Supporting Information for

FLEMO*flash* - Flood Loss Estimation Models for companies and households affected by *flash* floods

Apoorva Singh[1,2†], Ravi Kumar Guntu[1†*], Nivedita Sairam[1], Kasra Rafiezadeh Shahi[1], Anna Buch[3], Melanie Fischer[1], Chandrika Thulaseedharan Dhanya[2], and Heidi Kreibich[1]

[1]Section 4.4 Hydrology, GFZ Helmholtz Centre for Geosciences, 14473 Potsdam, Germany

[2]Department of Civil Engineering, Indian Institute of Technology Delhi, 110016 New Delhi, India

[3]Institute of Geography, University of Heidelberg, 69117 Heidelberg, Germany

[*]Correspondence to: Ravi Kumar Guntu (ravikumar.guntu@gfz.de)

[†]These authors contributed equally to this work

Text S1. Bayesian Network

We employed the 'Tabu Search' algorithm to learn the directed acyclic graph structure of the Bayesian networks from the company and private household loss data. This optimization routine searches the space of candidate Bayesian network structures for a directed acyclic graph that maximizes a predefined goodness-of-fit score. Specifically, we opted for the Bayesian Dirichlet equivalent score. Once we obtained the Bayesian network structure, we learned the conditional probability tables from the survey dataset through Bayesian parameter estimation.

To make predictions with the fitted Bayesian network, we used exact Bayesian inference by querying the conditional probability of the target variable, relative loss, conditioned on the observed predictors. For further details on Bayesian network theory, we refer to the literature (Jensen & Nielsen, 2007; Nagarajan et al., 2013). The proposed Bayesian networks were implemented in R using the packages ‘bnlearn’ (Scutari & Denis, 2021) and ‘gRain’(HøJsgaard, 2012).

Text S2. Model validation

The model validation process follows the methodology outlined in Schoppa et al., (2020). All models provide probabilistic predictions rather than deterministic loss estimates. However, they do not offer analytical predictive distributions but simulated approximations in the form of samples. For each model we sampled 1000 values from the conditional response distribution and evaluated this probabilistic response in terms of accuracy, sharpness, and calibration. Within each asset dataset, we estimated the model test errors through repeated cross validation to obtain robust estimates of true model performance. We initiated 100 independent runs of ten-fold cross validation with varying, random data partitioning. In each of the tenfold cross-validation runs, every company is held out of the training set for prediction exactly once. We validate model performance for each cross-validation fold using three performance metrics:

1. The mean absolute error (MAE) for the mean of the predictive distribution. The MAE evaluates the accuracy of a point forecast and averages the absolute differences between the observed response and predicted point estimate over the number of observations.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|x_i - y_i| \tag{1}$$

Where n is the number of observations, x_i is the predicted point estimate and y_i is the observed response for observation i.

2. The mean bias error (MBE), which quantifies model overestimation and underestimation in the mean of predictive distributions.

$$MBE = \frac{1}{n}\sum_{i=1}^{n}(x_i - y_i) \tag{2}$$

Where n is the number of observations, x_i is the predicted point estimate and y_i is the observed response for observation i. Unlike the MAE, since the absolute value of error is not taken in the MBE, it consists of both positive and negative values, and it serves as a measure of average prediction bias.

3. The continuous ranked probability score (CRPS), which is a proper scoring rule that evaluates the entire continuous distribution of a probabilistic forecast. It jointly assesses the sharpness and calibration of the predictive distribution and generalizes the absolute error. Hence the error can be compared directly to the MAE. The CRPS for one observation y_i is defined as

$$CRPS_i(F_i, y_i) = \int_{-\infty}^{\infty}(F_i(x) - 1\{y_i \leq x\})^2 dx \tag{3}$$

Where $F_i(x)$ is the CDF of the predictive distribution $f_i(x)$ and $1\{.\}$ Is the indicator function. We compute the CRPS with an empirical CDF estimated from samples of $f_i(x)$. For details on the numerical implementation of the CRPS for simulated forecasts, we refer to the corresponding literature (Gneiting & Katzfuss, 2014; Jordan et al., 2019; Krüger et al., 2021). For the proportional response variable, $rloss$, the CRPS is defined on the interval $[0,1]$ with the optimum at 0. Note that the CRPS is calculated individually for each observation. For the comparison with the MAE, we computed the mean CRPS value in each cross-validation fold.

