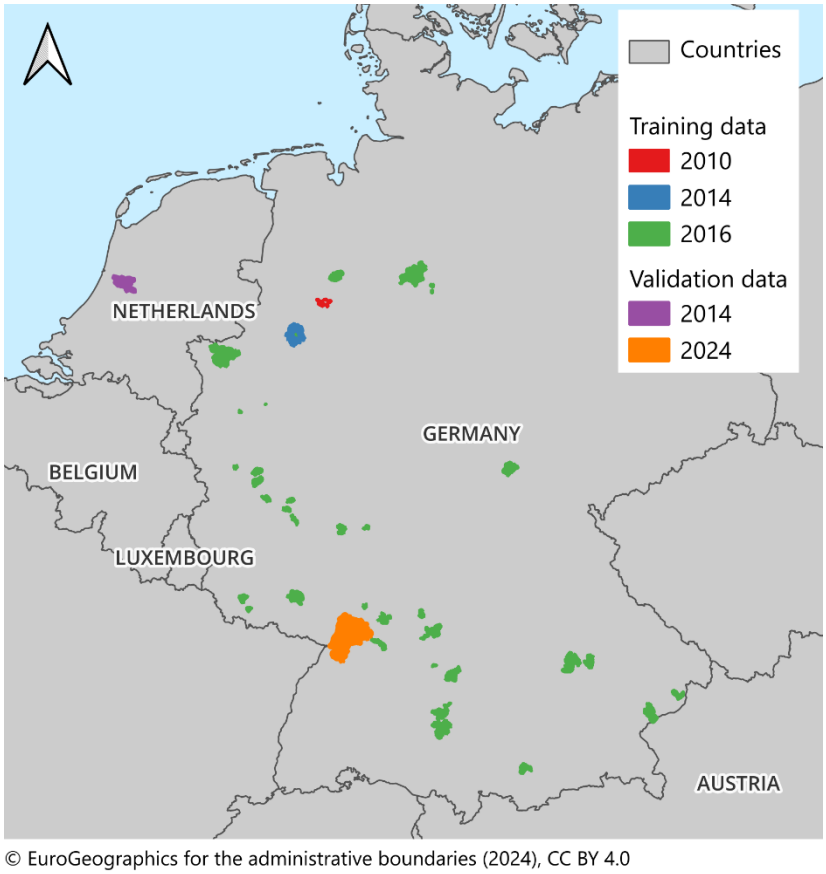
Decision tree algorithms have been further improved to enhance their robustness against e.g. data variability, skewness, and imbalances between continuous and categorical variables, among other challenges, for example via ensemble methods (James et al., 2021). Boosting algorithms build an ensemble of decision trees sequentially, where each subsequent tree aims to correct errors made by the previous ones (Chen and Guestrin, 2016; Friedman, 2001). This approach uses differential weighting of the data, giving more importance to misclassified samples to improve prediction where it is most needed. In this development,

https://doi.org/10.5194/egusphere-2026-3462
Preprint. Discussion started: 1 July 2026

we used XGBoost (eXtreme Gradient Boosting), a multi-purpose gradient boosted regression tree with regularization and column sub-sampling to reduce overfitting, a sparsity-aware split finding algorithm to deal with missing data, and further improvements for computer efficiency when dealing with large datasets (Chen and Guestrin, 2016). The Python implementation with native syntax was adopted (Python Version 3.11.9; XGBoost Version 3.0.5 [https://github.com/dmlc/xgboost]).

Different flood pathways lead to distinct damaging processes (e.g., Mohor et al., 2020; Thieken et al., 2022). Thus, we fed the model with data from previous events known to be of the specific flood pathway SWF. To train the model, we have used 551 data points collected via surveys of flood affected households in the aftermath of previous pluvial events in Germany (Figure 2:), namely: 2010 in Osnabruck (Rözer et al., 2016), 2014 in Münster and Greven (Spekkers et al., 2017), and 2016 in several municipalities across Germany (Thieken et al., 2022). Further survey datasets were used for validation: 44 data points from the events of 2014 in Amsterdam (Spekkers et al., 2017) and 2024 in and around the city of Karlsruhe, collected for this model development.

The surveys took place at least five months after the events (up to 21), providing enough time for the households to repair or replace damaged items and thus have a more accurate estimate of the damage (see Thieken et al., 2017).



https://doi.org/10.5194/egusphere-2026-3462
Preprint. Discussion started: 1 July 2026

Figure 2: Municipalities of the respondents from surveys used in this work

Because there is a dependency between the presence of predictor variables, their form (e.g. log-scaled), and the model's hyperparameters (Table 1), one single optimisation procedure was performed allowing for all variations, that is the inclusion and dismissal of each predictor variable as well as the definition of hyperparameters drawn from a range of initial values (see Supplementary Material S2). Four rounds of grid search were performed to shorten the range of potential hyperparameters. The last optimisation, varying both predictors and hyperparameters, was solved in Python with the package Optuna Version 4.5.0 (Akiba et al., 2019) using a random sampler through 10,000 trials, with early stopping rounds set to 15 (i.e. boosting will stop if for 15 rounds there is no improvement; the best boosting round, and not the last one, is then selected).

Based on other existing models (e.g. Mohor et al., 2021; Rözer et al., 2019; Vogel et al., 2018) as well as on insurer's practice, we selected a set of potential predictor variables that are relevant for damage estimation and, mostly, available with reasonable effort. An initial set was tested, including variations of the variables, for example with or without log-scale. This full set of tested variables is presented in Supplementary Material (Table S1), whilst the reduced listed of adopted variables is presented in Table 2. To characterize the flood hazard, we use water depth, flood intensity (an equivalent composite of water depth and velocity, see below), and the presence of contamination in the flood waters. To characterize the affected building, we use the building's size, its year of construction (period), and a score representing implemented precautionary measures at the property level (PLPM).

Flood intensity (FI) is defined in different ways, most commonly based on hydraulic information. The methodology adopted by (Defra / Environment Agency, 2006) splits the hazard rating [water depth*(velocity + 0.5)] into four levels representing probability of losing stability. Because velocity cannot be measured without specialized equipment, in the surveys flood intensity is only indirectly inquired as whether 'an adult could easily stand in the flooding waters'. The hazard rating levels were then mapped to the equivalent classes collected in the survey, namely: "Could easily stand", "Would struggle to stand", "Would have been carried away" or "Water was too deep to stand" (see Table 2). The latter classes are descriptive as to whether an adult could still stand in the flood stream (Nicklisch (2004) based on U.S. Bureau of Reclamation (1988)).

