Rözer, V., Kreibich, H., Schröter, K., Müller, M., Sairam, N., Doss-Gollin, J., Lall, U., Merz, B. (2019): Probabilistic models significantly reduce uncertainty in Hurricane Harvey pluvial flood loss estimates. - *Earth's Future*, *7*, 4, pp. 384—394.

DOI: http://doi.org/10.1029/2018EF001074

Earth's Future

RESEARCH ARTICLE
10.1029/2018EF001074

Probabilistic Models Significantly Reduce Uncertainty in Hurricane Harvey Pluvial Flood Loss Estimates

Viktor Rözer[1,2], Heidi Kreibich[1], Kai Schröter[1], Meike Müller[3], Nivedita Sairam[1,4], James Doss-Gollin[5], Upmanu Lall[5,6], and Bruno Merz[1,2]

[1]Section Hydrology, Helmholtz Centre Potsdam GFZ German Research Centre for Geosciences, Potsdam, Germany, [2]Institute for Environmental Sciences and Geography, University Potsdam, Potsdam, Germany, [3]Deutsche Rückversicherung AG, Düsseldorf, Germany, [4]Geography Department, Humboldt University of Berlin, Berlin, Germany, [5]Columbia Water Center, Columbia University, New York, NY, USA, [6]Department of Earth and Environmental Engineering, Columbia University, New York, NY, USA

Abstract Pluvial flood risk is mostly excluded in urban flood risk assessment. However, the risk of pluvial flooding is a growing challenge with a projected increase of extreme rainstorms compounding with an ongoing global urbanization. Considered as a flood type with minimal impacts when rainfall rates exceed the capacity of urban drainage systems, the aftermath of rainfall-triggered flooding during Hurricane Harvey and other events show the urgent need to assess the risk of pluvial flooding. Due to the local extent and small-scale variations, the quantification of pluvial flood risk requires risk assessments on high spatial resolutions. While flood hazard and exposure information is becoming increasingly accurate, the estimation of losses is still a poorly understood component of pluvial flood risk quantification. We use a new probabilistic multivariable modeling approach to estimate pluvial flood losses of individual buildings, explicitly accounting for the associated uncertainties. Except for the water depth as the common most important predictor, we identified the drivers for having loss or not and for the degree of loss to be different. Applying this approach to estimate and validate building structure losses during Hurricane Harvey using a property level data set, we find that the reliability and dispersion of predictive loss distributions vary widely depending on the model and aggregation level of property level loss estimates. Our results show that the use of multivariable zero-inflated beta models reduce the 90% prediction intervalsfor Hurricane Harvey building structure loss estimates on average by 78% (totalling U.S.$3.8 billion) compared to commonly used models.

Key Points:
- Recent severe pluvial flood events highlight the need to integrate pluvial flooding in urban flood risk assessment
- Probabilistic models provide reliable estimation of pluvial flood loss across spatial scales
- Beta distribution model reduces the 90% prediction interval for Hurricane Harvey building loss by U.S.$3.8 billion or 78%

Supporting Information:
- Supporting Information S1
- Figure S1
- Figure S2
- Figure S3
- Figure S4
- Figure S5

Correspondence to:
V. Rözer,
vroezer@gfz-potsdam.de

Citation:
Rözer, V., Kreibich, H., Schröter, K., Müller, M., Sairam, N., Doss-Gollin, J., et al. (2019). Probabilistic models significantly reduce uncertainty in Hurricane Harvey pluvial flood loss estimates. *Earth's Future*, *7*, 384–394. https://doi.org/10.1029/2018EF001074

Received 18 OCT 2018
Accepted 15 MAR 2019
Accepted article online 27 MAR 2019
Published online 9 APR 2019

1. Introduction

Quantifying the future economic risk of pluvial flooding is critical for climate change adaptation of an increasing urban population. Pluvial, or often referred to as surface water flooding, is directly caused by extreme rainstorms with rainfall rates exceeding the capacity of the urban drainage system. Cities around the globe have been impacted by recent pluvial flood events. Large-scale pluvial flooding in the Houston area in Texas during Hurricane Harvey has led to 68 deaths and estimated total damage in the range of U.S.$90 to 160 billion, making it the second most expensive natural disaster in the history of the United States (Blake & Zelinsky, 2018). Other examples include flooding after a rainstorm in Copenhagen 2011 causing total economic damage of U.S.$1 billion (Wojcik et al., 2013) or in Beijing 2012 causing total economic damage of U.S.$1.86 billion and 79 fatalities (Wang et al., 2013). An increasing pluvial flood risk caused by an expected increase of intensity and frequency of heavy precipitation events (Donat et al., 2016; Kundzewicz et al., 2014) combined with an ongoing urbanization with a concentration of population and assets in cities (Skougaard Kaspersen et al., 2015) motivates the need to assess the current and future risk of pluvial flooding. A review by Rosenzweig et al. (2018) identified the lack of knowledge in the quantification of present and future pluvial flood impacts as one of three key research areas for the development of flood resilient cities. However, pluvial flood risk is mostly excluded or neglected in flood risk analysis, although there is evidence that the high frequency of these events lead to long-term cumulative losses comparable to less frequent but severe flood events (Veldhuis, 2011). This lack of knowledge includes risk management and mitigation plans. With few exceptions, official flood hazard maps are exclusively focused on fluvial and coastal flood risk. For the conterminous United States, Wing et al. (2018) found that the poor coverage of urban

catchments in flood hazard maps produced by the Federal Emergency Management Agency (FEMA) has lead to an underestimation of the population affected by pluvial and fluvial flooding by a factor of 2.6–3.1. With scarce information on the hazard, only few loss estimation models for pluvial floods have been developed. Existing approaches include adapting water depth-damage functions (also known as stage-damage models) from river floods (Freni et al., 2010; Olsen et al., 2015; Zhou et al., 2012), using multiple linear regression models (Van Ootegem et al., 2015), or by correlating rainfall measurements with insurance claims or survey data (Spekkers et al., 2014; Van Ootegem et al., 2018). However, the lack of data, the complex nature of the hazard and impact as well as the lack of a consistent quantification of the associated uncertainties, has so far hampered an extensive estimation of expected pluvial flood losses needed to decide on adaptation strategies in cities. Van Ootegem et al. (2015, 2018) construct different multivariate pluvial flood damage models from survey data of a study in Belgium based on water depth-damage and rainfall-damage relationships. Key findings of their study include the importance of additional nonhazard variables such as risk awareness and the effect of reported zero loss cases. However, the results do not provide information as to whether additional variables can also improve loss estimates.