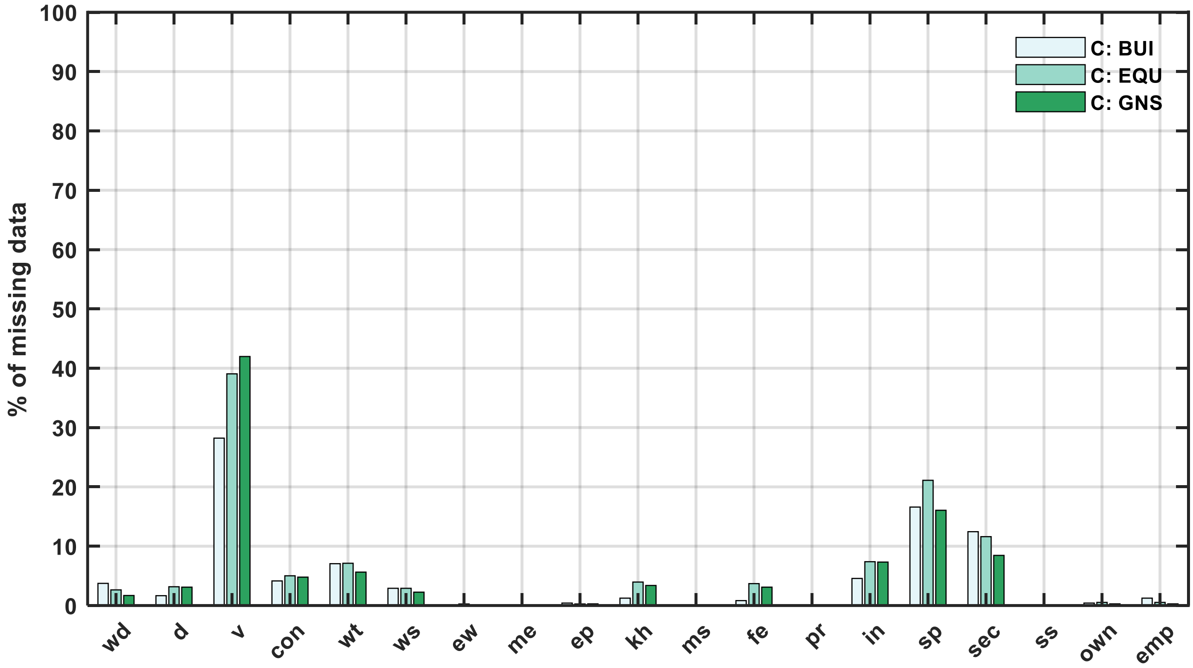

Figure S1: Illustration of missing data in each variable for developing company damage models. For instance, in the $C{:}BUI$ loss category, a value of 0.10 indicates that 24 observations out of 241 are missing for the respective variable.

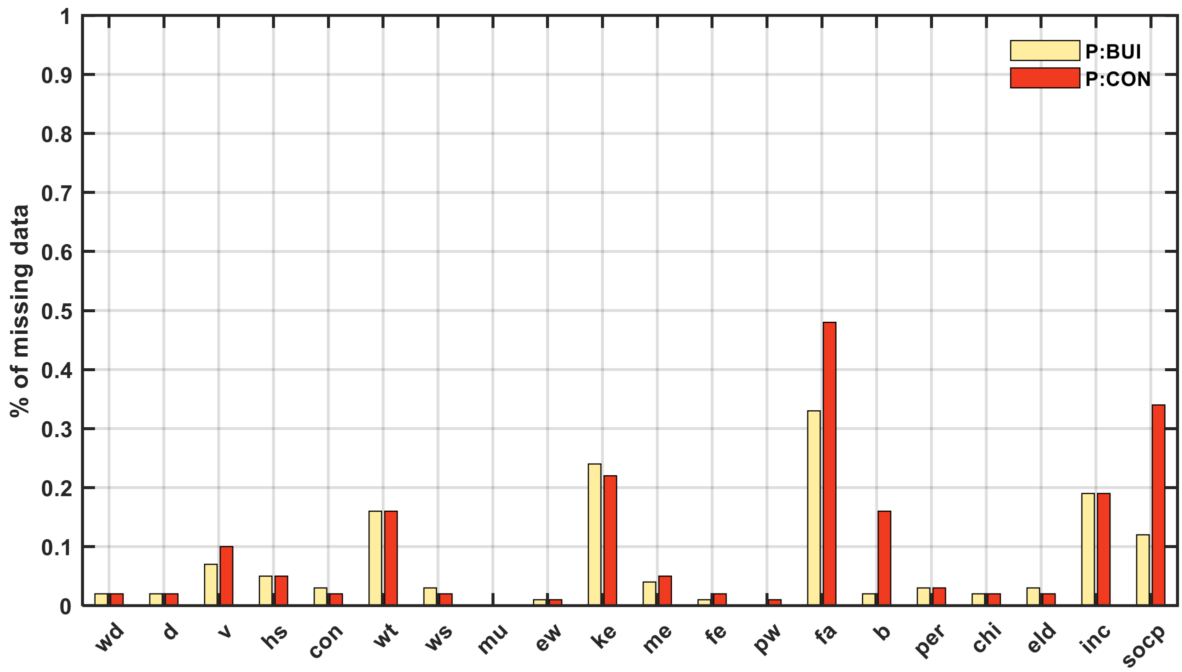

Figure S2: Same as Figure S1 but for private households.

