Nat. Hazards Earth Syst. Sci., 18, 1297–1314, 2018
https://doi.org/10.5194/nhess-18-1297-2018

Natural Hazards and Earth System Sciences
Open Access
EGU

Multi-model ensembles for assessment of flood losses and associated uncertainty

Rui Figueiredo[1,2], **Kai Schröter**[2], **Alexander Weiss-Motz**[2], **Mario L. V. Martina**[1], **and Heidi Kreibich**[2]

[1]Scuola Universitaria Superiore IUSS Pavia, Pavia, Italy
[2]GFZ German Research Centre for Geosciences, Sect. 5.4: Hydrology, Potsdam, Germany

Correspondence: Rui Figueiredo (rui.figueiredo@iusspavia.it)

Received: 2 October 2017 – Discussion started: 16 October 2017
Revised: 26 March 2018 – Accepted: 11 April 2018 – Published: 3 May 2018

Abstract. Flood loss modelling is a crucial part of risk assessments. However, it is subject to large uncertainty that is often neglected. Most models available in the literature are deterministic, providing only single point estimates of flood loss, and large disparities tend to exist among them. Adopting any one such model in a risk assessment context is likely to lead to inaccurate loss estimates and sub-optimal decision-making. In this paper, we propose the use of multi-model ensembles to address these issues. This approach, which has been applied successfully in other scientific fields, is based on the combination of different model outputs with the aim of improving the skill and usefulness of predictions. We first propose a model rating framework to support ensemble construction, based on a probability tree of model properties, which establishes relative degrees of belief between candidate models. Using 20 flood loss models in two test cases, we then construct numerous multi-model ensembles, based both on the rating framework and on a stochastic method, differing in terms of participating members, ensemble size and model weights. We evaluate the performance of ensemble means, as well as their probabilistic skill and reliability. Our results demonstrate that well-designed multi-model ensembles represent a pragmatic approach to consistently obtain more accurate flood loss estimates and reliable probability distributions of model uncertainty.

1 Introduction

Effective management of flood risk requires comprehensive risk assessment studies that consider not only the hazard component, but also the impacts that the phenomena may have on the built environment, economy and society (Messner and Meyer, 2006). This integrated approach has gained importance over recent decades, and with it so has the scientific attention given to flood vulnerability models describing the relationships between flood intensity metrics and damage to physical assets, also known as flood loss models. A large number of models have become available in the scientific literature. However, despite progress in this field, many challenges persist in their development, and flood loss models tend to be quite heterogeneous. This often results in practical difficulties when they are to be applied in risk assessment studies (Gerl et al., 2016; Jongman et al., 2012), as described below.

Flood damage mechanisms are complex, being dependent on different properties of flood events, such as water depth, flow velocity and flood duration, as well as on the physical characteristics of the exposed assets (Kelman and Spence, 2004). Precautionary and socio-economic factors can also influence their degree of vulnerability (Thieken et al., 2005). Building accurate and reliable flood loss models that account for all these factors is a challenging task. Model development is hampered by limited knowledge about damage-influencing factors, as well as limited data availability (Merz et al., 2010). It is therefore unsurprising that traditional flood loss models tend to be rather simple, often using water depth as the only explanatory variable to describe damage and loss to coarsely defined groups of assets (Green et al., 2011; Smith, 1994). However, the limited predictive ability and high degree of uncertainty associated with such models has been acknowledged (Krzysztofowicz and Davis, 1983; Merz et al., 2004), and more complex models that consider additional explanatory variables have been developed (Dottori et al., 2016;

Elmer et al., 2010; Merz et al., 2013). Regardless, uncertainty in flood loss modelling is to some extent inevitable (Schröter et al., 2014).

Furthermore, flood loss models are usually developed for specific regions, ranging from country to catchment or municipality level, with smaller scales making up the majority of models (Gerl et al., 2016). Lack of available flood loss models in many regions often leads to the transfer of models in space, resulting in their application to contexts with different built environments and/or socio-economic settings than originally intended. However, this is generally done with insufficient justification, and flood loss models have been shown to offer lower predictive ability under such circumstances (Cammerer et al., 2013; Jongman et al., 2012; Schröter et al., 2014).

In addition, flood loss models are most often constructed for specific flood types (e.g. fluvial flood, flash flood, coastal flood) and will usually be poorly suited to estimate loss due to flood events with other dominant damaging processes (Kreibich and Dimitrova, 2010; Kreibich and Thieken, 2008). Models also vary in the way loss is expressed, which can be either in monetary terms or as a fraction of the value of the element at risk (Messner et al., 2007). These are referred to respectively as absolute and relative flood loss models, the latter being better suited than the former for application across different study cases (Krzysztofowicz and Davis, 1983). Further differences may exist in terms of other model attributes.

Due to this large heterogeneity, it is difficult to identify flood loss models that, given their attributes, are potentially the most appropriate for application in specific risk assessment studies. Ideally, for any given application setting, a perfectly suited model (e.g. similar type of asset, no spatial transfer required, validated with local evidence) would be available and unambiguously identifiable, but unfortunately this is far from the case. The lack of an established procedure to select suitable flood loss models from the many available in the literature means that model selection is often done rather arbitrarily (Scorzini and Frank, 2015), which can negatively impact the quality of flood loss estimations and lead to suboptimal investment decisions based on model outcomes (Wagenaar et al., 2016).

A critical issue in flood loss modelling is uncertainty (Merz et al., 2004), which is usually high and can significantly contribute to overall uncertainty in flood risk analyses (de Moel and Aerts, 2011). Model uncertainty is mainly related with parameter representation, whereby fewer parameters than those theoretically needed to describe physical damage processes are used, and with insufficient data and/or knowledge about damage processes (Wagenaar et al., 2016). Quantifying uncertainty is imperative, as this information is required to make informed decisions in the context of flood risk management (Downton et al., 2005; Peterman and Anderson, 1999; USACE, 1992). However, the vast majority of flood loss models currently available in the literature are deterministic (Gerl et al., 2016), providing single point estimates of loss. Such estimates are unable to meet the decision needs of different stakeholders, who may have differing risk attitudes or cost-benefit ratios for risk mitigation measures (Merz and Thieken, 2009). Moreover, the uncertain nature of flood loss estimations means that the performance of any given deterministic model that appears appropriate for a certain application can be limited, as large disparities may exist even among seemingly comparable models (Jongman et al., 2012; Merz and Thieken, 2009). This makes flood risk estimates highly sensitive to loss model selection (Apel et al., 2009; Wagenaar et al., 2016). It is thus clear that adopting a single deterministic model for the estimation of flood losses is not recommended, as the information it provides is insufficient for optimal decision-making, and the results will potentially, and very likely, be inaccurate. Even though research on flood loss modelling has recently started to move into the probabilistic domain (Custer and Nishijima, 2015; Dottori et al., 2016; Kreibich et al., 2017; Schröter et al., 2014; Vogel et al., 2012), probabilistic models are still scarce.

Multi-model ensembles have been successfully applied in scientific fields such as hydrology or weather forecasting to tackle similar issues to those discussed above. Ensemble means have been shown to almost always outperform individual models (Georgakakos et al., 2004; Gleckler et al., 2008; Reichler and Kim, 2008), and the combination of the output of different models can be a pragmatic approach to estimate model uncertainty (Palmer et al., 2004; Weigel et al., 2008). However, in the context of vulnerability modelling, the concept of combining multiple models is relatively new. Rossetto et al. (2014) and Spillatura et al. (2014) have proposed the use of mean model estimates as part of their studies on respectively fragility and vulnerability curves for seismic risk assessment, but model performance is not evaluated and uncertainty quantification is not discussed. The potential use of multi-model ensembles in flood vulnerability assessment has not been addressed before.

This study therefore aims to answer the following research questions:

1. Can multi-model ensembles be used to improve the accuracy of flood loss estimations?

2. Are multi-model ensembles able to represent model uncertainty and provide reliable probabilistic estimates of flood loss?

3. How should such ensembles be constructed?

We first propose a framework to rate flood loss models according to their potential skill and suitability as participating members in such ensembles. We then construct various multi-model ensembles, based both on the rating framework and on a state of simulated non-informativeness, differing in terms of participating members, ensemble size, and weighting criteria, and evaluate their performance. Twenty

flood loss models available in the literature are adopted, and losses are modelled for residential buildings in two application cases, corresponding to flood events that took place in Germany in 2002 and in Italy in 2010. Based on the results, which are shown and discussed in Sect. 3, conclusions are drawn regarding the application of multi-model ensembles in flood loss estimations.

2 Setup of validation exercise

2.1 Flood loss models

The flood loss model catalogue developed by Gerl et al. (2016) was used as the basis for model selection in this study. We first identified all deterministic models describing loss to residential buildings, and then excluded models based on following criteria:

- the documentation is insufficient for model implementation;
- the model uses explanatory variables that are not available in most practical applications;
- the model has a functional form that is considered inappropriate (e.g. too simplistic or discretized);
- the model is based on the same dataset as another model deemed more appropriate for the application settings (this is to ensure model independence and avoid potential biases in the resulting ensembles).

Based on this procedure, 20 deterministic flood loss models for residential buildings were adopted. The catalogue developed by Gerl et al. (2016) provides information on the properties of each model, which is necessary to assess model suitability according to the framework proposed in Sect. 3.1. Table 1 shows the model properties relevant for this study, as well as the corresponding references, where model formulations can be consulted.

Each model is implemented to compute flood losses for the two application cases described in Sect. 2.3, for which the available hazard and exposure data are shown in Table 2. This consists in the largest application to date of different flood loss models within the scope of a scientific study on flood risk. In the estimation of losses for each asset, the best-matching function from each model is selected. In cases where this cannot be done unambiguously (e.g. due to mismatch in asset description between the exposure dataset and the model documentation), the selection is based on expert judgement. When models do not use some of the available hazard or exposure data, the unused variables are not considered. Losses given in absolute terms are adjusted for inflation. The modelled losses are provided as supplementary material.