Table 1. Fitted hyperparameters

Parameter	Default	Range	Tuned	Description
max_depth	6	3 – 32	6	Maximum depth of each tree. More complex trees are more likely to overfit.
num_boost_round	-	1 – 150	7	Number of boosting trees.
'monotone_constraints'	(not used)	None, -1, +1	Contamination: + PLPM: -	+/- Force higher/lower response values at higher input values. PLPM shall have a damage reducing effect.

https://doi.org/10.5194/egusphere-2026-3462
Preprint. Discussion started: 1 July 2026

				Contamination shall have a damage increasing effect
eta (learning_rate)	0.3	0.001 – 0.7	0.6564443183914798	Step size shrinkage used in update, to make boosting more conservative.
tree_method		exact, hist, approx	exact	Simplified algorithm can improve performance.
grow_policy		depthwise, lossguide	depthwise	Defines whether trees will grow up to depth (depthwise) or based on highest loss change (lossguide). 'lossguide' is only used if tree_method is not 'exact'.
max_leaves		10 – 256	-	Maximum number of nodes to be added. Not used if tree is 'exact'.
scale_pos_weight	(not used)	0.5 – 10	9.935662274906461	Control the balance of positive and negative weights, useful for unbalanced classes.

2.2 Model Validation

Using independent datasets, namely survey data from the SWFs of 2014 in Amsterdam-NL (n=113; max_rloss=0.173; Spekkers et al., 2017) and of 2024 in Karlsruhe (n=24; max_rloss=0.300; presented in this paper), a total of 44 valid data points (with water depth values, as this is the most important predictor variable) were used to validate the model. Note that the Amsterdam dataset is larger than that, but in this study damage entries that were flagged to be caused by water entrance through the windows, the roof, or the downpipes were dismissed.

All surveys employed about the same questionnaire, with minor changes in formulation, so that they are directly comparable. Absolute losses were reported in all cases; the calculation of relative losses depends on the value of the building. This value is usually calculated by the standard method of the German insurance market. The validation dataset however missed this information. The building value was therefore estimated through a linear mixed-effects model (R "lme4"; Bates et al., 2015) using the floor space, number of floors, type of roof, presence of a basement, and the building material as predictors, fitted to the available values calculated by the German insurance market standards from an expanded dataset (regardless of them being used in the damage modelling or not). With this procedure, one of the cases resulted in a building value much lower than the reported losses; this data point was therefore dismissed.

Table 2. Summary of adopted predictor variables

https://doi.org/10.5194/egusphere-2026-3462
Preprint. Discussion started: 1 July 2026

Variable	Description			Range		
				Training set (n=367)	Test set (n=184)	Validation set (n=44)
Water Depth (wd)	Maximum water depth outside the house [cm]			0 – 300	0 – 200	0 – 150
Flood Intensity (fi)	Hazard Rating (HR) = water depth*(velocity + 0.5)			0 – 2	0 – 2	0 – 2
	DEFRA classes	HR split	Surveyed classes (adopted)			
	Low	0 < HR < 0.75	0: "Could easily stand"			
	Medium	0.75 < HR < 1.25	1: "Would struggle to stand"			
	High	1.25 < HR < 2.5				
	Extreme	2.5 < HR	2: "Would have been carried away" OR "Water was too deep to stand"			
Contamination with heating oil (con_oil)	Indicator of contamination in the floodwater: 0=no contamination to 1=heavy contamination with heating oil [-]			y/n	y/n	y/n
Floor Space (fs)	Floor space of the whole building [m²]			50 – 12,000 m²	55 – 4900 m²	30 – 1100 m²
Cellar (cel)	Presence of a cellar [y/n]			y/n	y/n	y/n
Property-Level Precautionary Measures (plpm)	Score of implemented precautionary measures at the property, from 0 up to 48 (Laudan et al., 2020; Thieken et al., 2008) [-]			0 – 48	0 – 47	0 – 36

https://doi.org/10.5194/egusphere-2026-3462
Preprint. Discussion started: 1 July 2026

Building (by)	**Year**	Year of construction in classes: [-] 1: until 1924; 2: 1925-1948; 3: 1949-1964; 4: 1964-1990; 5: after 1990	1 – 5	1 – 5	1 – 5
Damage (dmgcl)	**class**	Reclassified building loss ratio (rloss): 1 (Low): 0 < rloss ≤ 3%; 2 (Medium): 3% < rloss ≤ 9%; 3 (High): 9% < rloss ≤ 16%; 4 (Very high): 16% < rloss	1 – 4	1 – 4	1 – 4

2.3 Application and Random Sampling – filling in the gaps

The model uses two sets of data: hazard, represented by the hydraulic information, and exposure, represented by the buildings' characteristics. While model training uses solely survey data, its application assumes that the hazard data comes from a hydraulic simulation of a scenario or a forecast is provided (Figure 1). In the application cases, only the maximal values of water depth and velocity in the simulation are used. Building characteristics are drawn from cadastral data, either authoritative data –such as the German ALKIS (Authoritative Real Estate Cadastre Information System)– or from any open data set, such as OpenStreetMap (OSM). Missing data, especially regarding the exposure, are imputed through random sampling based on the survey data.

In real-life applications, available data seldom matches the level of detail of the used survey data (Kellermann et al., 2020; Thieken et al., 2017). Although the variable selection procedure considered expert opinion on data availability, missing data is still expected. XGBoost is natively able to make a prediction with missing data by learning, during the training phase, after each partitioning which branch generally provides better predictive performance (Chen and Guestrin, 2016). However, the introduced error is relatively large. 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). Each predictor variable is fit to a statistical distribution –Beta, Gamma, or Multinomial if the variable is categorical. In the prediction phase, if one or more variables are missing, 100 samples are drawn from the respective distributions, generating 100 predictions. The outcome in turn is the average of the 100 predictions. A sensitivity analysis is run to assess whether the introduced errors are acceptable. In each assessment, one variable at a time is removed from the dataset, and the damage is estimated either using the (1) XGBoost handling of missing data or (2) our imputation through random sampling. The predictive performance is indicated by the root mean squared error (RMSE), a commonly used metric which gives higher weights to higher values and presents the error at the actual response variable unit.

https://doi.org/10.5194/egusphere-2026-3462
Preprint. Discussion started: 1 July 2026

2.4 Computational performance

Indicators of computational performance will only consider the time for data extraction (of hazard information for each building), random sampling (where necessary), and damage estimation. It is assumed that hydraulic forecasts or scenarios are readily available.

Besides the computational speed of the random sampling and the model estimation tasks, extracting the outputs of the (hydraulic model's) forecast to water depth and velocity (to calculate flood intensity) may consume valuable time in a real-time application. That is the case in the typical procedure using geographical information tools (e.g. zonal statistics) to search the maximal values per building unit. The function "zonal_stats" from the package "rasterstats" version 0.20 (Perry, 2024) is a popular Python solution for the Zonal Statistics of standard GIS software. To overcome this, considering that the application (e.g. a forecast) will always have a pre-defined domain, and that the building data basis does not need frequent updates, the geographical information may be removed; thus, the following step will only process a raw array data (a simple track max values per index, hereafter named "raster screening"; Figure 3).