In this study, we use probabilistic high-resolution loss models to estimate pluvial flood losses on different spatial scales. Unlike widely used deterministic stage-damage functions, these probabilistic univariable and multivariable loss models provide a consistent approach to quantify how certain a loss prediction is by providing predictive distributions instead of point estimates. Application and validation of different high-resolution probabilistic loss models in Harris County, Texas, reveal significant differences in the dispersion and reliability of property and county level pluvial flood loss predictions for Hurricane Harvey. Only two out of the six tested models reliably predicted the reported loss with a difference of 78% in the 90% prediction intervals between the two models equaling to an absolute difference of U.S.$3.8 billion for pluvial flood building structure loss in Harris County. These results have major implications for cost-benefit analysis of flood risk management and adaptation decisions in cities.

2. Background

With the need to adapt cities to an expected increase in pluvial flood risk, decision makers face the challenge to take appropriate decisions under the uncertainty of how the risk of pluvial flooding evolves in the future including the expected losses. As uncertainties in flood losses estimates are usually high, probabilistic loss models could greatly aid a comprehensive pluvial flood risk management (Todini, 2018). Unlike deterministic estimates, probabilistic predictions provide continuous predictive distributions where the dispersion of the distribution can provide the range an expected loss would fall in with a certain probability (e.g., 90%). The reliability of a probabilistic prediction can be expressed as the ability of the predictive distribution to cover the actual observed loss. Although probabilistic loss models have been developed for river floods (Dottori et al., 2016; Kreibich et al., 2017; Schröter et al., 2014), these models are the exception and deterministic estimates based on empirical or synthetic relationships between the water depth and the absolute or relative building loss are still widely used for loss estimations for all types of flooding (Gerl et al., 2016; Merz et al., 2010; Scawthorn et al., 2006). The resulting loss estimates in these stage-damage functions are commonly expressed as point estimates for the repair and/or replacement costs in monetary values (i.e., U.S.$) or percentage of the depreciated value of a building. Instead of a direct quantification of uncertainty inherent to probabilistic predictions, uncertainty in stage-damage functions is often based on expert judgment and/or by calculating a range of possible outcomes using different loss functions (Dittes et al., 2018). Missing information, and/or a lack of consistency in quantifying how certain a loss estimate is, makes it challenging for decision makers to, for example, evaluate the potential of an adaptation measure to reduce future losses. While the deviations of point estimates for deterministic loss models are often shown to be reasonably small for loss estimates on large spatial scales typical for river or coastal flooding, loss predictions become highly uncertain on smaller scales (i.e., individual buildings; Merz et al., 2004; Scawthorn et al., 2006). However, due to the local extent and small-scale variations, reliable small-scale loss models are required to quantify pluvial flood risk for a specific location. In this context, we use machine learning as well as different univariable and multivariable probabilistic approaches to investigate three main research objectives: we (i) identify important loss influencing variables and their effect on the uncertainty of loss predictions; (ii) analyze the potential of parametric and nonparametric probabilistic approaches on reducing the dispersion and increasing the reliability of building-level loss estimates; and (iii) evaluate the

applicability of probabilistic multivariable loss models in the context of new sensors and data sources for pluvial flood loss estimation on different spatial scales (Ford et al., 2016; Schröter et al., 2018).

3. Materials and Methods

3.1. Data

We construct a data set that consists of self-reported pluvial flood losses and related information of private households. The data were obtained through a standardized questionnaire using computer-aided telephone surveys after pluvial flood events in five cities in Germany between 2005 and 2014 (Rözer et al., 2016; Spekkers et al., 2017). Based on 120 items in the questionnaire, a data set with 56 predictors and two loss variables is constructed covering eight groups: reported loss, hazard, warning, emergency response, precaution, experience, building information, and social-economic information. The loss variables are represented as relative loss (*rloss*) and a variable with binary information if a building suffered from structural damage or not (*dam*). *rloss* is on the scale from 0 (*no loss*) to 1 (*total loss*), normalizing the reported direct building loss in Euro [EUR] with the total replacement cost value less depreciation of the respective building. We exclude observations where *rloss* could not be derived due to missing information on the building replacement value or the reported loss itself resulting in a total of 431 observations. Out of 56 predictors in the data set, 12 are excluded from the analysis, because of their zero or near-zero variance, resulting in 44 variables to be considered for further analysis. To address the issue of censoring zero loss observations, pluvial flood affected households without direct building loss are included in the data set if water intrusion into the building was reported (9% of observations; see Van Ootegem et al., 2015). Missing values in other variables were imputed using complementary information available in the questionnaire (i.e., missing information of the total living area through building footprint and number of habitable floors). In few cases where causal inference was not possible, missing values are imputed using nearest neighbor imputation. A more detailed description of the data including a table describing all 56 predictors, the two loss variables, the variables excluded from the analysis, and the percentage of imputed missing values is provided in the supporting information (SI; Data section).

3.2. Detection of Important Loss Influencing Variables

Prior to the actual model development, we screen the previously described data set for variables with the highest predictive power given the complex correlations and interdependencies in the data set using machine learning. A reduced set of variables out of the full 44 variables is then used to develop the multivariable probabilistic models described in the following section. The most important loss influencing variables are detected by using an ensemble of variable importance measures of two tree-based (Bagging [cRF; Strobl et al., 2007] and Boosting [GBM; Friedman, 2001]) and two linear regression-based (Ridge [Hoerl & Kennard, 1970 and LASSO [Tibshirani, 1996]) machine learning algorithms. The four different types of algorithms are used in two different settings: a binary classification between *loss/no loss (dam)* and a regression analysis modeling the *degree of loss (rloss)* of a building. Based on the variable importance score of each variable, its rank within each ensemble member as well as its overall rank is determined. The top five variables with the highest overall rank for *rloss* and *dam* are further considered in the model development process.

For details on the variable selection procedure, see SI (Materials and Methods section).