2.2 Evaluation methods

2.2.1 Deterministic predictions

The predictive performance of single loss models and ensemble means is evaluated in terms of accuracy and systematic bias, using respectively the root mean squared error (RMSE) and the mean bias error (MBE). These are given by

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\boldsymbol{X}}_i - \boldsymbol{X}_i\right)^2} \tag{1}$$

and

$$\mathrm{MBE} = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\boldsymbol{X}}_i - \boldsymbol{X}_i\right), \tag{2}$$

where $\hat{\boldsymbol{X}}$ is a vector of n predictions and $\boldsymbol{X}$ is the vector of observed values of flood loss.

2.2.2 Ensemble predictions

The probabilistic skill of ensembles is evaluated using the continuous ranked probability score (CRPS), which is defined as the integrated squared difference between the cumulative distributions of predictions and observations (Weigel, 2012). We adopt the expression for the CRPS derived by Hersbach (2000), which is described as follows. Consider a set of n elements affected by a flood with corresponding observed losses $x_1, \ldots, x_n$. Let there be m ensemble members, and let $\hat{x}_{t,i}$ be the prediction of loss given by ith ensemble member for the tth element, sorted in ascending order. Define $\hat{x}_{t,0} = -\infty$ and $\hat{x}_{t,m+1} = +\infty$. The CRPS is given by

$$\mathrm{CRPS} = \frac{1}{n}\sum_{t=1}^{n}\left[\sum_{i=1}^{m}\alpha_{t,i}\left(\frac{i}{m}\right)^2 + \sum_{i=0}^{m-1}\beta_{t,i}\left(1-\frac{i}{m}\right)^2\right], \tag{3}$$

where

$$\alpha_{t,i} = \begin{cases} 0 & \text{if } x_t \le \hat{x}_{t,i} \\ x_t - \hat{x}_{t,i} & \text{if } \hat{x}_{t,i} < x_t \le \hat{x}_{t,i+1} \\ \hat{x}_{t,i+1} - \hat{x}_{t,i} & \text{if } \hat{x}_{t,i+1} < x_t \end{cases}$$

and

$$\beta_{t,i} = \begin{cases} \hat{x}_{t,i+1} - \hat{x}_{t,i} & \text{if } x_t \le \hat{x}_{t,i} \\ \hat{x}_{t,i+1} - x_t & \text{if } \hat{x}_{t,i} < x_t \le \hat{x}_{t,i+1} \\ 0 & \text{if } \hat{x}_{t,i+1} < x_t \end{cases}.$$

The CRPS can be interpreted as an error measure, with lower values corresponding to higher probabilistic skill.

To assess ensemble reliability (i.e. whether ensemble predictions and observations are statistically indistinguishable), the rank histogram is adopted, which is constructed as

Table 1. Models included in this study, including some of their properties.

Name	Hazard variables[a]	Exposure variables[b]	Country	Region/catchment	Flood type	Damage metric	Reference
ANUFlood	wd	fa	Australia	–	fluvial	absolute	Department of Natural Resources and Mines (2002)
Budiyono	wd	bt	Indonesia	Ciliwung River	fluvial	relative	Budiyono et al. (2015)
DSM	wd	bt	the Netherlands	–	fluvial, coastal	relative	Klijn et al. (2007)
Dutta	wd	str	Japan	Ichinomiya River basin, Chiba prefecture	fluvial	relative	Dutta et al. (2003)
FLEMO	wd, con, rp	bt, bq, pre	Germany	Elbe, Danube	fluvial	relative	Elmer et al. (2010)
HAZUS-MH	wd	bt, nf, bas	USA	–	fluvial, coastal	relative	Scawthorn et al. (2006)
HOWAS	wd	bt, bas	Germany	–	fluvial	absolute	Buck and Merkel (1999)
HWS-GIS	wd	–	Germany	Lippe	fluvial	relative	Hydrotec (2002)
ICPR	wd	–	Switzerland, Germany, France, Netherlands	Rhine	fluvial	relative	ICPR (2001)
IKSE	wd	–	Germany	Elbe	fluvial	relative	IKSE (2003)
Luino	wd	–	Italy	Boesio basin, in the Lombardy Region	fluvial	relative	Luino et al. (2009)
MCM	wd, id	bt	England, Wales	–	fluvial, coastal	absolute	Penning-Rowsell et al. (2005)
MERK	wd	nf, bas	Germany	Coast of Schleswig-Holstein	coastal	relative	Reese et al. (2003)
Pistrika and Jonkman	wd, fv	–	USA	Mississippi River	fluvial, levee breach	relative	Pistrika and Jonkman (2010)
Riha and Marcikova	wd, id	bt, oth	Czech Republic	–	fluvial	relative	Riha and Marcikova (2009)
Toth	wd	bt, str, nf	Hungary	Körös corner flood area	fluvial	relative	Tóth et al. (2008)
TYROL	wd	–	Austria	Tyrol	fluvial	absolute	Huttenlau et al. (2010)
Vanneuville	wd	bt	Belgium	–	fluvial	relative	Vanneuville et al. (2006)
Vojinovic	wd	fa	St Maarten	–	fluvial	absolute	Vojinovic et al. (2008)
Yazdi and Neyshabouri	wd	–	Iran	Kan basin	fluvial	relative	Yazdi and Neyshabouri (2012)

[a] Hazard variables: wd: water depth; fv: flow velocity; id: inundation duration; con: contamination; rp: return period. [b] Exposure variables: bt: building type; str: building structure; bq: building quality; nf: number of floors; bas: presence of basement; fa: floor area; pre: precautionary measures.

follows. Consider an m-member ensemble prediction $\hat{x} = (\hat{x}_1, \ldots, \hat{x}_m)$ and a corresponding observation x. The rank of x in relation to the ensemble members of $\hat{x}$ is given by $r = M + 1$, where M is the number of ensemble members that x exceeds ($M \leq m$). For example, if x is smaller than all ensemble members, the observation has rank $r = 1$, while if x exceeds all ensemble members, then $r = m + 1$. If an ensemble is reliable, for a set of n prediction-observation pairs there should be $n/(m+1)$ observations with each $m+1$ possible rank values, i.e. the histogram should be flat. Systematic deviations from flatness can indicate deficiencies in terms of ensemble dispersion and bias. Note that no ensemble is perfectly reliable, and random deviations from flatness are expected due to sampling uncertainty (Talagrand et al., 1998; Weigel, 2012).

2.3 Application cases

2.3.1 2002 flood along the Mulde River, Germany

Floods are a recurring natural hazard in the Mulde catchment (7400 km^2) located in Saxony, Germany. In recent years, this area has been severely affected by the June 2013 and August 2002 floods (Engel, 2004; Merz et al., 2014). The latter was triggered by record-breaking precipitation amounts in the Ore Mountains, which form the headwaters of the Mulde River. At the Zinnwald-Georgenfeld station, operated by the German Weather Service, 312 mm of rainfall were recorded within 24 h (Ulbrich et al., 2003). The flood caused many dike breaches and resulted in considerable loss in 19 municipalities in the German state of Saxony along the Mulde (Fig. 1).

The data used for this application case are listed in Table 2, and the results of individual model applications in terms of error statistics are shown in Table 3. The flood extension and water depths were estimated through hydronumeric simulations (Apel et al., 2009) and hydraulic trans-

Table 2. Input variables for the Mulde and Caldogno application cases.

		Resolution	
Component	Variable	Mulde	Caldogno
Hazard	Water depth (m)	10 m × 10 m grid cell	5 m × 5 m grid cell
	Flow velocity ([a])	Municipality	5 m × 5 m grid cell
	Inundation duration (h)	Municipality	–
	Return period (yr)	Catchment	–
	Contamination indicator	Municipality	–
Exposure	Building floor area (m^2)	Municipality [b]	Building
	Value (EUR)	10 m × 10 m grid cell	Building
	Building type	Municipality	Building
	Building quality	Municipality	Building
	Building structure	–	Building
	Number of floors	–	Building
	Presence of basement	–	Building
	Year of construction	–	Building
	Precautionary measures indicator	Municipality	–
Loss	Reported loss (EUR)	Municipality	Building

[a] indicator for Mulde; $m\,s^{-1}$ for Caldogno. [b] Mean value.

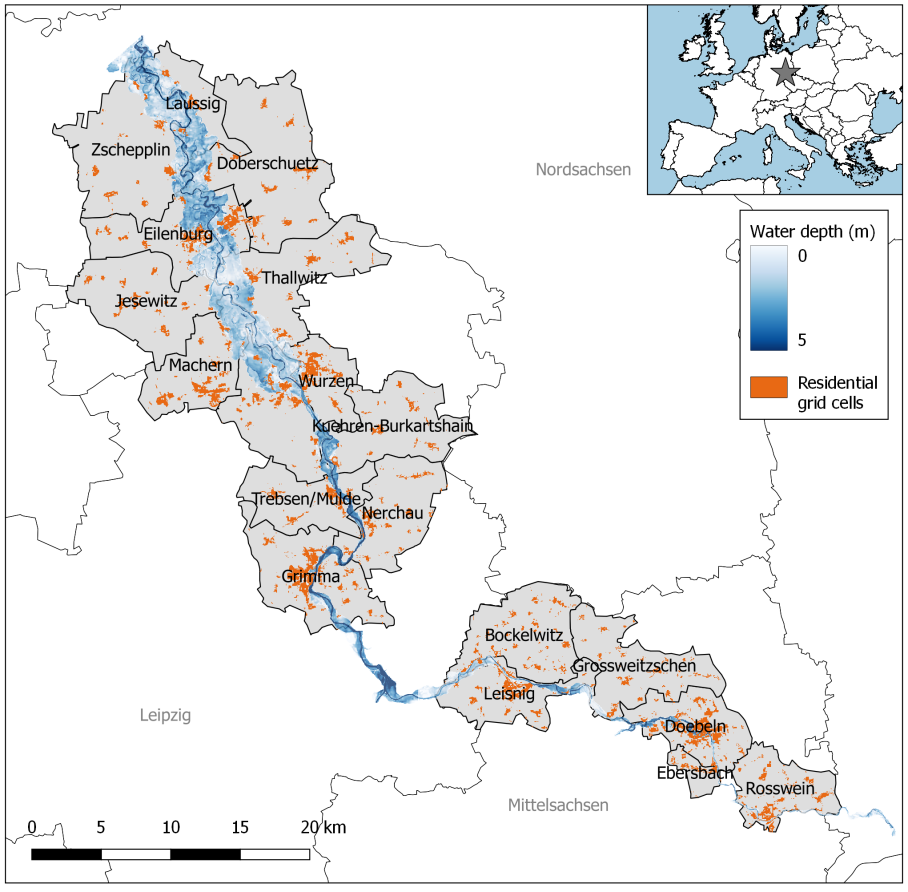

Figure 1. 2002 flood along the Mulde River, in Germany. The figure shows the municipalities considered in the case study (grey), the estimated flood extension and water depths (blue), and the location of the residential grid cells (orange).