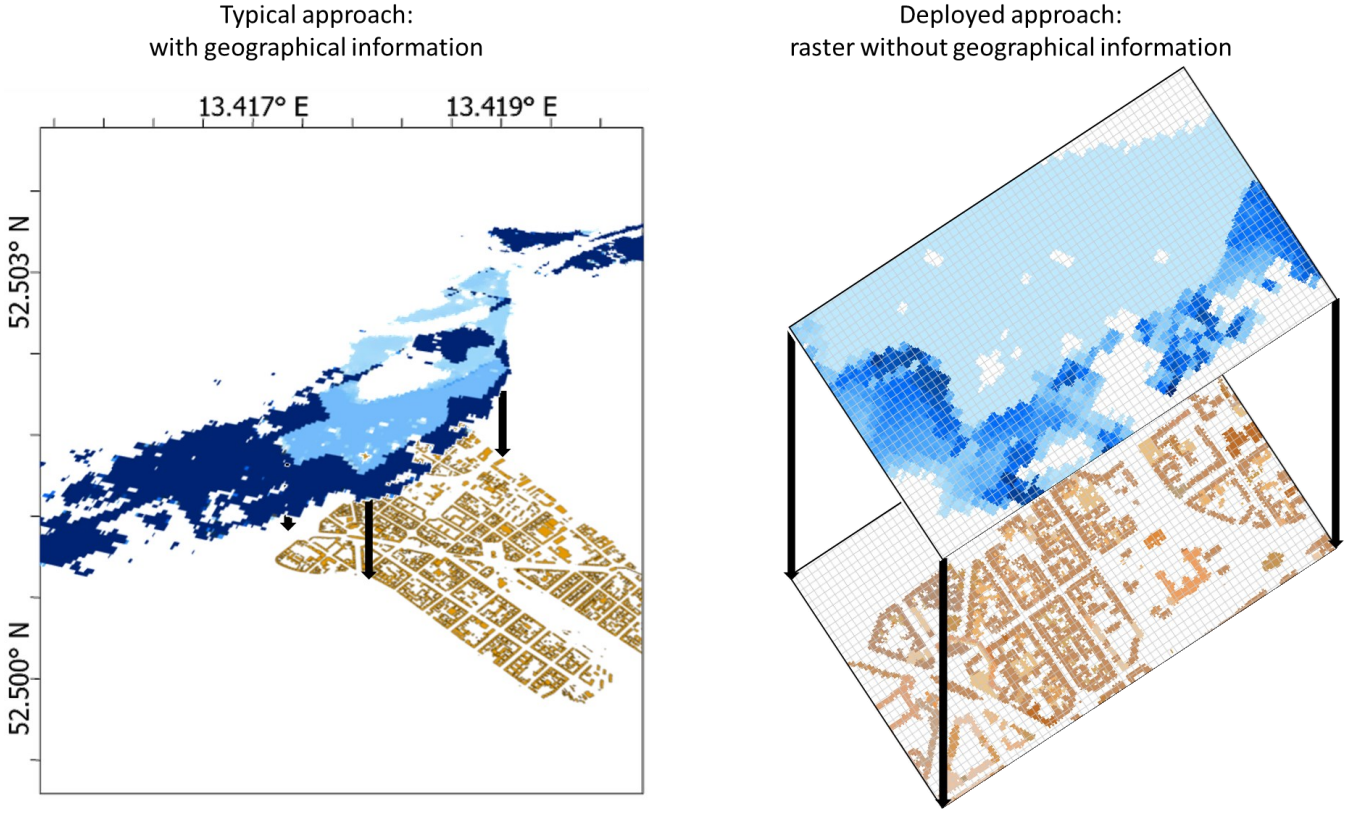


Figure 3. Illustration of the raster screening approach. The typical approach (left) uses geographical coordinate systems to cross overlapping information, whilst the approach deployed in FlooDEsT treats the raster layers, clipped to the relevant extent with the same cell resolution, as raw arrays, obliviating the need to cross the geographical coordinates.

https://doi.org/10.5194/egusphere-2026-3462
Preprint. Discussion started: 1 July 2026

3 Results and Discussion

3.1 Calibration performance

Of the 551 available data points, two thirds were used for model training, the remaining third for testing during the hyperparameter tuning (see Table 2).

Finally, the model was calibrated with the hyperparameters shown in Table 1. The number of boosting rounds is perhaps the most important parameter in a gradient boosting, as it can potentially lead to overfitting. Figure 4 shows the performance in RMSE in predicting the training and test datasets with the tuned hyperparameters, but varying boosting rounds beyond the optimised value of 7. It is noticeable that the gain with the training dataset reduces after the 5th round, whilst estimation of the test dataset loses performance after the 7th round, denoting overfitting. Therefore, the optimised 7 rounds of boosting were kept.

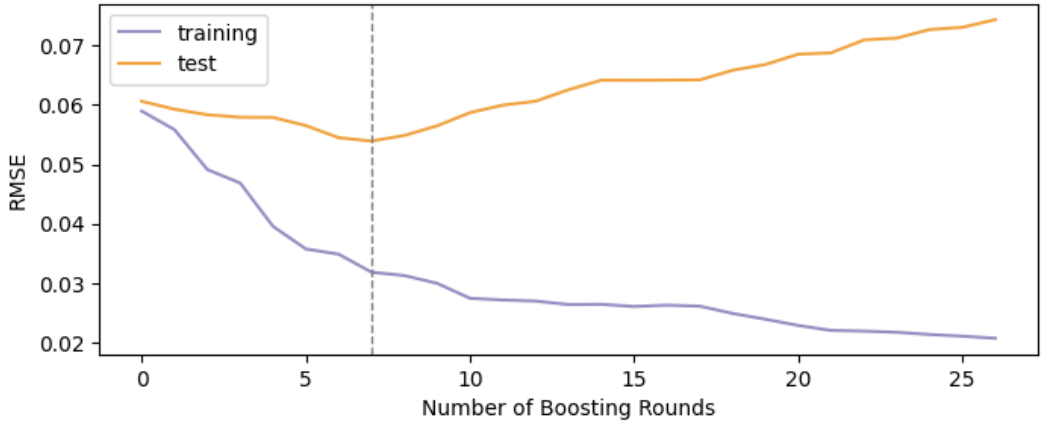


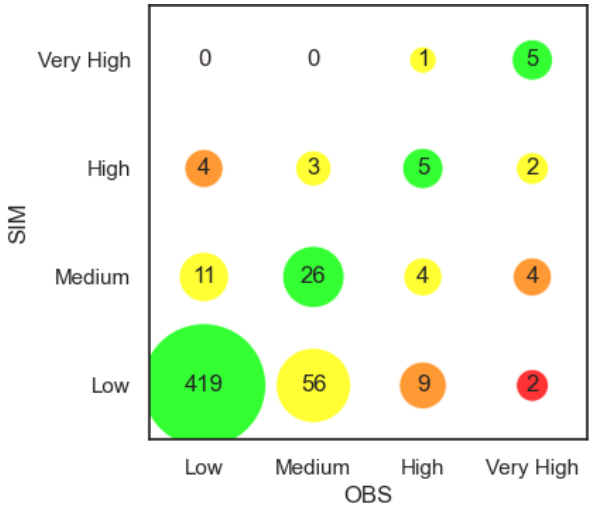
Figure 4. Performance indicator RMSE (root mean squared error) per boosting round for training and test datasets with the optimised hyperparameters