3.3. Probabilistic Loss Estimation Models

Bayesian zero-inflated beta regression (Ospina & Ferrari, 2010) is used to predict the relative loss to a building by pluvial flooding (*rloss*) using the previously selected important loss influencing variables. The probabilistic prediction y for *rloss* on the interval [0,1) is modeled as follows: We define z_i to be a binary variable for the occurrence of flooding in the ith observation and estimate it with a logistic regression:

$$z_i \sim \text{Bernoulli}\left(\gamma X_i\right) \tag{1}$$

where X_i is the vector of predictors for the ith observation, γ is the vector of coefficients, and Bernoulli(θ) indicates a Bernoulli trial with probability θ. Once z_i is known, then we can calculate y_i following a zero-inflated Beta regression model

$$y_i = \begin{cases} \text{Beta}(\alpha_i, \beta_i), & z_i = 1 \\ 0, & z_i = 0 \end{cases} \tag{2}$$

Table 1
Mean Variable Importance Scores of the Five Most Important Predictors for rloss and dam on the Scale (0, 100) for Each Ensemble Member (Tree-Based Bagging [cRF] and Boosting [GBM]; Penalized Regression With L1 [LASSO] and L2 [Ridge] Regularization)

Name	Variable	cRF	GBM	LASSO	Ridge	Avg. rank	Corr
Degree of loss (*rloss*)							
Water depth	wd	100^{1}	100^{1}	94^{1}	97^{1}	1	+
Duration	d	38^{2}	50^{2}	81^{3}	90^{2}	2	+
Basement [Y/N]†	bu	12^{9}	11^{13}	84^{2}	85^{3}	6	+
Contamination [Y/N]	con	15^{8}	9^{17}	77^{4}	81^{4}	6	+
Household size†	hs	17^{4}	17^{8}	45^{7}	64^{5}	6	-
Loss/no loss (*dam*)							
Water depth	wd	99^{1}	100^{1}	89^{1}	90^{2}	1	+
Household size	hs	84^{2}	14^{2}	67^{3}	93^{1}	2	-
Knowledge hazard	pre1	72^{3}	6^{4}	48^{7}	81^{3}	3.5	-
Age of respondent†	age	69^{4}	13^{3}	3^{32a}	42^{9}	6.5	+
Multifamily home [Y/N]	bt	49^{7}	1^{11a}	50^{6}	51^{6}	6.5	-

Note. Corr indicates direction of the trend: "+" increasing; "-" decreasing. Superscript numbers indicate rank within each ensemble member. Avg. rank indicates the overall rank based on the median rank of each ensemble member. Variables marked with a "†" showed no improvement in the predictive performance of the probabilistic loss models and were therefore not considered in the final models.
[a] Importance scores not stable.

where $\alpha_i > 0$ and $\beta_i > 0$ are the shape and scale parameters, respectively, of the Beta distribution. To estimate these parameters, we define

$$\begin{aligned} \alpha_i &= \mu_i \phi \\ \beta_i &= (1 - \mu_i)\, \phi \end{aligned} \tag{3}$$

following Ferrari and Cribari-Neto (2004). This parameterization allows us to define

$$\mu_i = X_i \beta \tag{4}$$

where β is the coefficient vector for the Beta regression. In summation, our zero-inflated beta regression model conducts simultaneous inference on the vector γ, the vector β, and the scalar ϕ, given observations of flood occurrence z, flood damage y (i.e., the variable *rloss*), and predictive variables X.

The probabilistic predictions of *rloss* from the Bayesian zero-inflated beta model (*Beta*) are compared with probabilistic predictions of two additional model types used for empirical flood loss estimation in previous studies. A simple Bayesian parametric model based on a *Gaussian* response distribution is used as a probabilistic representation of a model type widely used in flood loss estimation (Gerl et al., 2016; Van Ootegem et al., 2015) and a nonparametric model based on the *RandomForest* algorithm, used for probabilistic flood loss estimation in previous studies (Schröter et al., 2016). The three model types (*Beta*, *Gaussian*, and *RandomForest*) are fit as univariable and multivariable models (i.e., with a single predictor in X or with multiple predictors) to investigate the effect of additional variables on the predictive performance, resulting in six different models in total. The univariable models are fit using water depth *wd* as their only predictor, reflecting the current standard in flood loss estimation (Gerl et al., 2016; Merz et al., 2010). The univariable parametric models (*Beta* and *Gaussian*) are fit with the square root of the water depth to be comparable with reference functions in previous studies (Merz et al., 2013; Schröter et al., 2014; Wagenaar et al., 2017). All multivariable models use the set of predictors shown in Table 1. For more details on the models including details on the priors of the Bayesian models, see SI (Materials and Methods section).

3.4. Model Validation and Comparison

We validate the probabilistic loss predictions on the building level for the previously described models and data using 10-fold cross validation. For determining the error of the point estimate (median of the predictive distribution), the root-mean-square error (RMSE) and the mean bias error (MBE) are used. For validating

and comparing the reliability of the loss estimate, we calculate the hit rate (HR), meaning the percentage of cases where the observed value lies within the 90% highest density interval (HDI) of the predictive distribution. We use the width of the 90% HDI to evaluate the dispersion of the predictive distribution. In addition, we calculate the interval score, a combined dispersion and reliability score, penalizing predictions based on the width of the 90% HDI and the percentage of observations that are outside the 90% HDI of the respective predictive distributions (Gneiting & Raftery, 2007). To evaluate the effect of including the option to have no building loss in the model, we validate and compare the different models for three scenarios: one where zero-loss observations are removed from the data set prior to fitting the model, one where the zero-loss observations are kept in the data set (zero-loss proportion 9%), and one where the proportion of zero-loss observations is upsampled to 20%. Details on the validation procedure and the different scores used to compare the models are provided in SI (Materials and Methods section).

3.5. Application Harris County, TX

We apply the previously trained probabilistic loss models in Harris County, TX, to analyze the potential for reducing the dispersion and improving the reliability of probabilistic loss estimates for direct building damage of private households caused by pluvial flooding during Hurricane Harvey. To demonstrate the feasibility of probabilistic building-level loss estimation, we construct a high-resolution data set from publicly available data sources for Harris County, TX. Based on refined pluvial flood inundation maps for Hurricane Harvey provided by *JBA Risk Management* (2017), detailed information of affected properties are gathered from the *Harris County Appraisal District Real & Personal Property Database* including the type and value of each affected building (HCAD, 2018). In addition, census information is used to derive the average household size on the block level (U.S. Census Bureau, 2016). Besides this information, the constructed data set contains data on the knowledge about the flood hazard based on if a property is within the 100-year flood zone derived by FEMA (Zone A) and the probability of a property being affected by contamination. The contamination data was created by spatially interpolating reported point sources of contamination from the *National Response Center of United States Coast Guard* and volunteered geographic information using 2-D kernel density interpolation (National Response Center, 2018; Sierra Club). The resulting data set for Harris County contains information of more than 304,000 individual buildings affected by pluvial flooding during Hurricane Harvey. For validation and visualization the property level loss distributions of each model are aggregated on the zip code as well as on the county level. The aggregated loss estimates are validated using the sum and average total building damage from FEMA's Housing Assistance Program available on the zip code level as well as for the entire county for Hurricane Harvey (Federal Emergency Management Agency, 2018). Details on the data sets and models used in Harris County including the validation data are provided in SI (Materials and Methods section).