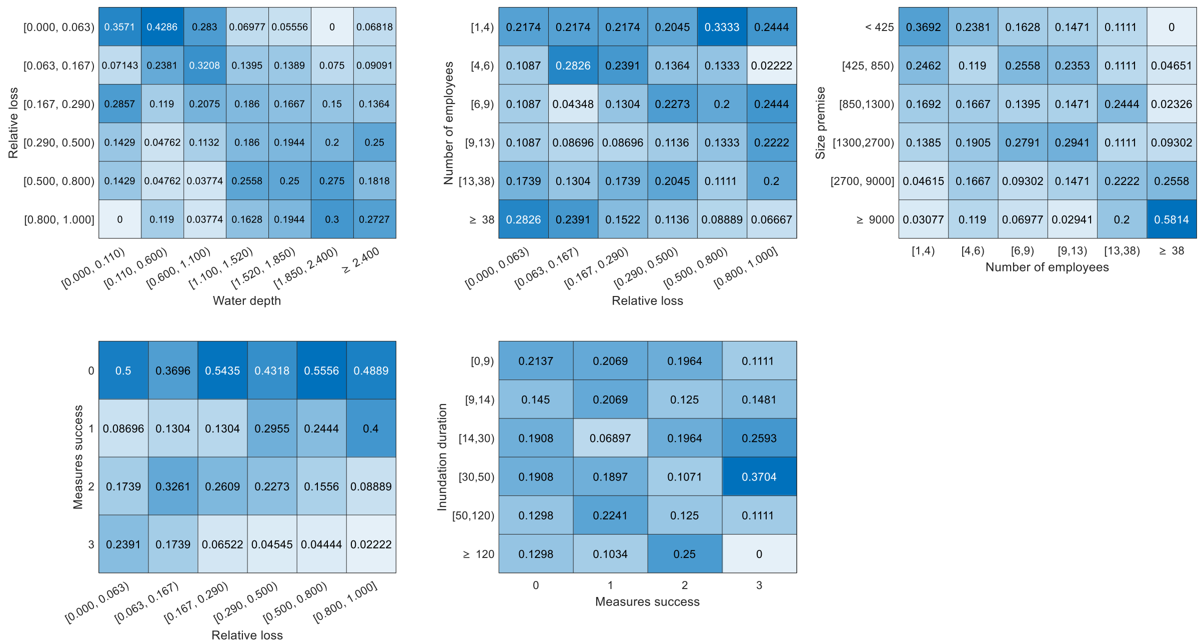

Figure S3: Conditional probability table of C:BUI Bayesian network. X-axis denotes the parent node, while Y-axis denotes the children node within the network. Each heatmap illustrates the conditional probabilities between a child and its parent node, with increasing intensity of blue indicating higher probability values. Numerical values are displayed in each cell, and a colorbar in each subplot shows the corresponding probability scale.