After 7 boosting rounds, the performance with the training dataset achieved a RMSE of about 0.032 loss ratio (or 3.2%) and of 0.054 (or 5.4%) the test dataset. This is acceptable in damage modelling, as seen by other residential building flood damage models (not necessarily for SWF, such as Merz et al. (2013): between 6% and 16%; Wagenaar et al. (2017) models absolute damage, with MAE of 48% or higher; Mohor et al. (2021): 10-11 %).

Beyond the loss ratio (percentage of damage from building value), a better communication format is to convert them in damage classes, providing a more immediate information with less nuance. Based on observed loss ratios across our dataset, we defined the damage classes as: Low: $0 < rloss \leq 3\%$; Medium: $3\% < rloss \leq 9\%$; High: $9\% < rloss \leq 16\%$; Very high: $16\% < rloss$.

Thus, Figure 5 shows the performance of the model with the training and test datasets, aggregated per damage class. The figure shows a cross-tabulation (count) of simulated and observed damage classes, with each grid point accompanied by a bubble whose size is proportional to the count.

https://doi.org/10.5194/egusphere-2026-3462
Preprint. Discussion started: 1 July 2026



205 **Figure 5. Bubble table confronting observed and simulated damage classes (Damage classes are defined as follows: Low: 0 < rloss ≤ 3%; Medium: 3% < rloss ≤ 9%; High: 9% < rloss ≤ 16%; Very high: 16% < rloss). Green bubbles show agreement between simulated and observed classes; yellow show a 1-class different, orange, a 2-class difference, and red a 3-class difference.**

In general, water depth is the most important variable for flood damage modelling. Yet, in many cases, the second most
210 important variable is not too far away in relevance. Figure 6 shows that in FlooDEsT the floor space (fs) even surpasses water depth (wd) in terms of total gain importance metric, showing that building characteristics are as important as the hazard magnitude when modelling relative damage caused by SWF.

https://doi.org/10.5194/egusphere-2026-3462
Preprint. Discussion started: 1 July 2026

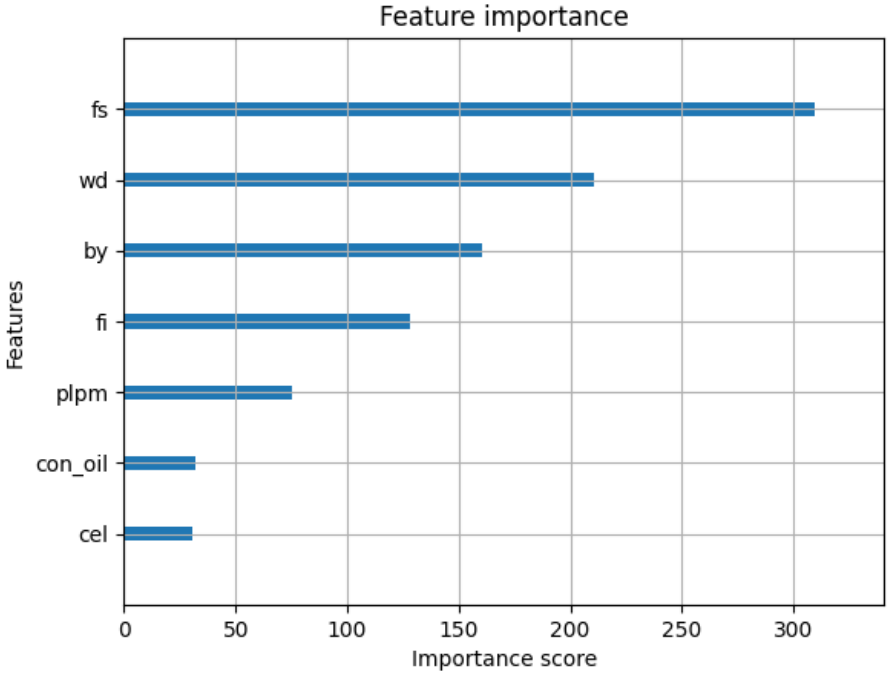


Figure 6. Feature importance plot, as the total gain across all splits each feature is used in. Abbreviations explanation can be found in Table 2.

Shapley values are one method of illustrating and understanding AI methods (also called "explainable AI" - XAI). Shapley values represent the contribution of each predictor variable, as would the coefficient of a linear model, but intrinsically considering the combinations of other predictors' values. In complex, non-linear models, Shapley values provide a more intuitive picture of the effects of a predictor variable onto the dependent variable. Figure 7 show the SHAP values –Shapley Additive exPlanations (Lundberg et al., 2019)– based on the training dataset. The SHAP values show that, in a non-linear fashion, higher water depth and higher flood intensity lead to higher relative damage. The effects of PLPM implementation and the presence of oil contamination had been set as monotonic effects to a damage decreasing and increasing effect, respectively, during the hyperparameters tuning, which can be confirmed in Figure 7. Larger values of the building age (represented in classes, Table 2) reflect higher relative loss, whilst larger floor space generally reflect lower relative damage, although the variability of floor space values does not reflect a monotonic effect across the sample space. 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.

https://doi.org/10.5194/egusphere-2026-3462
Preprint. Discussion started: 1 July 2026

Figure 7. SHAP values of FlooDEsT based on the training dataset

The model diagnostics seem reasonable, with relatively low RMSE and feature importance as well as feature effect agreeing to prior expectations.