4. Results

4.1. Important Loss Influencing Variables

Screening the high-dimensional data set for the most important loss influencing variables to be considered in the probabilistic loss models, we find that the drivers for having loss or not having any loss (*dam*) and the drivers for the degree of loss (*rloss*) to a building are different, indicating different damaging mechanisms. While both cases share the water depth as the most important predictor, other important predictors hardly overlap. Looking at the second to fifth most important predictors for *dam*, the resistance of a building and its inhabitants is decisive. Given a low inundation depth, larger households, multifamily buildings, younger residents, and residents who previously informed themselves about pluvial flooding have a lower probability of having any loss. In contrast, the second and fourth most important predictors influencing *rloss* are directly related to the flood intensity. Higher inundation depths, longer flood duration, and contamination of the flood water lead to higher losses. The variable importance scores of the five most important predictors of the four machine learning algorithms their rank within each ensemble member and the median rank of all ensemble members are summarized in Table 1. Starting with the most important predictor both the overall rank and the importance scores drop sharply. Of the five preselected important loss influencing variables shown in Table 1, we find three variables for *rloss* and four variables for *dam* to improve the predictive performance in the probabilistic loss models. Variable importance values for all 44 variables and differences between the machine learning algorithms are shown in SI (Results section).

Table 2
Performance of Loss Model Predictions for Out of Sample Observations (Median)

Model type	Variables	RMSE	MBE	Hitrate (90% PI)	Interval Score (90% PI)
Gaussian	univariable	0.028 (0.018)	0.015 (0.008)	0.91 (0.01)	0.26 (0.01)
	multivariable	0.027 (0.017)	0.013 (0.007)	0.91 (0.02)	0.25 (0.02)
RandomForest	univariable	0.028 (0.017)	0[a](0.009)	0.49[a], [b](0.07)	0.17[a](0.11)
	multivariable	0.025 (0.016)	0.005 (0.008)	0.67[a], [b](0.08)	0.11[a](0.08)
Beta	univariable	0.027 (0.017)	0.010 (0.008)	0.97 (0.06)	0.09[a](0.08)
	multivariable	0.025 (0.017)	0.009 (0.008)	0.95 (0.07)	0.08[a](0.08)

Note. Standard deviation in brackets. RMSE = root-mean-square error; MBE = mean bias error.
[a] Significantly different from Gaussian model for the 0.05 significance level (univariable and multivariable models, respectively). [b] Significantly different from univariable models for the 0.05 significance level for each model type.

4.2. Predictive Performance of Probabilistic Models

The prediction performance of the six probabilistic models (univariable and multivariable models for *Gaussian*, *RandomForest*, and *Beta*) for the cross-validated predictions are summarized in Table 2. Looking solely on the error of the point estimate of the predictions (median of the predictive distribution), we find only a minor nonsignificant reduction in root-mean-square error for the three models for both the univariable and multivariable versions. However, for the 90% HDI of each predictive distribution, the parametric *Beta* and *Gaussian* models are significantly more reliable with an average HR of 97% and 95% for the univariable and multivariable *Beta* models and 91% for both *Gaussian* models compared to 67% and 49% for the *RandomForest* counterparts. However, when we control the HR of the predictive distributions for dispersion and distance to missed observations using the interval score, the high HR scores of the *Gaussian* models can be attributed to consistently wider 90% HDI's (see Figure 1b) compared to the other two models. The difference in shape and width of the predictive distributions of the different models is illustrated in Figure 1a, for the example of a loss estimate for a single building with an observed *rloss* of 0.016. While the *RandomForest* models tend to give very sharp predictive distributions with shapes close to a normal distribution, the predictive distributions of the *Gaussian* and *Beta* models both have longer tails. The almost lognormal shape of the *Gaussian* models is caused by the backtransformation of the logit-transformed predictive distribution. Although the sharp predictive distributions of the *RandomForest* models lead to considerably narrower prediction intervals it significantly increases the risk of the 90% HDI not covering the actual observed loss (see Table 2). With its flexibility in shape and clearly defined interval of the response distribution, we find the *Beta* models to provide the best trade-off between reliability and dispersion. Compared to the widely used reference function (univariable *Gaussian*), the univariable and multivariable have between 47% and 50% narrower HDI's with HRs above 90%. Comparing the difference between the univariable and multivariable models, we find an increase in the variability in shape and width of the predictive distributions for all multivariable models. Although this increase in variability only show a minor, nonsignificant improvement in accuracy, reliability, and dispersion (see Table 2), we find that multivariable models perform significantly better compared to models using the water depth as only predictor when individual predictions are aggregated (see Figure 3c).

4.3. Effect of Zero-Loss Cases on the Damage Estimates

The often low water levels of pluvial flooding compared to river or coastal flooding increases the chances that direct building loss can be completely avoided, although water entered the building. Analyzing different zero-loss proportions, we find that not explicitly accounting for these cases can considerably affect model predictions in terms of reliability and dispersion of the predictive distribution. For the *Gaussian* models, none, and for multivariable *RandomForest* model, 28 of the 38 zero-loss observations in the data set were inside the respective 90% HDI. For increasing the zero-loss proportions we observe a significant increase in the reliability of the *RandomForest* model and a significant increase in the width of the 90% HDI of the loss prediction for the *Gaussian* model (Figure 2). The increase in reliability of the *RandomForest* model reflects the capability of the model to learn implicitly to account for zero-loss cases, when the learning sample becomes large enough. Without the possibility to consider zero-loss cases, a higher proportion of zero-loss observation simply adds additional variability, which the *Gaussian* models cannot explain. Bias caused by varying zero-loss proportions is found to be reduced to a minimum by explicitly accounting for zero-loss