Table 3. Results of individual model applications in the Mulde case: root mean square error (RMSE) and mean bias error (MBE), sorted by RMSE.

Model name	Error metrics (EUR million)	
	RMSE	MBE
Luino	8.143	−1.230
IKSE	9.160	−2.433
Dutta	9.177	1.870
DSM	9.469	1.359
FLEMO	10.918	−3.850
HAZUS-MH	10.964	3.998
Riha and Marcikova	11.449	2.986
Vanneuville	13.608	−5.302
Toth	13.906	−6.050
MCM	14.405	−4.266
HWS-GIS	15.796	−7.237
ICPR	15.888	−7.201
MERK	16.497	−7.656
Pistrika and Jonkman	16.883	8.235
Yazdi and Neyshabouri	17.174	7.398
Budiyono	18.258	8.190
Vojinovic	19.095	−8.667
HOWAS	20.982	−9.863
TYROL	21.160	−9.979
ANUFlood	21.559	−10.273

formation (Grabbert, 2006). Return periods of flood peak discharges were derived from annual maximum series of mean daily discharges by Elmer et al. (2010). For the estimation of contamination indicators, inundation durations, flow velocity indicators and precautionary measures indicators, computer aided telephone interviews with affected households have been used (Thieken et al., 2005). The average floor areas of residential buildings and average building values are based on official statistical data about total living area for different types of residential buildings per district, and standard construction costs per square metre gross floor area (Kleist et al., 2006). Asset values with a spatial resolution corresponding to the inundation map (i.e. 10 m × 10 m) have been derived by applying a binary disaggregation method and using the digital basic landscape model ATKIS as ancillary information (Wünsch et al., 2009). Residential building type composition and mean residential building quality per municipality were derived by Thieken et al. (2008) using geomarketing data from INFAS GEOdaten GmbH from 2001. Flood losses to residential buildings have been documented by the Saxon Relief Bank on the municipality level (Saxon Relief Bank, personal communication, 2005) and amount to a total of EUR 240.6 million. For more details, see Kreibich et al. (2017).

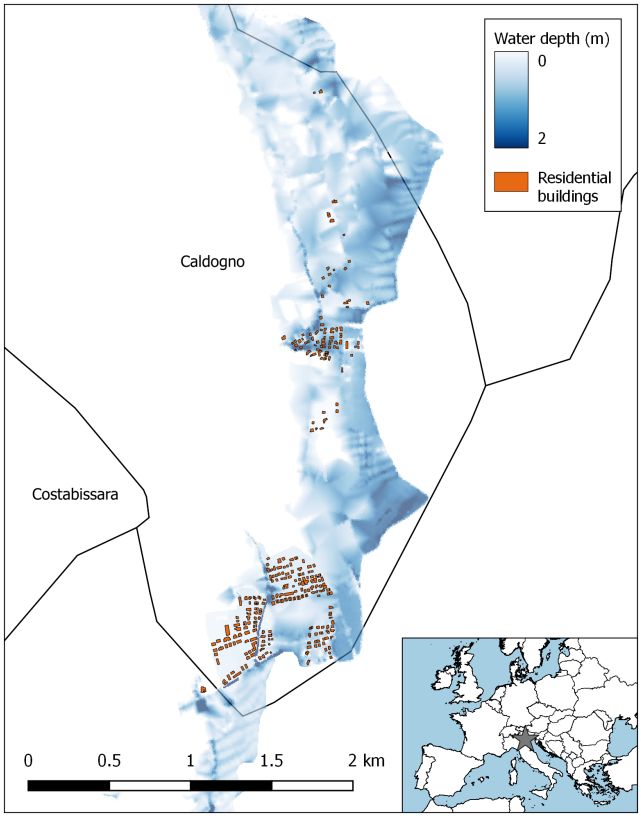

Figure 2. 2010 Bacchiglione river flood in Caldogno, Italy. The figure shows the estimated flood extension and water depths (blue), and the location of the residential buildings considered in the study (orange).

2.3.2 2010 flood in Caldogno, Italy

From 31 October to 2 November 2010, the Veneto Region was affected by persistent rain, particularly in the prealpine and foothill areas, with accumulated rainfall exceeding 500 mm in some locations (Regione del Veneto, 2011a). This caused multiple rivers to overflow, resulting in floods that inundated an area of 140 km^2 and had a considerable human and economic impact. Three people lost their lives and 3500 had to evacuate their homes. Flood losses to residential, commercial and public assets were estimated to be EUR 426 million. Caldogno, a municipality with a population of about 11 000 located in the province of Vicenza, was among the most affected, with reported losses to those sectors reaching EUR 25.7 million (Regione del Veneto, 2011b). In this study, we adopt it as the second application case (Fig. 2).

The data used for this application case are listed in Table 2, and the results of individual model applications in terms of error statistics are shown in Table 4. The inundation characteristics were estimated using a coupled 1-D/2-D model of the study area between the municipalities of Cal-

Table 4. Results of individual model applications in the Caldogno case: root mean square error (RMSE) and mean bias error (MBE), sorted by RMSE.

Model name	Error metrics (EUR)	
	RMSE	MBE
IKSE	28 324.2	3742.0
Toth	28 381.9	−6154.0
HWS-GIS	28 901.2	−7974.5
FLEMO	29 147.5	2899.9
DSM	29 950.1	8437.0
Riha and Marcikova	30 248.8	−12 084.7
MCM	30 798.3	−11 106.9
HAZUS-MH	30 829.7	13 131.2
Luino	31 050.4	12 688.3
Dutta	31 242.9	11 470.0
MERK	32 078.9	−16 228.1
Vojinovic	32 605.8	−15 581.1
TYROL	33 798.5	−17 867.9
ANUFlood	34 010.5	−18 510.8
Vanneuville	34 925.9	−20 809.0
HOWAS	34 954.3	−19 213.4
ICPR	35 356.4	−21 224.3
Yazdi and Neyshabouri	40 614.3	25 441.6
Budiyono	43 112.8	6602.7
Pistrika and Jonkman	109 444.7	101 296.1

dogno and Vicenza, and validated using data from sources such as aerial surveys and interviews with the local population. Building areas were derived from the cadastral map issued by the Veneto region. Building properties (i.e. building type, structural type, quality, number of floors, and year of construction) were assessed through direct surveys to each damaged building. Building values were estimated based on data from the Chamber of Commerce of Vicenza. Losses to residential buildings were provided by the municipality of Caldogno and amount to a total of EUR 7.55 million. These correspond to actual restoration costs that were collected and verified within the scope of the loss compensation process by the state. Further details can be found in Scorzini and Frank (2015).

3 Ensemble construction and evaluation

Ensembles are finite sets of deterministic realisations of a random variable, whereby the prediction given by each ensemble member is assumed to represent an independent sample from an underlying true probability distribution (Hamill and Colucci, 1997). Ensembles can be used to account for various sources of uncertainty in physical processes, namely initial conditions, parameter, and model uncertainty. The latter can be achieved by combining the output of different models to create a so-called multi-model ensemble (Weigel, 2012). In this section, we investigate how best to translate this concept to the field of flood loss modelling, and to which extent multi-model ensembles can improve the skill and usefulness of flood loss estimations.

3.1 Model rating

3.1.1 Method

The first challenge in constructing a multi-model ensemble to estimate flood loss for a certain future application is identifying models that are better suited to be participating members. One of the requirements for the construction of successful multi-model ensembles is that participating models are skilful; if a model is consistently worse than the others in terms of prediction quality, it should not be included (Hagedorn et al., 2005). Unfortunately, testing the level of skill of a model in predicting loss, for a certain type of asset and application setting, is often not possible. Such exercise would involve applying each candidate model to estimate loss for a past flood event with similar characteristics, and quantifying its performance based on past loss observations for the same assets. However, data required to perform such assessments are usually not available, as scarcity of data is still a major problem in the field of flood risk (Merz et al., 2010). Moreover, exposure and vulnerability tend to change over time, which is likely to affect loss estimates (Tanoue et al., 2016). Another issue of a more practical nature is that collecting, implementing and comparing flood loss models is laborious and time consuming. Because of the economic constraints that inevitably exist in any practical application, most users will likely have limited time to invest in that task. This becomes more problematic as the already large number of models available in the literature continues to increase.