3.2 Sensitivity analysis of (missing) predictors

In the pilot regions of Berlin, whilst the hazard variables (water depth and velocity, used to derive flood intensity) were completely available from given scenarios, other variables were not always available. More specifically, floor space (fs) was only partially missing, whilst contamination (con), precautionary measures (plpm), and building year (by) were unavailable. We investigated the effect of the random sampling procedure and the importance of each variable in practice, by removing each variable from the dataset one-by-one and observing the model performance when using random sampling for that variable. This is a very practical measure, as our random sampling considers the observed data to draw representative values instead of fully random values, and shows the model performance in its actual application. A comparison of the model predictions with the XGBoost native handling of missing data or with our missing data imputation is presented at Figure 8. The figure shows that the XGBoost native handling of missing data leads to much higher uncertainty, especially at the presence/absence of a cellar, whilst the model performance with our random sampling imputation deviates little from the baseline. Considering only the results with the random sampling, it is noticeable that flood intensity, alone or combined with water depth, is the most

https://doi.org/10.5194/egusphere-2026-3462
Preprint. Discussion started: 1 July 2026

relevant variable. Floor space also appears as quite relevant, although results vary across training or test dataset, which is also the case for oil contamination. In turn, PLPM implementation and the presence of a cellar show in general only minor deviations. Because this analysis includes the missing data imputation, it is not surprising that the order is not the same as in the feature importance (Figure 6).

Figure 8. Model error comparison with missing data: XGBoost's native handling vs. data imputation. Results are shown for both the training data subset (left) and test data subset (right), with missing data introduced one variable at a time.

We also checked how the variability across the 100 loss ratio estimates is reflected in the damage classes. The uncertainty range, in reference to the resulting mean value, varies from a 0.07 to 19 times the mean. Despite this variability, after reclassifying the loss ratio into a damage class the model is rather stable: in 80% of the cases, all estimates fall in the same damage class. In 15% of the cases, estimates fall into two distinct classes. 3% saw the estimates varying across three damage classes, and in only 2% (n=7) of cases the 100 estimates fell across 4 damage classes. Additionally, in only 2% (n=6) of cases the mean damage class differs from the median one.

3.3 Validation with independent data

Whilst the test data is a subset of the survey data used in model training, validation was done with data from separate surveys, one in Germany (Karlsruhe-2024) and one in the Netherlands (Amsterdam-2014), which nonetheless include the required predictor variables.

https://doi.org/10.5194/egusphere-2026-3462
Preprint. Discussion started: 1 July 2026

Figure 9 shows the comparison between observed and simulated damage classes of the validation set (n=44). The model show a moderate agreement and a large majority of cases with a maximum of one damage class deviation. The model failed to predict the seven cases of 'very high' damage classes. This might be due to a surprising moderate water depth (up to 80 cm), which might have caused the model to predict rather lower damages. Althought the overall agreement is lower in the validation than in the calibration, it shows a balanced result, with no clear bias, raising the model's confidence.

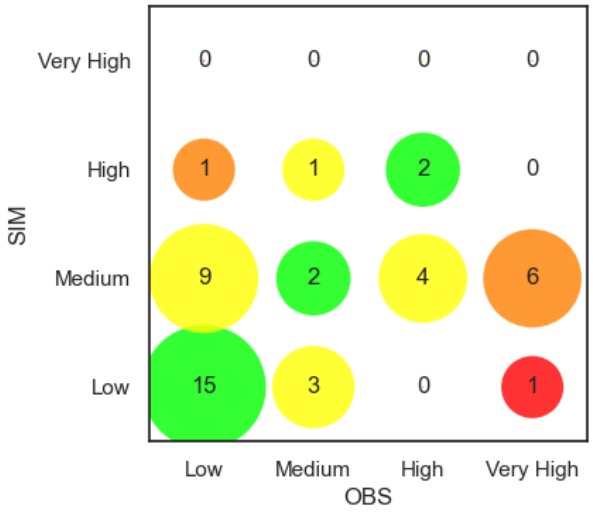


Figure 9. Bubble table confronting observed and simulated damage classes of validation dataset (n= 44). (Damage classes are defined as follows: Low: 0 < rloss ≤ 3%; Medium: 3% < rloss ≤ 9%; High: 9% < rloss ≤ 16%; Very high: 16% < rloss). Green bubbles show agreement between simulated and observed classes; yellow show a 1-class different, orange, a 2-class difference, and red a 3-class difference.

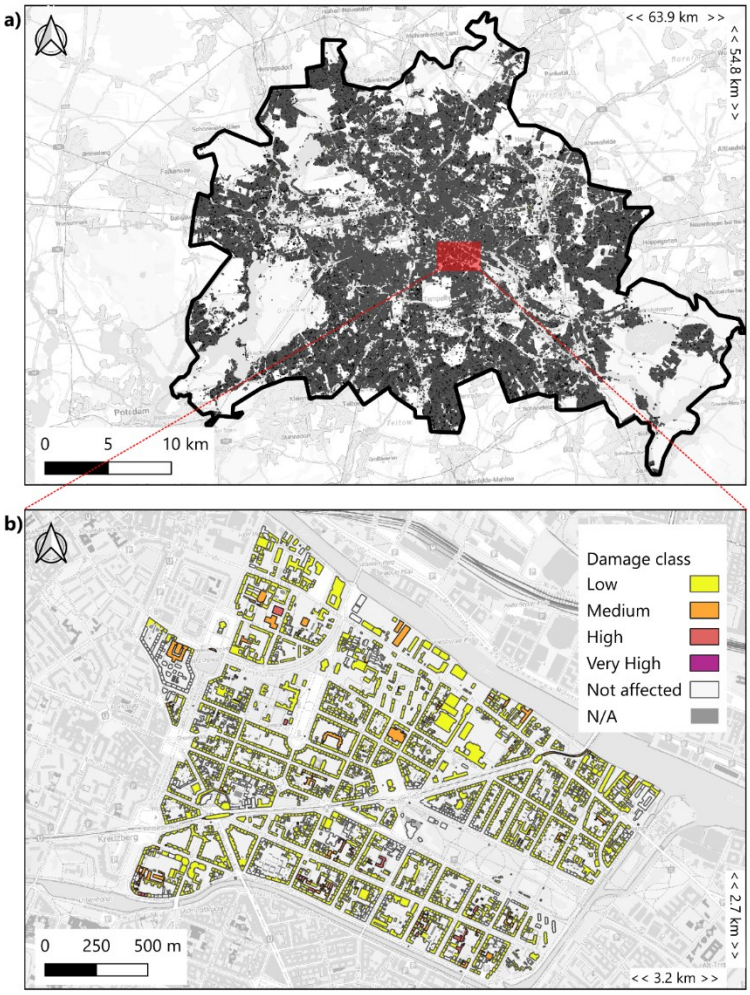
The validation set still only comprises cases from western Europe and represent similar building standards. A broader transferability of the model to different building types is yet to be tested.