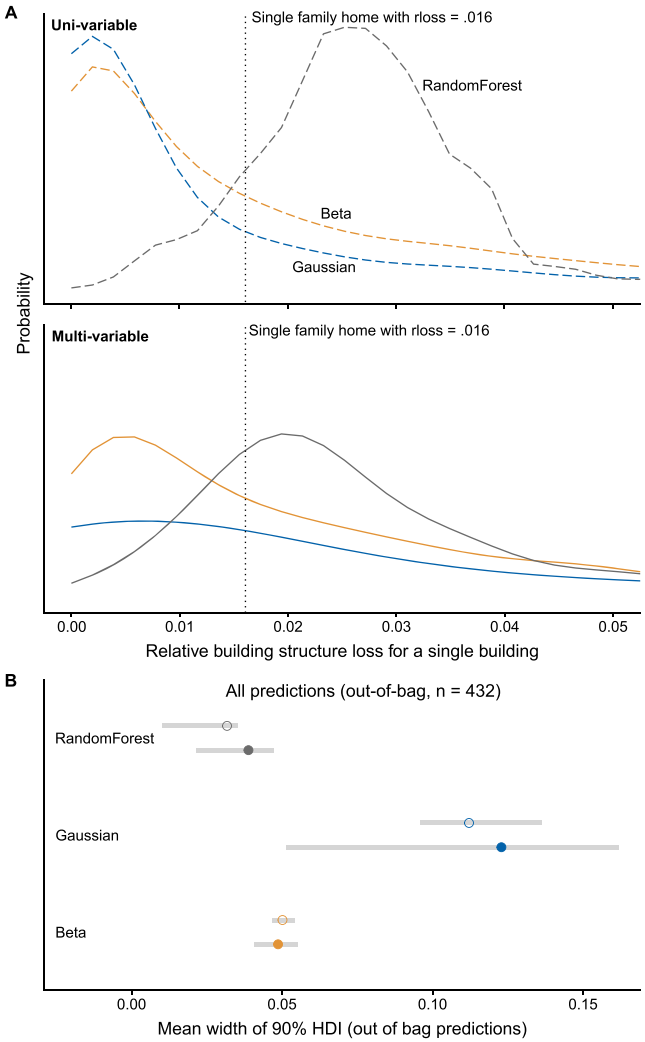

Figure 1. Probabilistic predictive distributions of different univariable and multivariable models (*RandomForest*, *Gaussian*, and *Beta*) for cross-validated observations. The predictive distributions for *Gaussian* and *Beta* models are based on 2000 MCMC samples from the respective posterior predictive distributions. The predictive distributions from *RandomForest* model are based on the predictions of 2,000 individual trees used for training the forest. (a) The different predictive distributions for a single household (single-family home) with a recorded relative loss of 0.016 (dotted vertical line). The upper plot of (a) shows the predictive distributions for three univariable models using the water level as only predictor (dashed lines). The lower plot of (b) shows the same three model types, but with five additional predictors (solid lines). (b) The widths of the 90% HDI for the predictive distributions of all cross-validated observations ($n = 431$) are summarized. The points show the medians for the univariable (hollow) and multivariable (solid) models for the three different model types.The gray boxes show the 25th to 75th percentile ranges for each model. HDI = highest density interval.

observation in the (zero-inflated) *Beta* models (see *Beta* model in Figure 2). Findings for the univariable models are, for the sake of readability, shown in SI (Results section).

4.4. Hurricane Harvey Building Loss for Harris County, TX

Modeled direct losses to the building structure caused by pluvial flooding during Hurricane Harvey in Harris County, TX, are summarized in Figure 3. Our main finding is that the width of the 90% HDI of the predictive distribution for individual buildings can be reduced by 21% or U.S.$3,685 on average when using the multivariable *Beta* model instead of the univariable *Gaussian* model representing the current standard in empirical flood loss estimation. Panel (b) shows the mean relative reduction in the width of the 90% HDI between the two models for individual buildings on the zip code level. For individual buildings we find spatial variations for the average building structure loss ranging from U.S.$544 to U.S.$10,134 with the majority

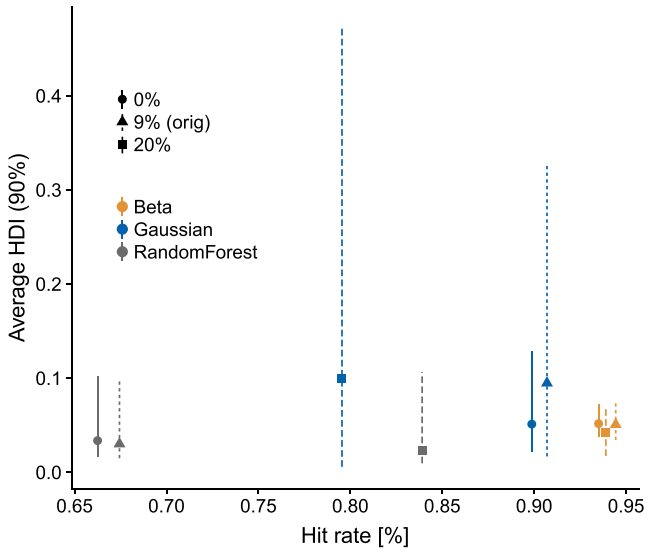

Figure 2. Trade-off between reduction in uncertainty and reliability for cross-validated predictions for different multivariable loss models and different proportions of zero-loss observations in the data set. Results for univariable models are shown in SI (Results section). Uncertainty is represented as mean width of the 90% HDI for all observations. Reliability is represented as proportion of the out-of-sample observation, which are inside the respective 90 % HDI. Error bars represent the 90% interval for the HDI width of all out-of-bag predictions. HDI = highest density interval.

of areas being in the range of U.S.$2,000 to U.S.$5,000. The highest average building structure loss with values above U.S.$7,500 are found west and southwest of Downtown Houston (panel a).

For the aggregated predictive distribution of the absolute loss to the building structure of over 304,000 affected residential buildings (single-family and multifamily homes) in Harris County, the corresponding samples of the individual predictive distributions of each building are summed up. This leads to an effect, know as the central limit theorem, where the Beta-distributed predictive distributions for individual buildings coming from the *Beta* model tend to form a normal distribution when enough individual predictive distributions are summed. In combination with a higher variability, introduced by the additional variables, the considerably higher reliability and lower dispersion of the multi-variable *Beta* model compared to the univariable *Gaussian* model on the building-level vanishes when the predictions are aggregated over a large amount of individual buildings (panel c).