A more practicable approach is to evaluate the suitability and potential performance of each model in estimating loss, for a given application setting, based on its properties. This is advantageous, as it does not require that each model be tested explicitly, and can instead be achieved by making use the information contained in a model metadata catalogue such as the ones developed by Gerl et al. (2016) or Pregnolato et al. (2015). However, models differ at various levels, and a model that is potentially superior regarding some of its properties may be inferior in terms of others (see Sect. 1). Consequently, directly evaluating the potential performance of flood loss models is arduous, and currently no established procedure exists to this end. In this subsection, we address this gap by proposing a framework to rate a set of flood loss models based on their properties. The framework is described as follows:

1. a probability tree of model properties is set up through expert elicitation. A set of N independent properties that characterize flood loss models and that are likely to be informative for model performance are identified (e.g. damage metric). For each property (i.e. tree node) n, a set of mutually exclusive and collectively exhaus-

tive categories are defined (e.g. relative and absolute). A subjective probability is then assigned to each category, corresponding to the degree of belief that a model that falls into that category will offer higher predictive performance than if it did in others. It follows that for each property n, the probabilities of the different categories sum to 1. Each path of the tree will have an associated probability that is obtained through the product of each node's probabilities p_n along the path, therefore reflecting the degree of belief that this combination of model properties is the one that should be used;

2. once the probability tree is set up, it can be used to assign scores to and rank flood loss models. Because the tree covers the entire space of possible categories within each property, all flood loss models will necessarily have a set of properties that matches one of the tree paths. Any model can thus be assigned a score that is equal to the probability of its respective path. When assigned to a certain number of models rather than to all the possible combinations of model properties, such scores no longer have a specific probabilistic meaning, nor are they intended to. Instead, the scores of different candidate models in a pool can be used to establish a relative degree of belief among them. This effectively provides users with information on their potential performance, in relation to the other models in the pool, through a structured and simple to use procedure.

3.1.2 Application

We apply this framework to the models and test cases presented in Sect. 2. We first propose a probability tree referring to flood loss models for buildings. It condenses expert knowledge and current state of the art in flood vulnerability of buildings, as well as experience from previous model transfer studies. The selection of properties and categories aims to balance comprehensiveness, objectivity and simplicity. Figure 3 presents the different properties, a succinct justification of their potential relevance in assessing model performance, and the respective categories and assigned subjective probabilities. Note that the maximum partial score that can be assigned to a model for properties 1 and 2 (shown in Fig. 3) depends not only on the model but also on the hazard and exposure data sets. For example, when in a certain application case only water depth data is available, loss models that use additional explanatory variables (e.g. velocity) should not be rated higher. We then use this setup to rate the flood loss models. The results are shown in Tables 5 and 6.

While model properties are expected to be informative for performance, they are not presumed to explain it fully. However, if model properties do have usefulness in assessing the performance of models in relation to one other, some degree of correlation between model scores and different performance metrics should exist. We evaluate this using the Spearman's rank correlation coefficient r_s respectively between the scores shown in Tables 5 and 6 and the error metrics shown in Tables 3 and 4. Results show a significant strong negative correlation between the variables ($-0.79 < r_s < -0.51$, $p < 0.01$), which suggests that model rating based on expert judgement is indeed informative for model performance. Note that no attempt was made to maximize correlations by fine-tuning the subjective probabilities, as not only would those not correspond to the experts' degrees of belief, but more importantly, because that would be no more than an exercise in overfitting to these two case studies. This topic is revisited in Sect. 3.2.2.

1. Flood intensity measures

Flood damage processes are influenced by multiple factors. Although water depth is considered the most important intensity measure, additional variables tend to improve model predictive skill.

Water depth and additional variables	0.65
Water depth only	0.35

2. Characterization of exposed assets

The degree of characterization of assets in flood loss model is directly related with potential performance, as insufficient distinction may result in their inappropriate application to assets for which they are not suited.

Building type and physical properties (e.g. material, no. of floors)	0.45
Building type only (e.g. single family house)	0.35
Occupancy type only (e.g. residential)	0.20

3. Similarity of local context with application setting

Models generally perform better in regions with a socio-economic context comparable to the one for which they were developed, as they tend to have more similarities in terms of construction quality and practices.

Same region	0.40
Same country	0.30
Same WESP classification[1]	0.20
Different WESP classification[1]	0.10

[1] According to UN World Economic Situation and Prospects 2016. A: developed economies; B: economies in transition; C: developing countries; D: least developed countries.

4. Flood type in relation to application setting

Flood loss models are usually constructed for specific types of inundation (e.g., fluvial flood), and will usually perform worse when applied to flood events with different dominant damaging processes.

Identical	0.70
Different	0.30

5. Damage metric

Relative loss models, which express loss as a fraction of the total asset value, offer better transferability, whereas absolute loss models have little applicability outside the specific case for which they were developed.

Relative	0.70
Absolute	0.30

Figure 3. Proposed set of properties (probability tree nodes) that are considered relevant to assess the performance of flood loss models for buildings, and respective categories and subjective probabilities.

Table 5. Model scores for the Mulde application case.

Model name	Node probabilities					Score (10^{-2})	Score rank
	p_1	p_2	p_3	p_4	p_5		
FLEMO	0.65	0.35	0.30	0.70	0.70	3.34	1
IKSE	0.35	0.20	0.40	0.70	0.70	1.37	2
Riha and Marcikova	0.65	0.20	0.20	0.70	0.70	1.27	3
HAZUS-MH	0.35	0.35	0.20	0.70	0.70	1.20	4
HWS-GIS	0.35	0.20	0.30	0.70	0.70	1.03	5
MCM	0.65	0.35	0.20	0.70	0.30	0.96	6
DSM	0.35	0.20	0.20	0.70	0.70	0.69	7
Dutta	0.35	0.20	0.20	0.70	0.70	0.69	7
ICPR	0.35	0.20	0.20	0.70	0.70	0.69	7
Luino	0.35	0.20	0.20	0.70	0.70	0.69	7
Toth	0.35	0.20	0.20	0.70	0.70	0.69	7
Vanneuville	0.35	0.20	0.20	0.70	0.70	0.69	7
Pistrika and Jonkman	0.65	0.20	0.20	0.30	0.70	0.55	13
HOWAS	0.35	0.20	0.30	0.70	0.30	0.44	14
MERK	0.35	0.20	0.30	0.30	0.70	0.44	14
Budiyono	0.35	0.20	0.10	0.70	0.70	0.34	16
Yazdi and Neyshabouri	0.35	0.20	0.10	0.70	0.70	0.34	16
ANUFlood	0.35	0.20	0.20	0.70	0.30	0.29	18
TYROL	0.35	0.20	0.20	0.70	0.30	0.29	18
Vojinovic	0.35	0.20	0.10	0.70	0.30	0.15	20

Table 6. Model scores for the Caldogno application case.

Model name	Node probabilities					Score (10^{-2})	Score rank
	p_1	p_2	p_3	p_4	p_5		
FLEMO	0.65	0.35	0.20	0.70	0.70	2.23	1
Riha and Marcikova	0.65	0.20	0.20	0.70	0.70	1.27	2
HAZUS-MH	0.35	0.35	0.20	0.70	0.70	1.20	3
Toth	0.35	0.35	0.20	0.70	0.70	1.20	3
Luino	0.35	0.20	0.30	0.70	0.70	1.03	5
MCM	0.65	0.35	0.20	0.70	0.30	0.96	6
DSM	0.35	0.20	0.20	0.70	0.70	0.69	7
Dutta	0.35	0.20	0.20	0.70	0.70	0.69	7
HWS-GIS	0.35	0.20	0.20	0.70	0.70	0.69	7
ICPR	0.35	0.20	0.20	0.70	0.70	0.69	7
IKSE	0.35	0.20	0.20	0.70	0.70	0.69	7
Vanneuville	0.35	0.20	0.20	0.70	0.70	0.69	7
Pistrika and Jonkman	0.65	0.20	0.20	0.30	0.70	0.55	13
MERK	0.35	0.35	0.20	0.30	0.70	0.51	14
Budiyono	0.35	0.20	0.10	0.70	0.70	0.34	15
Yazdi and Neyshabouri	0.35	0.20	0.10	0.70	0.70	0.34	15
ANUFlood	0.35	0.20	0.20	0.70	0.30	0.29	17
HOWAS	0.35	0.20	0.20	0.70	0.30	0.29	17
TYROL	0.35	0.20	0.20	0.70	0.30	0.29	17
Vojinovic	0.35	0.20	0.10	0.70	0.30	0.15	20