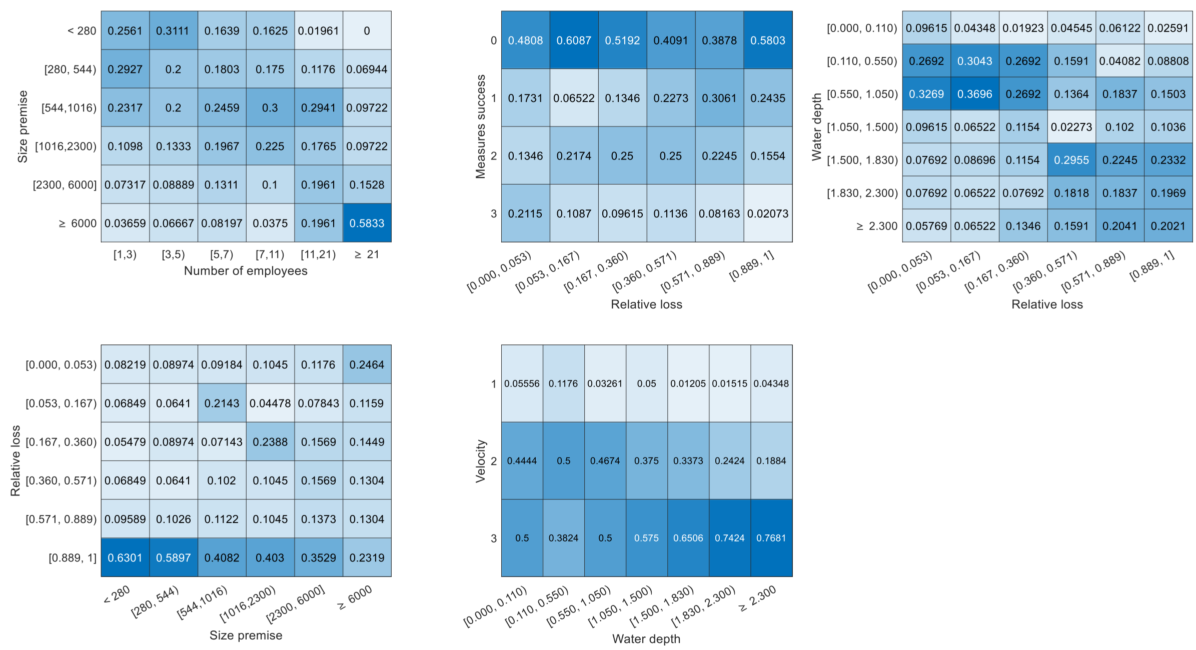

Figure S4: same as Fig.S1, but for C:EQU

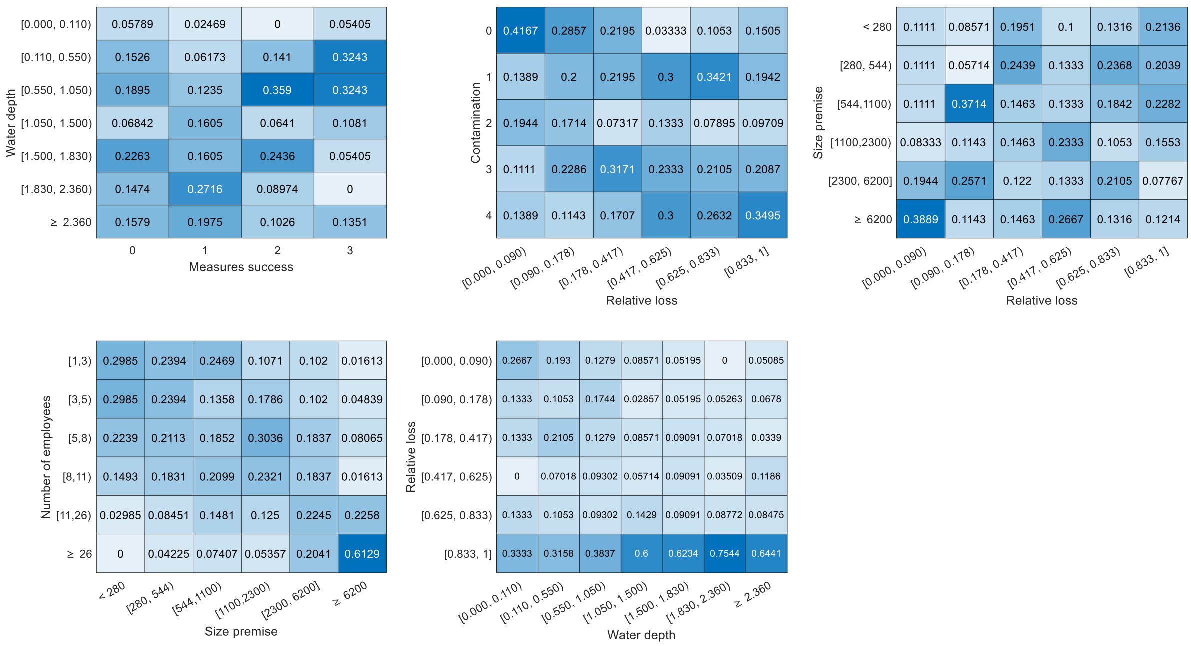

Figure S5: Same as Fig.S1 but for C: GNS

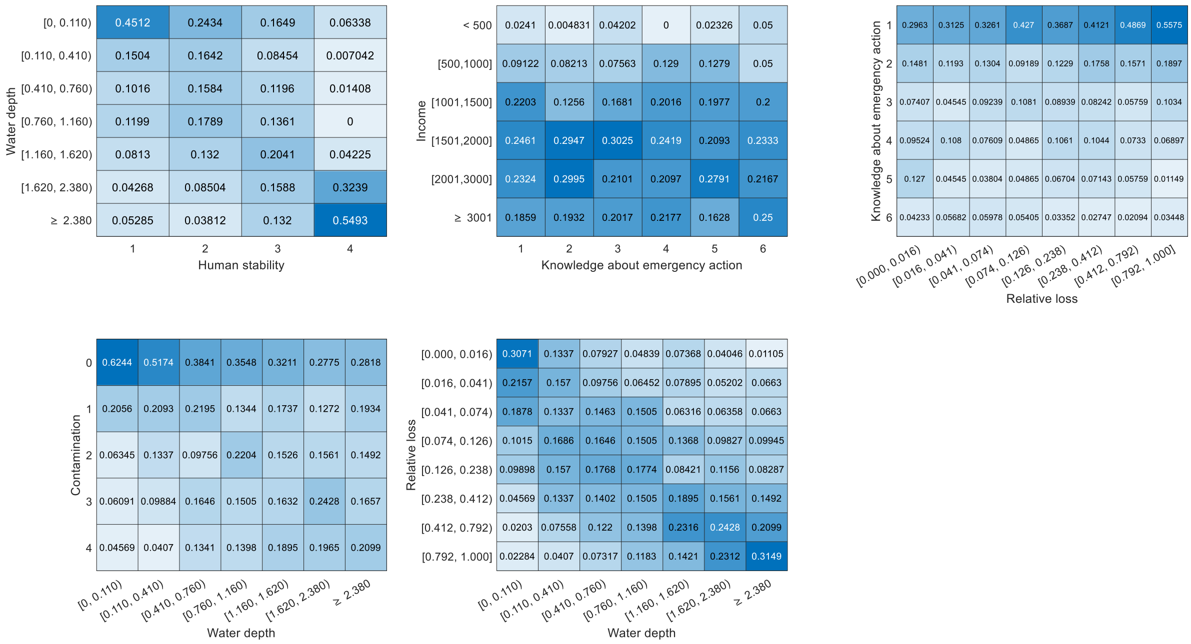

Figure S6: same as Fig.S1, but for P:CON

Table S1: Prior distributions for probabilistic stage-damage function classified according to model parameters and predictors. Each prior contains a note motivating the prior choice.

Model parameters	**Predictors**	**Prior**	**Notes**
α – Intercept	-	$Student-t(3,0,10)$	Weakly informative standard prior for intercepts in brms-package.
β – regression parameter	wd	$Normal(1,1.5)$	Higher water depth causes larger flood loss.
λ – zero-and-one-inflation	-	$Beta(1,1)$	The parameter represents a probability and, hence, is constrained to the interval [0,1].
γ – conditional one-inflation	-	$Beta(1,1)$	The parameter represents a probability and, hence, is constrained to the interval [0,1].
ϕ – beta precision	-	$Gamma(0.01,0.01)$	The precision of the beta distribution has to be positive.

References

Gneiting, T., & Katzfuss, M. (2014). Probabilistic Forecasting. *Annual Review of Statistics and Its Application*, *1*(1), 125–151. https://doi.org/10.1146/annurev-statistics-062713-085831