This effect is also described by Sieg (2019) and provides further evidence why univariable stage damage functions based on *Gaussian* response distributions yield sufficiently accurate loss predictions on larger scales while the same model produces highly uncertain loss estimates on the building level. For results aggregated to the county level, we find univariable and multivariable *Gaussian* models to overestimate the absolute building structure losses by U.S.$0.7 and U.S.$3.4 billion, respectively. This can be partly attributed to the underestimation of zero-loss cases described in the previous chapter, which leads to higher intercepts in the model. For the multivariable model this effect is considerably stronger as the model is fit as a linear instead of a square root function (see section 3.3). Of the six models tested, none of the univariable models, and only the aggregated predictive distributions of the multivariable *RandomForest* and *Beta* models are covering the reported loss from FEMA's Housing Assistance Program (U.S.$1.04 billion). Here the multivariable *Beta* performs significantly better with a total reduction in width of the 90% HDI of U.S.$3.8 billion (or 78%) compared to the multivariable *RandomForest* model, providing the best trade-off between dispersion and reliability.

5. Discussion and Conclusions

Despite causing severe losses in cities around the globe, pluvial flooding is still widely neglected when estimating the current and future flood risk in urban areas. This results in a widespread underestimation of flood risk especially in urban areas where fluvial or coastal floods are not the dominant sources of flooding (Rosenzweig et al., 2018). One key limitation in reliably quantifying pluvial flood risk is the local extend of

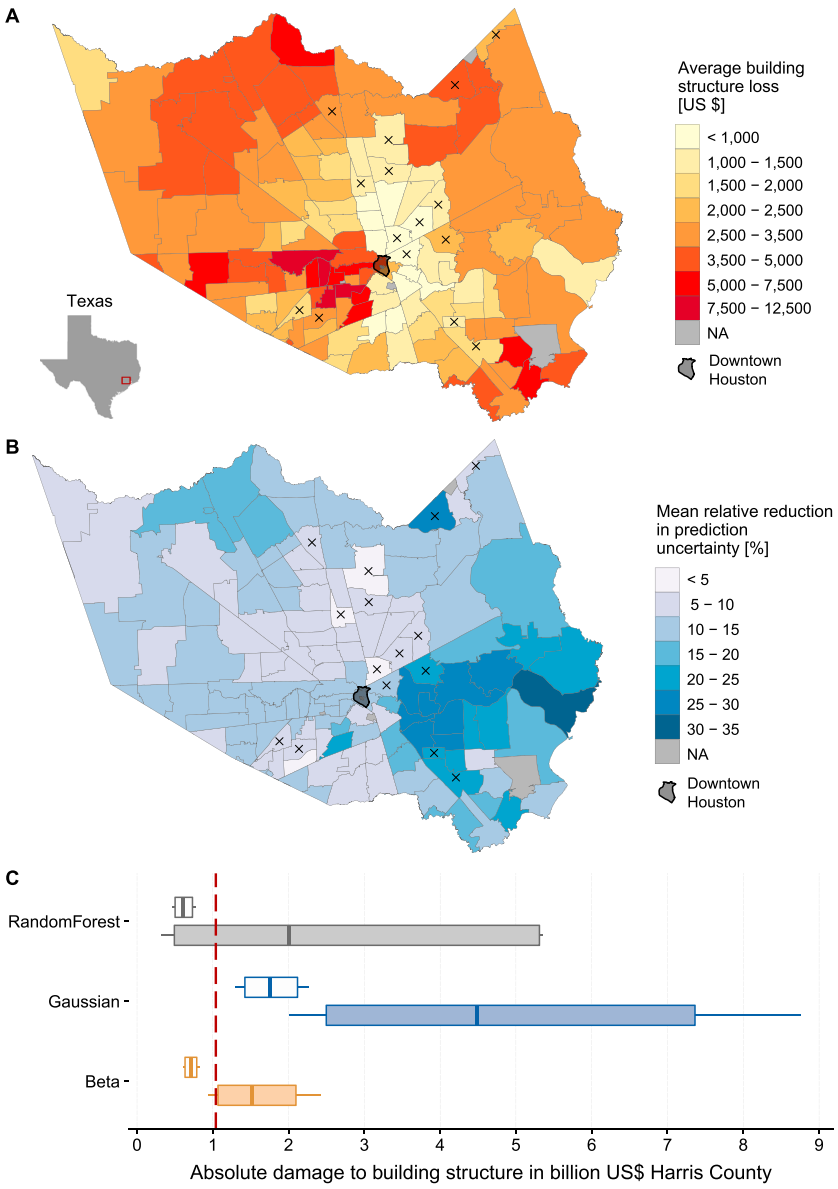

Figure 3. Modeled direct building structure losses for Harris County, TX, caused by pluvial flooding during Hurricane Harvey. (a) The modeled average building structure loss per building aggregated on the zip code level using the multivariable *Beta* model. (b) The average relative reduction in uncertainty (expressed through the width of the 90% HDI) per building between the univariable *Gaussian* model (reference function) and the multivariable *Beta* model in percent aggregated on the zip code level. Crosses in (a) and (b) indicate zip code areas where the reported average building loss is outside the 90% HDI of the modeled average building loss. (c) Box plots of the aggregated predictive distributions of the absolute direct building structure damage for the entire county for three different model types (*RandomForest*, *Gaussian*, and *Beta*) in their univariable (hollow) and multivariable (solid) versions. Bars indicate the median absolute loss, boxes the 90% HDI, and whiskers the 98% HDI of the absolute direct building loss for Harris County. The red dashed line represents the official reported absolute building structure loss based on data from the Federal Emergency Management Agency Housing Assistance Program. HDI = highest density interval.

pluvial floods, requiring loss estimates on spatial scales where damaging processes are still hardly understood and the associated uncertainties are often unknown. We present the first consistent quantification of uncertainties in pluvial flood loss models for private buildings in the shape of predictive distributions using a fully probabilistic modeling approach. We train and validate different univariable and multivariable probabilistic loss models with a local training data set and use these models for a probabilistic estimate of building structure losses of over 304,000 individual buildings in Harris County during Hurricane Harvey. Our analysis reveal significant differences in the dispersion and reliability of the continuous predictive distribu-

tions between different models depending on (i) the use of additional predictors, (ii) the choice of response distribution, (iii) the ability of the model to account for zero-loss cases, and (iv) the spatial scale of the analysis. We find that the assumption of a normal or lognormal distribution of uncertainties in loss estimates, which most loss models implicitly use today, results in unnecessarily wide prediction intervals. In the case of property level predictive distributions, we find that the width of the 90% HDI exceeds the median of the prediction by factor 30 on average. Our results suggest that the with of the 90% HDI for pluvial flood loss estimates on the property level can be significantly reduced by 47% when using a zero-inflated beta distribution instead of normal response distributions without sacrificing the reliability (Table 2). While not evident on the property level, we find that using water depth as only predictor results in an underestimate of the prediction intervals leading to unreliable loss estimates when spatially aggregating loss predictions (Figure 3c). Here, we find additional predictors to improve the pluvial flood loss predictions in two ways: (i) by increasing the variability of individual predictive distributions leading to a more realistic representation of uncertainties when aggregating estimates and (ii) by improving the detection of cases where water entered the building but did not cause any monetary damage to the structure (Figure 2). For the latter our analysis indicate the ability of households to prevent direct damage to their homes should be included in loss models.