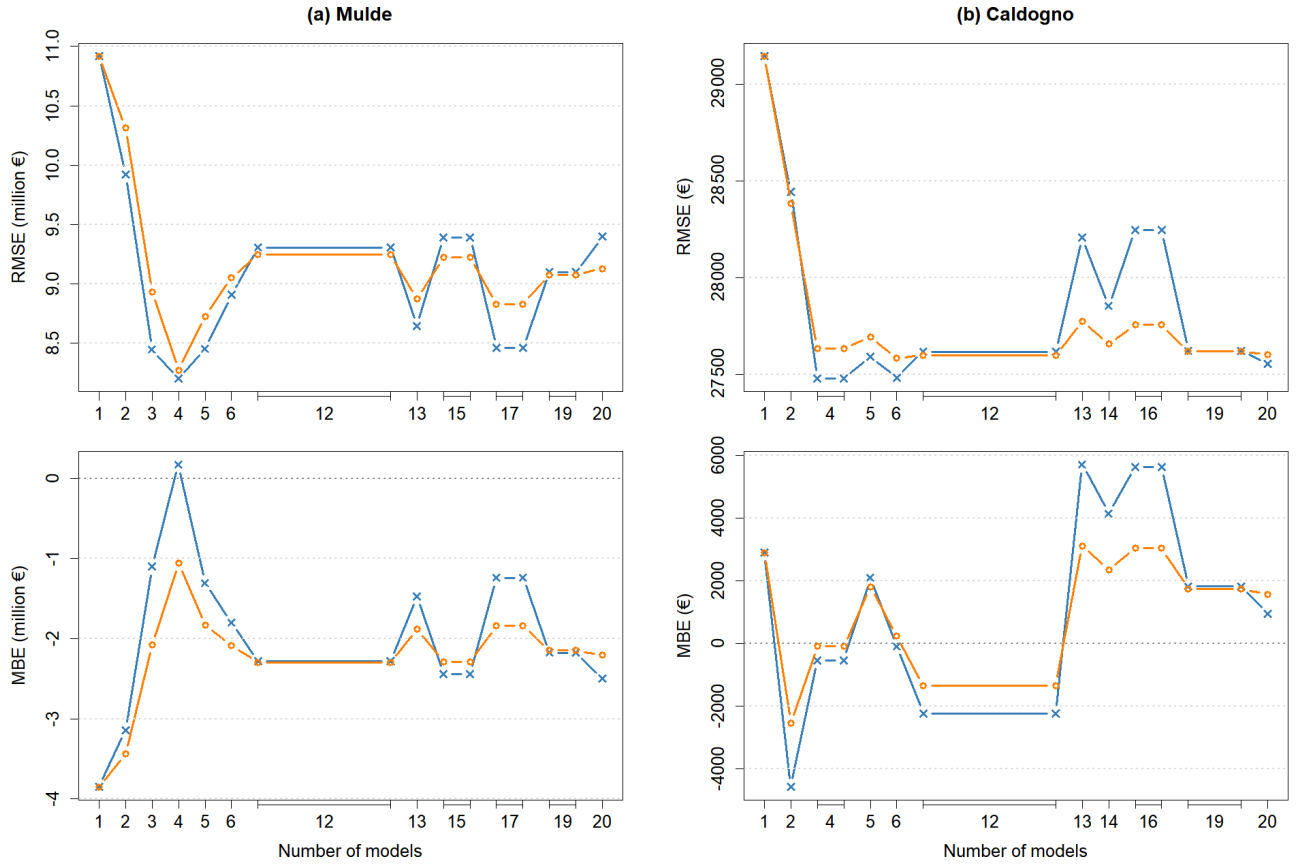

Figure 4. Root mean square error (RMSE) and mean bias error (MBE) of the means of ensembles of increasing size, with models included sequentially from highest to lowest score, starting with the highest ranked single model. Blue crosses and orange circles refer to ensembles weighted equally and differently, respectively.

3.2 Ensemble-mean performance

The objective of the analyses presented in this section is twofold: to assess to which extent ensemble-means are able to improve skill in the estimation of flood losses, and to investigate how such ensembles should be constructed. Regarding the latter, two questions require particular attention: firstly, which and how many models to include as participating members, and secondly, how to weight those members. Both the ensemble size and the model weighting scheme are likely to have an effect on skill.

3.2.1 Based on model rating

In this exercise, the models and application cases described in Sect. 2 are used. For the construction of the various multi-model ensembles, we mimic the most common practical situation whereby it is necessary to estimate losses for a certain scenario for which past observational data is not available. Because in such situation, the skill of the individual models is not known, the potential suitability of each model for inclusion in a multi-model ensemble is evaluated through their properties, following the framework proposed in Sect. 3.1. Accordingly, different ensembles with increasing number of members are built, by including models sequentially from highest to lowest scores, according to Tables 5 and 6. Models with the same score are added to the ensemble simultaneously. The ensembles of different sizes constructed for each case study are shown in the x axes of Fig. 4, where 1 refers to the highest-ranked single model.

Losses given by ensemble means are estimated using two approaches: firstly, by assigning equal weights to all models, and secondly, by weighting them differently. Concerning this point, we now present some considerations. In the construction of an equal-weighted multi-model ensemble, the underlying hypothesis is that each model is independent and equally skilful, whereas this condition is most often not satisfied. For this reason, adopting different weights may increase the quality of multi-model predictions. However, finding optimal weights is not straightforward, and previous studies show that weighting models differently may result in different outcomes ranging from slight increases to degradation in performance (Doblas-Reyes et al., 2005; Hagedorn et al., 2005; Knutti et al., 2010). Here, we aim to assess how weights affect ensemble-mean performances in estimat-

ing flood loss, again by reproducing a practical situation where the skill of models in a certain future application is not known. Therefore, assigned weights instead reflect the user's confidence in each model (Marzocchi et al., 2015). Because the framework proposed in Sect. 3.1 provides scores that are proportional to relative degrees of belief among models, in principle they may be used as weights. This is achieved by normalizing the weights of the participating models in each ensemble so that they sum to 1 (Spillatura, 2014). As mentioned in Sect. 2.1, in this study we aimed to ensure model independence by selecting a set of models developed independently, by different authors, using non-overlapping datasets (Cotton et al., 2006; Palmer et al., 2004). We therefore assume that possible model dependences are not relevant and have no bearing on the weighting scheme. Section 3.2.2 further discusses the effect of model weighting on ensemble-based loss estimation.

Ensemble-mean performances are calculated in terms of RMSE and MBE, which are shown in Fig. 4 for the ensembles of different sizes – starting with a single model, the highest ranked for each case – and using the two weighting schemes described above. A number of observations can be made from this figure. Firstly, multi-model ensembles of any size, built by adding models with the highest degrees of belief first, considerably outperform the highest ranked single model in terms of both RMSE and MBE. This is observed for both application cases, the only exception being the MBE of some ensembles in the Caldogno case. Secondly, the performances obtained using the two different weighting approaches is mixed; while in some cases there is improvement by weighting ensemble members differently, in others the opposite is observed. The weighting approach generally does not have a significant impact on error metrics, especially when compared to the model selection. Thirdly, in both cases, the largest improvements in ensemble-mean performances are obtained after the first few highest ranked models are added. In relative terms, the impact of including additional models after that is lower. For example, in the Mulde and Caldogno case studies, the best performances are obtained with ensembles using respectively the highest-scoring four and six models. From a practical point of view, this is a particularly interesting finding because, as mentioned previously, it may not be feasible to implement a large number of models, and users may therefore be interested in parsimonious ensembles with the least number of models that lead to high predictive skill. However, in terms of probabilistic estimates of loss, smaller ensembles are less useful, which also needs to be taken into account when deciding on which ensemble size to use, as further discussed in Sect. 3.3.

Note that from here on, the equal-weighted expert-based multi-model ensembles shown in Fig. 4 will be used as a basis for other analyses and further discussion, and for the sake of brevity will be referred to as EEM-ensembles.

Some of the above observations draw comparisons between multi-model ensembles and individual models, for which the highest ranked single model is used as reference. Even though that model may not necessarily correspond to the highest performing model (which it does not in either of the application cases used here; see Tables 3, 4 and 5, 6), in a practical application case, users have no way of knowing which model is the "best". The above results very clearly demonstrate that in such situation, using a multi-model ensemble is preferable. However, it is also insightful to assess how the constructed multi-model ensembles perform in relation to the other single models. Therefore, in Fig. 5, the error metrics of the predictions given by EEM-ensemble-means and single models are presented, showing that the former consistently outperform the latter. Note that ensembles are not expected to outperform every single model in every possible situation, and it is possible that in some application cases, certain models have such high accuracy that combining them with other models results in lower performances. The problem is that it is usually not possible to identify such models beforehand. For example, in the Mulde case, the Luino model slightly outperforms the constructed ensembles in terms of RMSE. This model consists in a simple stage-damage function that refers to a single building type, and was derived from data relative to a flood in Italy. Therefore, it is not expectable that it would consistently perform as well if applied to other analogous case studies. Overall, better performances should be obtained by using multi-model ensembles (Hagedorn et al., 2005).

3.2.2 Based on simulated non-informativeness

The framework proposed in Sect. 3.1 and the subjective probabilities proposed in Fig. 3 provide a basis for model selection and weighting in the development of multi-model ensembles. In Sect. 3.2.1, we constructed various ensembles using this approach and evaluated their performance in estimating loss. However, in principle, it is possible that multi-model ensembles developed differently, i.e. by selecting different models and/or assigning different weights, would have higher skill. To investigate this issue, we simulate a so-called state of non-informativeness in terms of model suitability. This consists in assuming we have no knowledge about how particular model characteristics might affect model predictive performance (Scherbaum and Kuehn, 2011), and therefore have no way of rating models. Accordingly, we implement a probabilistic sampling procedure that, for a large number of realisations, randomly generates weights for individual models regardless of their properties. On this basis, model ensembles are built and their predictive performance is calculated for the Mulde and the Caldogno case studies. The weight generation follows the stick-breaking method, whereby models are first randomly ordered and then assigned weights sequentially. For each model, the weight is drawn from a continuous uniform distribution with a minimum value of 0 and a maximum value of 1 minus the sum of weights that have already been assigned. This approach,

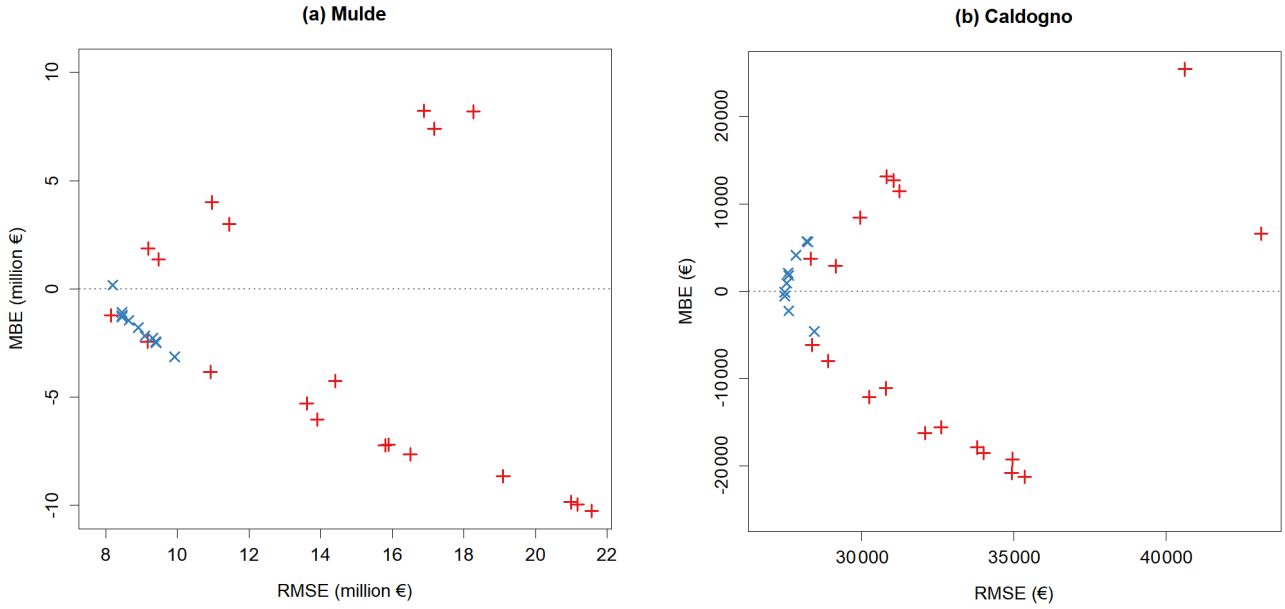

Figure 5. RMSE and MBE of the EEM-ensemble means, represented by blue crosses, and single model predictions, by red plus signs.

based on a large number of realisations, aims to cover all possible ensembles that can be constructed using the 20 flood loss models from Table 1, using not only different weighting approaches (i.e. ensemble members weighted both equally and differently) but also different combinations of models. The latter is because according to the stick-breaking method, once the model weights sum to 1, all other models receive a weight of 0 and are thus not included in the ensemble.