3.4 Operational performance

Since the damage is estimated per building unit, the amount of buildings is crucial for the model's runtime, whilst the domain area, or rather the amount of pixels in the raster layers, is relevant for the data extraction. Figure 10 shows the two pilot regions of Berlin, (a) with 455 thousand buildings, of which 440 thousand affected and (b) a subset of this area with 3217 buildings, 2258 of which affected. In both cases, the illustrating hazard scenarios are provided in a 1 m resolution for direct comparison. For details of the scenarios see Bronstert et al. (2025) and Lindenlaub et al. (2026). Because damage estimation is only performed for affected buildings, this is the relevant number reported from now on.The raster screening (without counting the pre-processing of the basis data, which is fixed for all runs afterwards) performed 6.1 times faster than the baseline function for the pilot region with 2.3 k buildings or 2.3 for the region with 440k buildings.

https://doi.org/10.5194/egusphere-2026-3462
Preprint. Discussion started: 1 July 2026

The prediction function is run for each building individually, which can be run in parallel. With help from the package "joblib" (The joblib developers, 2025), in a CPU AMD Ryzen™ AI 7 PRO 350 (8 cores, 2.00GHz), the prediction of the 2.3k buildings improved 8.0 times on average, or 7.8 for 440k buildings. This computational performance improvement shows that the parallelization is little sensitive to scaling, but this is not exhaustive, as not all functions are parallelized. Altogether, both improvements, i.e. parallelization and use of raster screening, represented a 2.7- to 6.6-fold improvement in speed compared to the baseline approach.



- Basemap: basemap.de Web Raster: ©GeoBasis-DE/BKG (2025), https://www.bkg.bund.de/ – CC BY 4.0, https://creativecommons.org/licenses/by/4.0/
- Administrative boundaries: © EuroGeographics (2024) – CC BY 4.0
- Buildings (a): ALKIS Berlin Gebäude (2025) – dl-de/by-2-0, https://www.govdata.de/dl-de/by-2-0
- Buildings (b): ©OpenStreetMap contributors (2024), ODbL 1.0, https://openstreetmap.org/copyright

https://doi.org/10.5194/egusphere-2026-3462
Preprint. Discussion started: 1 July 2026

Figure 10. Estimated damage classes in the pilot regions with (a) 440 thousand and (b) 2258 buildings affected buildings for extreme rainfall scenarios (see Lindenlaub et al., 2026 and Bronstert et al., 2025). Horizontal and vertical lengths denote the extent of the used raster files.

The raster screening and the damage prediction consume 90% to 95% of the application runtime. On average, raster screening takes 35 minutes and the damage prediction 4 minutes, for 440k buildings. An additional 2.1 minutes are required to run each scenario. For 2.2k buildings, the runtime takes about 0.08 minutes. Overall, the application was able to provide damage estimates in respectively 2.1 to 5.6 seconds per thousand buildings for the 2.2 and 440k pilot regions, on an upper-mid-range laptop.

The values depend on the device used, which indicates that running the script with more powerful devices could achieve real-time performance.

4 Conclusion

In this paper we presented the development of the Flood Damage Estimation Tool (FlooDEsT) to rapidly estimate building-level damage from urban pluvial floods, addressing critical gaps in current damage modelling that primarily focus on riverine flooding. The tool was training utilizing XGBoost, a machine learning algorithm well-suited for complex data relationships, along with innovative methods such as smart random sampling to handle missing data, achieving significant improvements over existing tools.

The model's ability to provide reliable damage estimates within seconds per thousand buildings, even when scaled to hundreds of thousands of objects, underscores its potential for real-time applications for risk management and disaster response, meeting the demands for rapid impact assessments.

Validation using independent datasets confirms FlooDEsT's robustness across different urban contexts, achieving an RMSE comparable to other flood damage models reported in the literature (e.g. Merz et al., 2013; Mohor et al., 2021; Wagenaar et al., 2017). This demonstrates the tool's practical applicability beyond the training data regions, offering confidence in its performance for operational use.

However, several limitations remain that should guide future improvements. While the tool performs well for low to moderate damage cases, its accuracy diminishes when predicting extreme losses, which are also rarer, indicating potential areas where additional data or model refinements are needed. The current dataset also reflects a limited geographic scope, primarily focusing on German urban contexts, which may affect its transferability to regions with different architectural characteristics. Future work should prioritize expanding the tool's geographic applicability, further validating its performance under diverse flood scenarios and building characteristics. With an expanded dataset and improved random sampling procedure, missing data imputation could be better contextualized.

Despite these challenges, FlooDEsT represents an important advancement in flood risk assessment by bringing machine learning capabilities to the practical needs of flood damage estimation in the context of SWF. Its integration into existing risk

https://doi.org/10.5194/egusphere-2026-3462
Preprint. Discussion started: 1 July 2026

management systems could enhance decision-making for emergency response and long-term urban planning, particularly in cities increasingly vulnerable to pluvial flooding due to global changes.

Acknowledgments

The 2010, 2014 and 2016 survey on building damage were conducted as a joint venture between the GeoForschungsZentrum Potsdam, the Deutsche Rückversicherung AG, Düsseldorf, and the University of Potsdam.
Data collection campaigns were financially supported by the German Ministry for Education and Research (BMBF) in the frame of the project EVUS (03G0846B) as well as by the German Science Foundation (DFG) in the frame of the Research Training Group NatRiskChange (2043/1).
We hereby declare that generative AI was used to support the Python scripting, restricted to syntax, and language polishing. All other tasks (e.g., logic development) were accomplished by the authors.

Financial support

The collection of the 2024 damage data and this study were financially supported by the German Ministry for Education and Research (BMBF) in the frame of the project Inno_MAUS (FKZ 02WEE1632A).

Code and data availability

The datasets from the 2010 and 2014 events are available via the German flood loss database HOWAS21 (https://doi.org/10.1594/GFZ.SDDB.HOWAS21). Data from the 2016 event will be included by 2029 and can currently be provided upon request only. The dataset from the 2014 Amsterdam case is available under Creative Commons Attribution-Non-Commercial license (CC BY-NC) and can be downloaded from the DANS archive (Spekkers, 2016; https://doi.org/10.17026/dans-x8n-vcbn). The dataset from the 2024 Karlsruhe case can currently be provided upon request only.

Author contribution

CRediT
GSM: Conceptualization, Development, Analysis, Writing of First Draft, Review
SL: Development, Collaborate in First Draft, Review
AT: Resources, Collaborate in First Draft, Review

https://doi.org/10.5194/egusphere-2026-3462
Preprint. Discussion started: 1 July 2026

References

Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M.: Optuna: A Next-generation Hyperparameter Optimization Framework, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD '19: The 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2623–2631, https://doi.org/10.1145/3292500.3330701, 2019.

Bernet, D. B., Prasuhn, V., and Weingartner, R.: Surface water floods in Switzerland: what insurance claim records tell us about the damage in space and time, Nat. Hazards Earth Syst. Sci., 17, 1659–1682, https://doi.org/10.5194/nhess-17-1659-2017, 2017.