The analysis of important loss influencing variables has further shown that the probability of a household to not have any monetary loss to the building structure is—other than for the degree of loss—strongly influenced by household characteristics such as the number of people living in a household and their prior knowledge about the pluvial flood hazard. This highlights the need to account for differences in the ability of households to reduce or avoid damage to their homes in loss models for pluvial floods.

For loss estimates in Harris County, the use of additional predictors in zero-inflated beta models considerably increases the reliability while at the same time significantly reduces the dispersion of the predictive distribution given validation data. For direct building losses aggregated on the county level this reduction accounts for U.S.$3.8 billion or 78% compared to loss models based on normal response distributions. These findings are relevant for a larger discussion on using probabilistic loss estimates for decision making in flood risk management. This includes the potential of probabilistic approaches to improve the spatial transferability of loss models. We further demonstrate the potential to significantly improve the dispersion and reliability of pluvial flood loss estimates using probabilistic models, which goes beyond previous studies considering only point estimates (Van Ootegem et al., 2015; Zhou et al., 2012). Although these results are limited to a quantification of uncertainties of loss predictions, the results can easily be extended for robust decision making on adaptation strategies based on exceeding probabilities, which can be directly derived from predictive distributions. While our results suggest that models that use a zero-inflated beta response distribution provide predictive distributions with a significantly lower dispersion and higher reliability, a general paradigmatic change toward probabilistic models would greatly aid a better understanding of uncertainties in loss models (Todini, 2018). Same is true for multivariable models, where emerging cloud-based reporting systems and open data portals now allow the use of high-dimensional data sets in flood loss modeling.

Acknowledgments
The data collection campaign after the flood event in Münster, Germany, in 2014 was supported by the project '"EVUS Real-Time Prediction of Pluvial Floods and Induced Water Contamination in Urban Areas" (BMBF, 03G0846B), the University of Potsdam, and Deutsche Rückversicherung AG. The data collection campaigns after the pluvial floods in Lohmar and Hersbruck in 2005 were undertaken within the project "URBAS - urban flash floods"; we thank the German Ministry of Education and Research (BMBF; 0330701C) for financial support. Data collection after the pluvial flood in Osnabrück in 2010 were funded by the University of Potsdam, the German Research Centre for Geosciences GFZ, and the Deutsche Rückversicherung AG. Additional financial support is gratefully acknowledged from the German-American Fulbright Commission for V. R. J. D.-G. thanks the NSF GRFP program for support(Grant DGE 16-44869). We would also like to acknowledge JBA Risk Management, who supported our work by providing the pluvial flood inundation map for Hurricane Harvey. The pluvial flood inundation map from JBA Risk Management is available via the OASIS Hub (https://oasishub.co/dataset/surface-water-flooding-footprint-hurricane-harvey-august-2017-jba). The data sets of the flood events in Germany from 2005 and 2010 are available via the German flood damage data base HOWAS21 (http://howas21.gfz-potsdam.de/howas21/). The data set from 2014 will be made available via the HOWAS21 database in June 2023. All other data sets used for the application in Harris County, TX, are openly available and cited in the text and SI. Detailed information on all data sets used for this study and how to access them are available in the supporting information (SI; Data section).

References

Blake, E. S., & Zelinsky, D. A. (2018). Tropical cyclone report Hurricane Harvey (Tech. Rep. No. AL092017). Miami, FL: US Department of Commerce National Oceanic and Atmospheric Administration National Hurricane Center. Retrieved from https://www.nhc.noaa.gov/data/tcr/AL092017_Harvey.pdf

Dittes, B., Kaiser, M., Špačková, O., Rieger, W., Disse, M., & Straub, D. (2018). Risk-based flood protection planning under climate change and modeling uncertainty: A pre-alpine case study. *Natural Hazards and Earth System Sciences*, *18*(5), 1327–1347.

Donat, M. G., Lowry, A. L., Alexander, L. V., O'Gorman, P. A., & Maher, N. (2016). More extreme precipitation in the world's dry and wet regions. *Nature Climate Change*, *6*(5), 508–513.

Dottori, F., Figueiredo, R., Martina, M. L., Molinari, D., & Scorzini, A. (2016). INSYDE: A synthetic, probabilistic flood damage model based on explicit cost analysis. *Natural Hazards and Earth System Sciences*, *16*, 2577–2591.

Federal Emergency Management Agency (2018). Housing assistance data. Retrieved from https://www.fema.gov/media-library/assets/documents/34758, (Accessed: 2018-09-12).

Ferrari, S., & Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, *31*(7), 799–815.

Ford, J. D., Tilleard, S. E., Berrang-Ford, L., Araos, M., Biesbroek, R., Lesnikowski, A. C., et al. (2016). Opinion: Big data has big potential for applications to climate change adaptation. *Proceedings of the National Academy of Sciences*, *113*(39), 10,729–10,732.

Freni, G., La Loggia, G., & Notaro, V. (2010). Uncertainty in urban flood damage assessment due to urban drainage modelling and depth-damage curve estimation. *Water Science and Technology*, *61*(12), 2979–2993.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, *29*, 1189–1232.

Gerl, T., Kreibich, H., Franco, G., Marechal, D., & Schröter, K. (2016). A review of flood loss models as basis for harmonization and benchmarking. *PloS One*, *11*(7), e0159791.

Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, *102*(477), 359–378.
HCAD (2018). "Harris County appraisal district—Real and personal property database". Retrieved from http://pdata.hcad.org/download/index.html, (Accessed: 2018-09-12).
Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55–67.
JBA Risk Management (2017). Oasis hub—Pluvial flooding footprint—Hurricane Harvey—28th August 2017. Retrieved from https://oasishub.co/dataset/surface-water-flooding-footprint-hurricane-harvey-august-2017-jba, (Accessed: 2018-09-12).
Kreibich, H., Botto, A., Merz, B., & Schröter, K. (2017). Probabilistic, multivariable flood loss modeling on the mesoscale with BT-FLEMO. *Risk Analysis*, *37*(4), 774–787.
Kundzewicz, Z. W., Kanae, S., Seneviratne, S. I., Handmer, J., Nicholls, N., Peduzzi, P., et al. (2014). Flood risk and climate change: Global and regional perspectives. *Hydrological Sciences Journal*, *59*(1), 1–28.
Merz, B., Kreibich, H., & Lall, U. (2013). Multi-variate flood damage assessment: A tree-based data-mining approach. *Natural Hazards and Earth System Sciences*, *13*(1), 53–64.
Merz, B., Kreibich, H., Schwarze, R., & Thieken, A. (2010). Review article "Assessment of economic flood damage". *Natural Hazards and Earth System Sciences*, *10*(8), 1697–1724.
Merz, B., Kreibich, H., Thieken, A., & Schmidtke, R. (2004). Estimation uncertainty of direct monetary flood damage to buildings. *Natural Hazards and Earth System Science*, *4*(1), 153–163.
National Response Center (2018). United States Coast Guard—National Response Center Database. Retrieved from http://www.nrc.uscg.mil/FOIAFiles/CY17.xlsx, (Accessed: 2018-09-12).
Olsen, A. S., Zhou, Q., Linde, J. J., & Arnbjerg-Nielsen, K. (2015). Comparing methods of calculating expected annual damage in urban pluvial flood risk assessments. *Water*, *7*(1), 255–270.
Ospina, R., & Ferrari, S. L. (2010). Inflated beta distributions. *Statistical Papers*, *51*(1), 111.
Rosenzweig, B. R., McPhillips, L., Chang, H., Cheng, C., Welty, C., Matsler, M., et al. (2018). Pluvial flood risk and opportunities for resilience. *Wiley Interdisciplinary Reviews: Water*, *5*, e1302.
Rözer, V., Müller, M., Bubeck, P., Kienzler, S., Thieken, A., Pech, I., et al. (2016). Coping with pluvial floods by private households. *Water*, *8*(7), 304.
Scawthorn, C., Flores, P., Blais, N., Seligson, H., Tate, E., Chang, S., et al. (2006). Hazus-MH flood loss estimation methodology. II. Damage and loss assessment. *Natural Hazards Review*, *7*(2), 72–81.
Schröter, K., Kreibich, H., Vogel, K., Riggelsen, C., Scherbaum, F., & Merz, B. (2014). How useful are complex flood damage models? *Water Resources Research*, *50*, 3378–3395. https://doi.org/10.1002/2013WR014396
Schröter, K., Lüdtke, S., Redweik, R., Meier, J., Bochow, M., Ross, L., et al. (2018). Flood loss estimation using 3D city models and remote sensing data. *Environmental Modelling & Software*, *105*, 118–131.
Schröter, K., Lüdtke, S., Vogel, K., Kreibich, H., & Merz, B. (2016). Tracing the value of data for flood loss modelling. In *E3s web of conferences*, (Vol. *7*, pp. 05005).
Sieg, T. (2019). Reliability of flood damage estimations across spatial scales (Doctoral dissertation). University of Potsdam, Potsdam, Germany.
Skougaard Kaspersen, P., Høegh Ravn, N, Arnbjerg-Nielsen, K., Madsen, H., & Drews, M. (2015). Influence of urban land cover changes and climate change for the exposure of european cities to flooding during high-intensity precipitation. *Proceedings of the International Association of Hydrological Sciences*, *370*, 21–27.
Spekkers, M., Kok, M., Clemens, F., & Ten Veldhuis, J. (2014). Decision-tree analysis of factors influencing rainfall-related building structure and content damage. *Natural Hazards and Earth System Sciences*, *14*(9), 2531–2547.
Spekkers, M., Rözer, V., Thieken, A., ten Veldhuis, M.-C., & Kreibich, H. (2017). A comparative survey of the impacts of extreme rainfall in two international case studies. *Natural Hazards and Earth System Sciences*, *17*(8), 1337–1355.
Strobl, C., Boulesteix, A.-L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, *8*(1), 25.
Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, *58*, 267–288.
Todini, E. (2018). Paradigmatic changes required in water resources management to benefit from probabilistic forecasts. *Water Security*, *3*, 9–17.
U.S. Census Bureau (2016). Occupancy characteristics 2012–2016 American Community survey 5-year estimates. Retrieved from https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=ACS_16_5YRB25010&prodType=table, (Accessed: 2018-09-12).
Van Ootegem, L., Van Herck, K., Creten, T., Verhofstadt, E., Foresti, L., Goudenhoofdt, E., et al. (2018). Exploring the potential of multivariate depth-damage and rainfall-damage models. *Journal of Flood Risk Management*, *11*, S916–S929.
Van Ootegem, L., Verhofstadt, E., Van Herck, K., & Creten, T. (2015). Multivariate pluvial flood damage models. *Environmental Impact Assessment Review*, *54*, 91–100.
Veldhuis, J. (2011). How the choice of flood damage metrics influences urban flood risk assessment. *Journal of Flood Risk Management*, *4*(4), 281–287.
Wagenaar, D., de Jong, J., & Bouwer, L. M. (2017). Multi-variable flood damage modelling with limited data using supervised learning approaches. *Natural Hazards and Earth System Sciences*, *17*(9), 1683.
Wang, K., Wang, L., Wei, Y.-M., & Ye, M. (2013). Beijing storm of July 21, 2012: Observations and reflections. *Natural Hazards*, *67*(2), 969–974.
Wing, O. E., Bates, P. D., Smith, A. M., Sampson, C. C., Johnson, K. A., Fargione, J., & Morefield, P. (2018). Estimates of present and future flood risk in the conterminous united states. *Environmental Research Letters*, *13*(3), 034023.
Wojcik, O., Holt, J., Kjerulf, A., Müller, L, Ethelberg, S., & Mølbak, K (2013). Personal protective equipment, hygiene behaviours and occupational risk of illness after July 2011 flood in Copenhagen, Denmark. *Epidemiology & Infection*, *141*(8), 1756–1763.
Zhou, Q., Mikkelsen, P. S., Halsnæs, K., & Arnbjerg-Nielsen, K. (2012). Framework for economic pluvial flood risk assessment considering climate change effects and adaptation benefits. *Journal of Hydrology*, *414*, 539–549.