Scatter plots of the RMSE and MBE that result from the above procedure are presented in Fig. 6 for both case studies. The same error metrics regarding the EEM-ensembles and the single models are also included. The plots show that a wide range of possible outcomes in terms of RMSE and MBE exist when random weights are assigned to models within the framework of a state of non-informativeness. While the lower bounds of the resulting convex hull are defined by the error metrics of the lowest-performing models, the upper bounds (i.e. highest performances) are given not by any single model, but instead by multi-model ensembles, as expected. In this regard, it is clear that the model rating framework based on expert judgement and subjective probabilities proposed in Sect. 3.1 add value to the ensemble development process. Indeed, ensembles that are constructed by adding models prioritized in terms of potential suitability (shown in Fig. 4) are among the highest performing ensembles, considering all the existing possibilities. It is interesting to highlight that the simple unweighted mean of all models also performs relatively well, which suggests that if no knowledge is available on model properties and/or on how they influence performance, it is better to include all models than to wrongly select them.

The plots also show that it is possible to create certain ensembles that lead to better skill in relation to the ones developed based on expert judgement. However, the potential relative degree of improvement is very low in both test cases, more markedly so in the Caldogno case, which reinforces the idea that the approach proposed in Sect. 3.1 provides a good basis for ensemble construction. We do not attempt to maximize the performance of the constructed multi-model ensembles based on the results obtained in this exercise, as this would be of little relevance. Analogously with the "best" model discussion in Sect. 3.2.1, in a practical application the ensembles cannot be tested beforehand. Finding specific weights that maximize performance for the Mulde and the Caldogno case studies would consist in pointless overfitting, as such weights necessarily vary from case to case. In addition, it is likely that such weights would not make sense from the perspective of an expert. Instead, the objective here is that ensembles are constructed in a manner that leads to good performances in all situations, which the results support. Finally, Fig. 6 corroborates that correctly selecting models for an ensemble is more important than weighting them. The EEM-ensembles, which result from model selection only, display error metrics close to the minimum obtainable from a wide range of possible outcomes. In comparison, further improvements that could possibly be achieved by assigning different weights to ensemble members are very small.

3.3 Probabilistic application

In Sect. 3.2, multi-model ensemble-means have been shown to provide more skilful estimates of flood losses than single models. Another motivation for the use of such ensembles

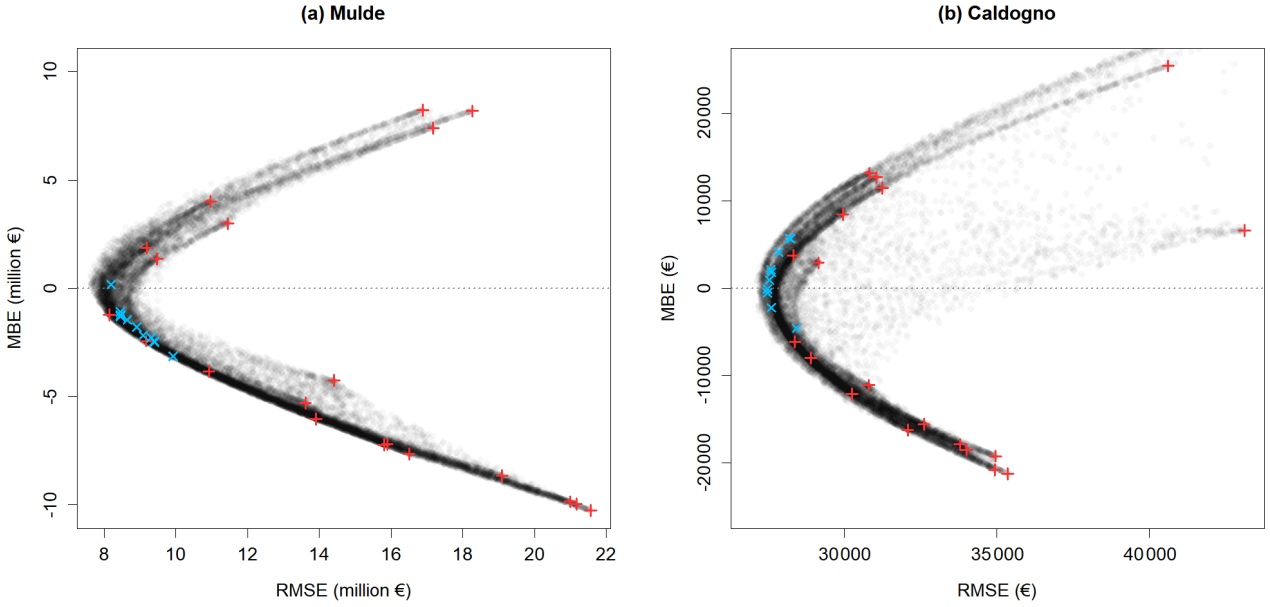

Figure 6. RMSE and MBE of 20 000 multi-model ensemble means, generated by simulating a state of non-informativeness, whereby each participating member is assigned a random weight. Blue crosses and red plus signs refer respectively to the EEM-ensemble means and the single model predictions.

is that they may be used to quantify model uncertainty and obtain probabilistic distributions of possible outcomes rather than single point estimates, which is, as discussed previously, required for optimal decision-making. In this section, we offer some discussion on this topic.

It is first necessary to make clear what the probabilistic meaning of a multi-model ensemble is. Multi-model ensembles do not directly provide probability distributions of a certain variable; instead, ensemble predictions are a priori only finite sets of deterministic realisations of that variable. The question then arises how a probability distribution can be obtained from such ensembles. The simplest approach is to adopt a frequentist interpretation of the ensembles, whereby the probability of a certain event to happen is estimated by the fraction of ensemble members predicting it. However, such approach can only produce reasonable probabilistic estimates if many ensemble members are available. Better probabilistic estimates may in principle be obtained by dressing the ensemble members with kernel functions or by fitting a suitable parametric distribution to them, provided that this is done in an appropriate manner (Weigel, 2012).

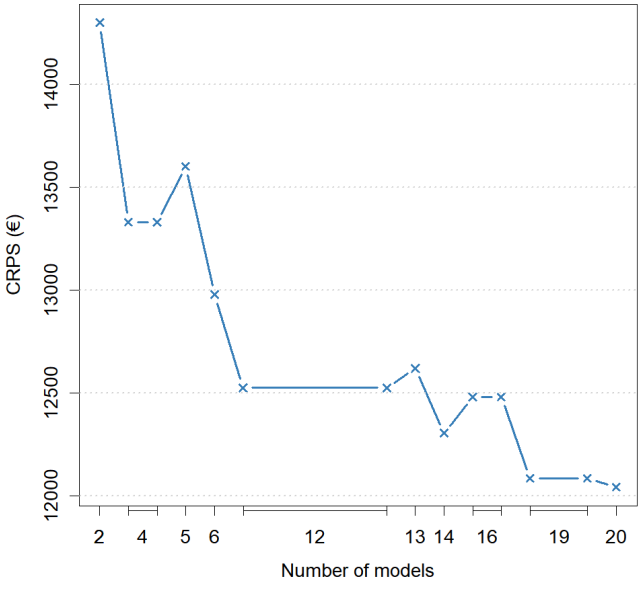

Figure 7. Continuous ranked probability score (CRPS) of the EEM-ensembles for the Caldogno application case.

3.3.1 Skill and reliability

Regardless of the method that is used to obtain probabilistic estimates from multi-model ensembles, it is first important to evaluate the "raw" ensembles, with minimum interference from the ensemble interpretation model that is used. This can be achieved using the continuous ranked probability score (CRPS) (Bröcker, 2012; Hersbach, 2000). We calculate the CRPS for the EEM-ensembles, and present the results in Fig. 7. This is done for the Caldogno case study, as the low number of data points in the Mulde case (19) are insufficient for such analysis.

The probabilistic skill of the ensembles is observed to have an increasing trend (i.e. decreasing CRPS) with the number

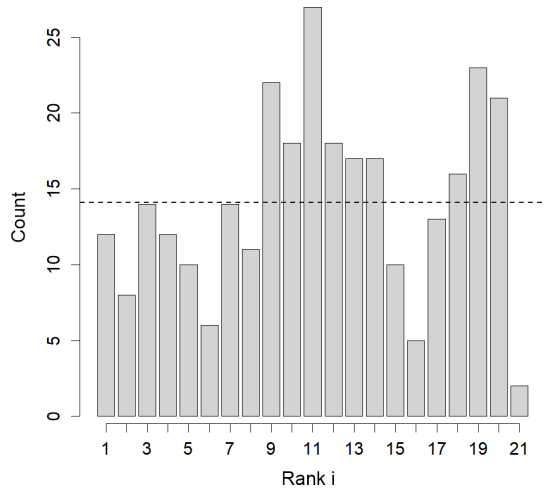

Figure 8. Rank histogram relative to the 20-model ensemble for the Caldogno application case.

of participating members. This is to some extent expected, as ensemble size is known to have an effect on probabilistic skill scores, which is explained by the fact that probabilistic estimates derived from ensembles become more unreliable as the size of the ensemble gets smaller (Weigel, 2012). This highlights the need of using a considerable number of models when the objective is to obtain reliable (i.e. statistically consistent) probabilistic estimates of flood loss. Another requirement to achieve this is that the ensemble itself is reliable, in the sense that ensemble members and observations are sampled from the same underlying probability distributions or, in other words, that they are statistically indistinguishable from each other (Leutbecher and Palmer, 2008). Even an ensemble of infinite size is unable to yield reliable probabilistic estimates if its members are not reliable (e.g. if they are heavily biased). For illustration, we assess reliability considering an ensemble comprising all 20 models implemented in this study using the rank histogram, which is shown in Fig. 8. As expected, the ensemble is not perfectly reliable; however, the counts do tend to oscillate around $\frac{n}{m+1} = \frac{296}{21}$, which suggests a reasonable degree of reliability. In addition, the ensemble appears to be slightly over-dispersive, due to an overpopulation of central ranks of the histogram.