Bronstert, A., Chan, E. N. H., Chen, Y., De Vos, F., Dobkowitz, S., Jarajapu, C. D., Kiss, A., Lindenlaub, S., Samprogna Mohor, G., Nijzink, R., Seleem, O. M. A., Thieken, A. H., Thiemann, M., and Xu, Q.: Innovative Instrumente zum Management des urbanen Starkregenrisikos, Hydrol. Wasserbewirtsch. BfG - 692025, 773 KB, 20 Seiten, https://doi.org/10.5675/HYWA_2025.6_6, 2025.

Centre for Research on the Epidemiology of Disasters (CRED): 2024 Disasters in Numbers, Brussels, Belgium: CRED, 2025.

Chen, T. and Guestrin, C.: XGBoost: A Scalable Tree Boosting System, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16: The 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794, https://doi.org/10.1145/2939672.2939785, 2016.

Copernicus Climate Change Service (C3S) and World Meteorological Organization (WMO): European State of the Climate 2024, https://doi.org/10.24381/14J9-S541, 2025.

Cortès, M., Turco, M., Llasat-Botija, M., and Llasat, M. C.: The relationship between precipitation and insurance data for floods in a Mediterranean region (northeast Spain), Nat. Hazards Earth Syst. Sci., 18, 857–868, https://doi.org/10.5194/nhess-18-857-2018, 2018.

Defra / Environment Agency: The Flood Risks to People Methodology, 2006.

Dobkowitz, S., De Vos, L. F., Jarajapu, D. C., Lindenlaub, S., Samprogna Mohor, G., Seleem, O., and Bronstert, A.: The potential of green infrastructure in urban pluvial flood mitigation – a scenario-based modelling study in Berlin, https://doi.org/10.5194/egusphere-2025-5466, 21 November 2025.

Friedman, J.: Greedy function approximation: a gradient boosting machine, Ann. Stat., 29, 1189–1232, 2001.

Gerl, T., Kreibich, H., Franco, G., Marechal, D., and Schröter, K.: A Review of Flood Loss Models as Basis for Harmonization and Benchmarking, PLOS ONE, 11, e0159791, https://doi.org/10.1371/journal.pone.0159791, 2016.

Gradeci, K., Labonnote, N., Sivertsen, E., and Time, B.: The use of insurance data in the analysis of Surface Water Flood events – A systematic review, J. Hydrol., 568, 194–206, https://doi.org/10.1016/j.jhydrol.2018.10.060, 2019.

Hundecha, Y., Parajka, J., and Viglione, A.: Assessment of past flood changes across Europe based on flood-generating processes, Hydrol. Sci. J., 65, 1830–1847, https://doi.org/10.1080/02626667.2020.1782413, 2020.

James, G., Witten, D., Hastie, T., and Tibshirani, R.: An Introduction to Statistical Learning: with Applications in R, Springer US, New York, NY, https://doi.org/10.1007/978-1-0716-1418-1, 2021.

https://doi.org/10.5194/egusphere-2026-3462
Preprint. Discussion started: 1 July 2026

Kellermann, P., Schröter, K., Thieken, A. H., Haubrock, S.-N., and Kreibich, H.: The object-specific flood damage database HOWAS 21, Nat. Hazards Earth Syst. Sci., 20, 2503–2519, https://doi.org/10.5194/nhess-20-2503-2020, 2020.

Laudan, J., Zöller, G., and Thieken, A. H.: Flash floods versus river floods – a comparison of psychological impacts and implications for precautionary behaviour, Nat. Hazards Earth Syst. Sci., 20, 999–1023, https://doi.org/10.5194/nhess-20-999-2020, 2020.

Lindenlaub, S., Samprogna Mohor, G., Creutzfeldt, B., and Thieken, A. H.: Neue Impulse für das kommunale Starkregenrisikomanagement: Kartierung potenzieller Auswirkungen auf ausgewählte Risikoelemente, Hydrol. Wasserbewirtsch. BfG - 692025, 500 KB, 16 Seiten, https://doi.org/10.5675/HYWA_2025.6_5, 2025.

Lindenlaub, S., Samprogna Mohor, G., Creutzfeldt, B., Floto, H., and Thieken, A.: ICFM10_Lindenlaub_etal_ch2.1_7, in: International Conference on Flood Management, 10, London, Ontario, Canada, 193–201, 2026.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., and Lee, S.-I.: Explainable AI for Trees: From Local Explanations to Global Understanding, https://doi.org/10.48550/arXiv.1905.04610, 11 May 2019.

Merz, B., Kreibich, H., and Lall, U.: Multi-variate flood damage assessment: a tree-based data-mining approach, Nat. Hazards Earth Syst. Sci., 13, 53–64, https://doi.org/10.5194/nhess-13-53-2013, 2013.

Merz, B., Kuhlicke, C., Kunz, M., Pittore, M., Babeyko, A., Bresch, D. N., Domeisen, D. I. V., Feser, F., Koszalka, I., Kreibich, H., Pantillon, F., Parolai, S., Pinto, J. G., Punge, H. J., Rivalta, E., Schröter, K., Strehlow, K., Weisse, R., and Wurpts, A.: Impact Forecasting to Support Emergency Management of Natural Hazards, Rev. Geophys., 58, https://doi.org/10/gjdvt5, 2020.

Merz, R. and Blöschl, G.: A process typology of regional floods, Water Resour. Res., 39, https://doi.org/10.1029/2002wr001952, 2003.

Meyer, V., Becker, N., Markantonis, V., Schwarze, R., Van Den Bergh, J. C. J. M., Bouwer, L. M., Bubeck, P., Ciavola, P., Genovese, E., Green, C., Hallegatte, S., Kreibich, H., Lequeux, Q., Logar, I., Papyrakis, E., Pfurtscheller, C., Poussin, J., Przyluski, V., Thieken, A. H., and Viavattene, C.: Review article: Assessing the costs of natural hazards – state of the art and knowledge gaps, Nat. Hazards Earth Syst. Sci., 13, 1351–1373, https://doi.org/10.5194/nhess-13-1351-2013, 2013.

Mohor, G. S., Hudson, P., and Thieken, A. H.: A Comparison of Factors Driving Flood Losses in Households Affected by Different Flood Types, Water Resour. Res., 56, e2019WR025943, https://doi.org/10.1029/2019WR025943, 2020.

Mohor, G. S., Thieken, A. H., and Korup, O.: Residential flood loss estimated from Bayesian multilevel models, Nat. Hazards Earth Syst. Sci., 21, 1599–1614, https://doi.org/10.5194/nhess-21-1599-2021, 2021.

Molinari, D., Menoni, S., and Ballio, F.: Flood damage survey and assessment: new insights from research and practice, AGU/Wiley, USA, 266 pp., 2017.