HøJsgaard, S. (2012). Graphical Independence Networks with the **gRain** Package for *R*. *Journal of Statistical Software*, *46*(10). https://doi.org/10.18637/jss.v046.i10

Jensen, F. V., & Nielsen, T. D. (2007). *Bayesian Networks and Decision Graphs*. New York, NY: Springer New York. https://doi.org/10.1007/978-0-387-68282-2

Jordan, A., Krüger, F., & Lerch, S. (2019). Evaluating Probabilistic Forecasts with **scoringRules**. *Journal of Statistical Software*, *90*(12). https://doi.org/10.18637/jss.v090.i12

Krüger, F., Lerch, S., Thorarinsdottir, T., & Gneiting, T. (2021). Predictive Inference Based on Markov Chain Monte Carlo Output. *International Statistical Review*, *89*(2), 274–301. https://doi.org/10.1111/insr.12405

Nagarajan, R., Scutari, M., & Lèbre, S. (2013). *Bayesian Networks in R: with Applications in Systems Biology*. New York, NY: Springer New York. https://doi.org/10.1007/978-1-4614-6446-4

Schoppa, L., Sieg, T., Vogel, K., Zöller, G., & Kreibich, H. (2020). Probabilistic Flood Loss Models for Companies. *Water Resources Research*, *56*(9), e2020WR027649. https://doi.org/10.1029/2020WR027649

Scutari, M., & Denis, J.-B. (2021). *Bayesian Networks: With Examples in R* (2nd ed.). Boca Raton: Chapman and Hall/CRC. https://doi.org/10.1201/9780429347436