3.3.2 Loss estimation

Finally, we illustrate the simplest approach to obtain a probabilistic distribution of flood losses using a multi-model ensemble. For each building, a value of loss is randomly generated using the reverse transform sampling method, whereby a number $u \sim [0, 1]$ is sampled from the standard uniform distribution, and the corresponding quantile is sampled from the empirical cumulative distribution function (ECDF) of losses given by ensemble members through linear interpolation. The losses for each building are then summed up, and a total loss is obtained. This process is repeated a large number of times, yielding a loss distribution for the flood event. The results for the Caldogno application, based on 10 000 realisations, are shown in Fig. 9 in the form of a histogram and ECDF of total loss.

Statistical post-processing techniques may be used to improve the reliability of probabilistic predictions. This is common practice in the field of numerical weather prediction, for example. However, in that case, relatively long time series of past observational data for a certain variable (e.g. temperature) at a certain location are usually available, and such data continue to be collected, which allows the predictive system to be calibrated and the forecasts verified. This is in contrast with the case of flood loss estimations, where loss models necessarily need to be transferred due to the rarity of the events and the difficulty in obtaining data. In the particular case of probabilistic loss estimates based on ensembles, it is therefore necessary to investigate how best to improve their reliability for future applications by considering data from previous flood events often occurring in different contexts. In addition, as mentioned previously, the reliability of probabilistic estimates may also be improved by using a more sophisticated ensemble interpretation method (i.e. kernel dressing or parametric distribution fitting). However, the most appropriate approach to do this in the case of flood loss modelling also needs to be investigated. These topics are beyond the scope of this article.

4 Conclusions

Flood loss modelling is associated with considerable uncertainty that is often neglected. In fact, most currently available flood loss models are deterministic, providing only single point estimates of loss. Users interested in performing a risk assessment will typically select one such model from the large number available in the literature, based on their perception of which one is the most suitable for the application case at hand. However, this is generally done rather arbitrarily. Moreover, the uncertain nature of flood loss estimations means that the performance of any single deterministic model may vary considerably from case to case, as large disparities in model outcomes exist even among apparently comparable models. This approach is therefore flawed at two main levels: first, flood risk estimates are highly sensitive to the selection of the flood loss model, and second, deterministic estimates of loss do not lead to optimal decision-making. In this study, we have proposed a novel approach to tackle these issues and advance the state of the art of flood loss modelling, based on the application of the concept of multi-model ensembles. This technique, which is widely used in fields such as weather forecasting, consists in combining the outcomes of different models in order to improve prediction skill and sample model uncertainty.

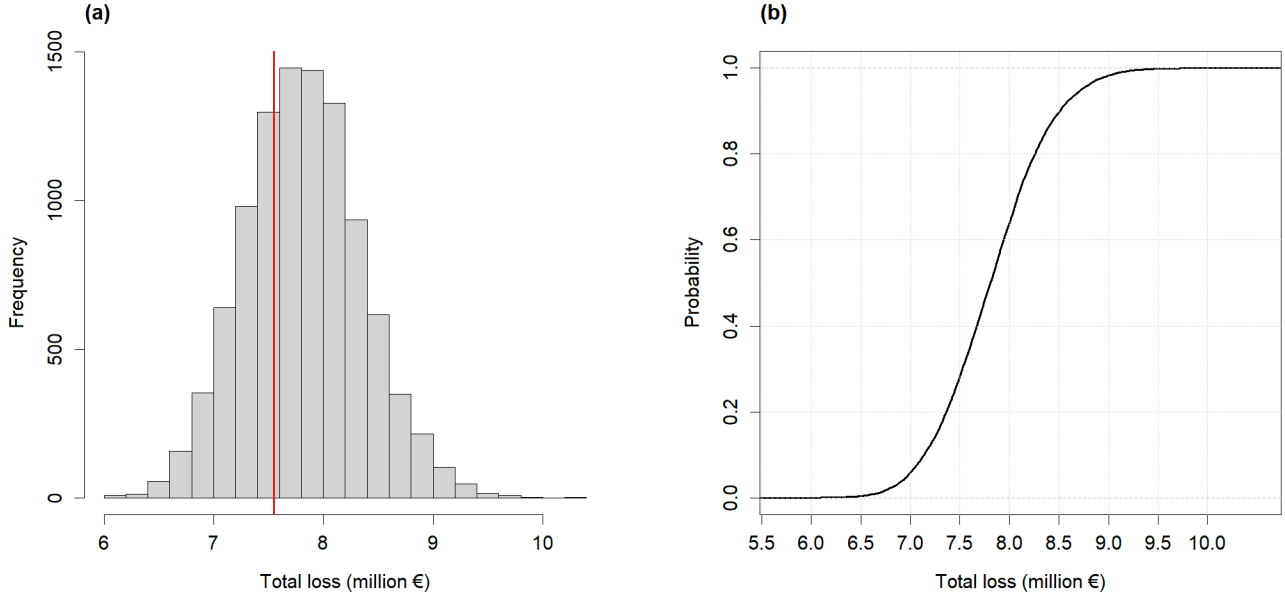

Figure 9. Probabilistic estimates of total loss, relative to model uncertainty, for the Caldogno application case, based on 10 000 realisations of loss to each building. **(a)** Histogram, with observed loss shown by the vertical red line. **(b)** Empirical cumulative distribution function (ECDF).

In order to support ensemble construction, we have first proposed a framework to assess the suitability of flood loss models to specific application cases, based on some of their main properties, through expert knowledge. This approach is advantageous as it does not require that all candidate models are implemented beforehand, which is often not achievable in practice. Based on such framework, we have proposed a scoring scheme for flood loss models for residential buildings, and applied it to the 20 models and two applications cases used in this study. The obtained model scores show significant strong negative rank correlation with error metrics, suggesting that the proposed approach is useful, and that expert judgement is informative for model performance and selection.

The constructed ensembles have been shown to considerably outperform the highest ranked single models in the estimation of flood losses. This demonstrates that in a practical application, where model performances cannot be tested beforehand, using multi-model ensembles will result in more skilful loss estimates. Ensemble-means were also tested against all single models, consistently showing higher accuracy. Equal-weighted ensembles generally displayed performances comparable to the score-weighted ones. The largest improvements in ensemble-mean performances were observed after the first few highest ranked models were added to the ensembles, which is a useful finding for practical applications, where it is not always feasible to implement a large number of models. We have also simulated a state of non-informativeness and randomly generated a large set of multi-model ensembles, representative of all possible ensembles that can be constructed using the 20 flood loss models adopted in this study. The ensembles based on expert-based scoring approach were among the most skilful, highlighting its value in the construction of multi-model ensembles. Results also suggest that model selection is more important than weighting. Further insight may be gained by testing the approach in other application cases and using a different set of flood loss models.

Larger ensembles showed higher probabilistic skill than smaller ones, which results from the increased intrinsic unreliability of ensembles as the number of participating members decreases. Therefore, if on the one hand only a limited number of models is necessary to obtain accurate mean estimates of loss, on the other hand, additional effort in model implementation is recommended when the objective is to derive a probabilistic distribution of loss that captures model uncertainty. For the Caldogno case study, we have illustrated how such a distribution can be constructed, adopting a simple equal-weighted ensemble comprising all 20 models. The results demonstrate that the use of multi-model ensembles represents a simple and pragmatic way of obtaining reliable flood loss distributions, which are more useful for decision-making than single point estimates of loss. Reliability may be further improved by calibrating the ensembles and/or adopting more sophisticated ensemble interpretation models, which warrants further research.

Data availability. The observed and modelled losses for both case studies are available in the Supplement.

The Supplement related to this article is available online at https://doi.org/10.5194/nhess-18-1297-2018-supplement.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. This research was partly supported by the European Union's Horizon 2020 research and innovation programme, through the IMPREX project (grant agreement no. 641811) and the H2020 Insurance project (grant agreement no. 730459). Further support has been received from Guy Carpenter and Company Ltd. (www.guycarp.com).

Edited by: Margreth Keiler
Reviewed by: two anonymous referees

References

Apel, H., Aronica, G. T., Kreibich, H., and Thieken, A. H.: Flood risk analyses – how detailed do we need to be?, Nat. Hazards, 49, 79–98, https://doi.org/10.1007/s11069-008-9277-8, 2009.

Bröcker, J.: Evaluating raw ensembles with the continuous ranked probability score, Q. J. Roy. Meteor. Soc., 138, 1611–1617, https://doi.org/10.1002/qj.1891, 2012.

Buck, W. and Merkel, U.: Auswertung der HOWAS-Schadendatenbank, Institut für Wasserwirtschaft und Kulturtechnik der Universität Karlsruhe, 1999.

Budiyono, Y., Aerts, J., Brinkman, J. J., Marfai, M. A., and Ward, P.: Flood risk assessment for delta mega-cities: a case study of Jakarta, Nat. Hazards, 75, 389–413, https://doi.org/10.1007/s11069-014-1327-9, 2015.

Cammerer, H., Thieken, A. H., and Lammel, J.: Adaptability and transferability of flood loss functions in residential areas, Nat. Hazards Earth Syst. Sci., 13, 3063–3081, https://doi.org/10.5194/nhess-13-3063-2013, 2013.