Molinari, D., Scorzini, A. R., Arrighi, C., Carisi, F., Castelli, F., Domeneghetti, A., Gallazzi, A., Galliani, M., Grelot, F., Kellermann, P., Kreibich, H., Mohor, G. S., Mosimann, M., Natho, S., Richert, C., Schroeter, K., Thieken, A. H., Zischg, A. P., and Ballio, F.: Are flood damage models converging to "reality"? Lessons learnt from a blind test, Nat. Hazards Earth Syst. Sci., 20, 2997–3017, https://doi.org/10.5194/nhess-20-2997-2020, 2020.

Nicklisch, M.: Abschätzung von Fließgeschwindigkeiten und ihrer Bedeutung für Hochwasserschäden am Beispiel des August-Hochwassers 2002, Diplomarbeit, Humbod-Universität zu Berlin, Berlin, 109 pp., 2004.

https://doi.org/10.5194/egusphere-2026-3462
Preprint. Discussion started: 1 July 2026

Perry, M.: Rasterstats: Summary statistics of geospatial raster datasets based on vector geometries, 2024.

Porter, J. R., Marston, M. L., Shu, E., Bauer, M., Lai, K., Wilson, B., and Pope, M.: Estimating Pluvial Depth–Damage Functions for Areas within the United States Using Historical Claims Data, Nat. Hazards Rev., 24, 04022048, https://doi.org/10.1061/NHREFO.NHENG-1543, 2023.

Riese, M., Thieken, A. H., Müggenburg, E., and Bubeck, P.: Synergien und Hemmnisse einer möglichen Integration von Starkregen in die Bearbeitung der europäischen Hochwasserrisikomanagementrichtlinie, Hydrol. Wasserbewirtsch. BfG – Jahrg. 632019, 4ISSN 1439, https://doi.org/10.5675/HYWA_2019.4_1, 2019.

Rözer, V., Müller, M., Bubeck, P., Kienzler, S., Thieken, A., Pech, I., Schröter, K., Buchholz, O., and Kreibich, H.: Coping with Pluvial Floods by Private Households, Water, 8, 304, https://doi.org/10.3390/w8070304, 2016.

Rözer, V., Kreibich, H., Schröter, K., Müller, M., Sairam, N., Doss-Gollin, J., Lall, U., and Merz, B.: Probabilistic Models Significantly Reduce Uncertainty in Hurricane Harvey Pluvial Flood Loss Estimates, Earths Future, 7, 384–394, https://doi.org/10.1029/2018EF001074, 2019.

Rözer, V., Peche, A., Berkhahn, S., Feng, Y., Fuchs, L., Graf, T., Haberlandt, U., Kreibich, H., Sämann, R., Sester, M., Shehu, B., Wahl, J., and Neuweiler, I.: Impact-Based Forecasting for Pluvial Floods, Earths Future, 9, https://doi.org/10.1029/2020ef001851, 2021.

Sieg, T., Vogel, K., Merz, B., and Kreibich, H.: Seamless Estimation of Hydrometeorological Risk Across Spatial Scales, Earths Future, 7, 574–581, https://doi.org/10.1029/2018EF001122, 2019.

Sörensen, J. and Mobini, S.: Pluvial, urban flood mechanisms and characteristics – Assessment based on insurance claims, J. Hydrol., 555, 51–67, https://doi.org/10.1016/j.jhydrol.2017.09.039, 2017.

Spekkers, M., Rözer, V., Thieken, A., Ten Veldhuis, M.-C., and Kreibich, H.: A comparative survey of the impacts of extreme rainfall in two international case studies, Nat. Hazards Earth Syst. Sci., 17, 1337–1355, https://doi.org/10.5194/nhess-17-1337-2017, 2017.

Spekkers, M. H., Kok, M., Clemens, F. H. L. R., and Ten Veldhuis, J. A. E.: A statistical analysis of insurance damage claims related to rainfall extremes, Hydrol. Earth Syst. Sci., 17, 913–922, https://doi.org/10.5194/hess-17-913-2013, 2013.

Spekkers, M. H., Kok, M., Clemens, F. H. L. R., and Ten Veldhuis, J. A. E.: Decision-tree analysis of factors influencing rainfall-related building structure and content damage, Nat. Hazards Earth Syst. Sci., 14, 2531–2547, https://doi.org/10.5194/nhess-14-2531-2014, 2014.

Spekkers, M. H., Clemens, F. H. L. R., and Ten Veldhuis, J. A. E.: On the occurrence of rainstorm damage based on home insurance and weather data, Nat. Hazards Earth Syst. Sci., 15, 261–272, https://doi.org/10.5194/nhess-15-261-2015, 2015.

Strobl, C., Malley, J., and Tutz, G.: An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests., Psychol. Methods, 14, 323–348, https://doi.org/10.1037/a0016973, 2009.

The joblib developers: joblib (1.5.2), , https://doi.org/10.5281/ZENODO.16964648, 2025.

Thieken, A., Kreibich, H., Müller, M., and Lamond, J.: Data Collection for a Better Understanding of What Causes Flood Damage–Experiences with Telephone Surveys, in: Flood damage survey and assessment: new insights from research and practice, Wiley, 95–106, https://doi.org/10.1002/9781119217930.ch7, 2017.

https://doi.org/10.5194/egusphere-2026-3462
Preprint. Discussion started: 1 July 2026

Thieken, A. H., Olschewski, A., Kreibich, H., Kobsch, S., and Merz, B.: Development and evaluation of FLEMOps – a new *F* lood *L* oss *E* stimation *MO* del for the *p* rivate *s* ector, in: Flood Recovery, Innovation and Response I, FLOOD RECOVERY, INNOVATION AND RESPONSE 2008, 315–324, https://doi.org/10.2495/FRIAR080301, 2008.

Thieken, A. H., Samprogna Mohor, G., Kreibich, H., and Müller, M.: 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, 2022.

U.S. Bureau of Reclamation: Downstream Hazard Classification Guidelines, U.S. Department of the Interior, U.S.A., 1988.

Van Ootegem, L., Van Herck, K., Creten, T., Verhofstadt, E., Foresti, L., Goudenhoofdt, E., Reyniers, M., Delobbe, L., Murla Tuyls, D., and Willems, P.: Exploring the potential of multivariate depth-damage and rainfall-damage models: Exploring the potential of multivariate flood damage models, J. Flood Risk Manag., 11, S916–S929, https://doi.org/10.1111/jfr3.12284, 2018.

Vogel, K., Weise, L., Schröter, K., and Thieken, A. H.: Identifying Driving Factors in Flood-Damaging Processes Using Graphical Models, Water Resour. Res., 54, 8864–8889, https://doi.org/10.1029/2018WR022858, 2018.

Wagenaar, D., De Jong, J., and Bouwer, L. M.: Multi-variable flood damage modelling with limited data using supervised learning approaches, Nat. Hazards Earth Syst. Sci., 17, 1683–1696, https://doi.org/10.5194/nhess-17-1683-2017, 2017.