Cotton, F., Scherbaum, F., Bommer, J. J., and Bungum, H.: Criteria for selecting and adjusting ground-motion models for specific target regions: application to central Europe and rock sites, J. Seismol., 10, 137–156, https://doi.org/10.1007/s10950-005-9006-7, 2006.

Custer, R. and Nishijima, K.: Flood vulnerability assessment of residential buildings by explicit damage process modelling, Nat. Hazards, 78, 461–496, https://doi.org/10.1007/s11069-015-1725-7, 2015.

de Moel, H. and Aerts, J. C. J. H.: Effect of uncertainty in land use, damage models and inundation depth on flood damage estimates, Nat. Hazards, 58, 407–425, https://doi.org/10.1007/s11069-010-9675-6, 2011.

Department of Natural Resources and Mines: Guidance on the Assessment of Tangible Flood Damage, Department of Natural Resources and Mines, Queensland Government, Australia, 2002.

Doblas-Reyes, F. J., Hagedorn, R., and Palmer, T. N.: The rationale behind the success of multi model ensembles in seasonal forecasting – II. Calibration and combination, Tellus A, 57, 234–252, 2005.

Dottori, F., Figueiredo, R., Martina, M. L. V., Molinari, D., and Scorzini, A. R.: INSYDE: a synthetic, probabilistic flood damage model based on explicit cost analysis, Nat. Hazards Earth Syst. Sci., 16, 2577–2591, https://doi.org/10.5194/nhess-16-2577-2016, 2016.

Downton, M. W., Morss, R. E., Wilhelmi, O. V., Gruntfest, E., and Higgins, M. L.: Interactions between scientific uncertainty and flood management decisions: two case studies in Colorado, Global Environ. Chang., 6, 134–146, https://doi.org/10.1016/j.hazards.2006.05.003, 2005.

Dutta, D., Herath, S., and Musiake, K.: A mathematical model for flood loss estimation, J. Hydrol., 277, 24–49, https://doi.org/10.1016/S0022-1694(03)00084-2, 2003.

Elmer, F., Thieken, A. H., Pech, I., and Kreibich, H.: Influence of flood frequency on residential building losses, Nat. Hazards Earth Syst. Sci., 10, 2145–2159, https://doi.org/10.5194/nhess-10-2145-2010, 2010.

Engel, H.: The flood event 2002 in the Elbe River basin, causes of the flood, its course, statistical assessment and flood damages, Houille Blanche, 6, 33–36, 2004.

Georgakakos, K. P., Seo, D. J., Gupta, H., Schaake, J., and Butts, M. B.: Towards the characterization of streamflow simulation uncertainty through multimodel ensembles, J. Hydrol., 298, 222–241, https://doi.org/10.1016/j.jhydrol.2004.03.037, 2004.

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, https://doi.org/10.1371/journal.pone.0159791, 2016.

Gleckler, P. J., Taylor, K. E., and Doutriaux, C.: Performance metrics for climate models, J. Geophys. Res.-Atmos., 113, 1–20, https://doi.org/10.1029/2007JD008972, 2008.

Grabbert, J. H.: Analyse der schadensbeeinflussenden Faktoren des Hochwassers 2002 und Ableitung eines mesoskaligen Abschätzungsmodells für Wohngebädeschäden, University of Potsdam, 2006.

Green, C., Viavattene, C., and Thompson, P.: Guidance for Assessing Flood Losses, CONHAZ project, report no. 6.1, 2011.

Hagedorn, R., Doblas-Reyes, F. J., and Palmer, T. N.: The rationale behind the success of multi-model ensembles in seasonal forecasting – I. Basic concept, Tellus A, 57, 219–233, https://doi.org/10.1111/j.1600-0870.2005.00103.x, 2005.

Hamill, T. M. and Colucci, S. J.: Verification of Eta–RSM short-range ensemble forecasts, Mon. Weather Rev., 125, 1312–1328, 1997.

Hersbach, H.: Decomposition of the continuous ranked probability score for ensemble prediction systems, Weather Forecast., 15, 559–570, https://doi.org/10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2, 2000.

Huttenlau, M., Stötter, J., and Stiefelmeyer, H.: Risk-based damage potential and loss estimation of extreme flooding scenarios in the Austrian Federal Province of Tyrol, Nat. Hazards Earth Syst. Sci., 10, 2451–2473, https://doi.org/10.5194/nhess-10-2451-2010, 2010.

Hydrotec: Hochwasser-Aktionsplan Lippe – Grundlagen, Überflutungsgebiete, Schadenspotenzial, Defizite und Maßnahmen, Aachen, 2002.

ICPR: Atlas on the Risk of Flooding and Potential Damage Due to Extreme Floods of the Rhine, International Commission for the Protection of the Rhine, Koblenz, 2001.

IKSE: Aktionsplan Hochwasserschutz Elbe, Internationale Kommission zum Schutz der Elbe, Magdeburg, 2003.

Jongman, B., Kreibich, H., Apel, H., Barredo, J. I., Bates, P. D., Feyen, L., Gericke, A., Neal, J., Aerts, J. C. J. H., and Ward, P. J.: Comparative flood damage model assessment: towards a European approach, Nat. Hazards Earth Syst. Sci., 12, 3733–3752, https://doi.org/10.5194/nhess-12-3733-2012, 2012.

Kelman, I. and Spence, R.: An overview of flood actions on buildings, Eng. Geol., 73, 297–309, https://doi.org/10.1016/j.enggeo.2004.01.010, 2004.

Kleist, L., Thieken, A. H., Köhler, P., Müller, M., Seifert, I., Borst, D., and Werner, U.: Estimation of the regional stock of residential buildings as a basis for a comparative risk assessment in Germany, Nat. Hazards Earth Syst. Sci., 6, 541–552, https://doi.org/10.5194/nhess-6-541-2006, 2006.

Klijn, F., Baan, P., de Bruijn, K., and Kwadijk, J.: Overstromingsrisico's in Nederland in een veranderend klimaat: Verwachtingen, schattingen en berekeningen voor het project Nederland Later, Deltares (WL), 2007.

Knutti, R., Furrer, R., Tebaldi, C., Cermak, J., and Meehl, G. A.: Challenges in combining projections from multiple climate models, J. Climate, 23, 2739–2758, https://doi.org/10.1175/2009JCLI3361.1, 2010.

Kreibich, H. and Dimitrova, B.: Assessment of damages caused by different flood types, WIT Trans. Ecol. Envir., 133, 3–11, https://doi.org/10.2495/FRIAR100011, 2010.

Kreibich, H. and Thieken, A. H.: Assessment of damage caused by high groundwater inundation, Water Resour. Res., 44, 1–14, https://doi.org/10.1029/2007WR006621, 2008.

Kreibich, H., Botto, A., Merz, B., and Schröter, K.: Probabilistic, multivariable flood loss modeling on the mesoscale with BT-FLEMO, Risk Anal., 37, 774–787, https://doi.org/10.1111/risa.12650, 2017.

Krzysztofowicz, R. and Davis, D. R.: Category-unit loss functions for flood forecast-response system evaluation, Water Resour. Res., 19, 1476–1480, 1983.

Leutbecher, M. and Palmer, T. N.: Ensemble forecasting, J. Comput. Phys., 227, 3515–3539, https://doi.org/10.1016/j.jcp.2007.02.014, 2008.

Luino, F., Cirio, C. G., Biddoccu, M., Agangi, A., Giulietto, W., Godone, F., and Nigrelli, G.: Application of a model to the evaluation of flood damage, GeoInformatica, 13, 339–353, https://doi.org/10.1007/s10707-008-0070-3, 2009.

Marzocchi, W., Taroni, M., and Selva, J.: Accounting for epistemic uncertainty in PSHA: logic tree and ensemble model, B. Seismol. Soc. Am., 105, 2151–2159, https://doi.org/10.1785/0120140131, 2015.

Merz, B. and Thieken, A. H.: Flood risk curves and uncertainty bounds, Nat. Hazards, 51(3), 437–458, https://doi.org/10.1007/s11069-009-9452-6, 2009.

Merz, B., Kreibich, H., Thieken, A., and Schmidtke, R.: Estimation uncertainty of direct monetary flood damage to buildings, Nat. Hazards Earth Syst. Sci., 4, 153–163, https://doi.org/10.5194/nhess-4-153-2004, 2004.

Merz, B., Kreibich, H., Schwarze, R., and Thieken, A.: Review article "Assessment of economic flood damage", Nat. Hazards Earth Syst. Sci., 10, 1697–1724, https://doi.org/10.5194/nhess-10-1697-2010, 2010.

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., Elmer, F., Kunz, M., Mühr, B., Schröter, K., and Uhlemann-Elmer, S.: The extreme flood in June 2013 in Germany, Houille Blanche, 1, 5–10, https://doi.org/10.1051/lhb/2014001, 2014.

Messner, F. and Meyer, V.: Flood damage, vulnerability and risk perception – challenges for flood damage research, in: Flood Risk Management: Hazards, Vulnerability and Mitigation Measures, Springer Netherlands, Dordrecht, https://doi.org/10.1007/978-1-4020-4598-1_13, 149–167, 2006.

Messner, F., Penning-Rowsell, E., Green, C., Meyer, V., Tunstall, S., and van der Veen, A.: Evaluating Flood Damages: Guidance and Recommendations on Principles and Methods, FLOODsite Project Deliverable D9.1, 2007.

Palmer, T. N., Alessandri, A., Andersen, U., et al.: Development of a European multimodel ensemble system for seasonal-to-interannual prediction (DEMETER), B. Am. Meteorol. Soc., 85, 853–872, https://doi.org/10.1175/BAMS-85-6-853, 2004.

Penning-Rowsell, E., Johnson, C., Tunstall, S., Tapsell, S., Morris, J., Chatterton, J., and Green, C.: The Benefits of Flood and Coastal Risk Management: A Handbook of Assessment Techniques, Middlesex University Press, https://doi.org/10.1596/978-0-8213-8050-5, 2005.

Peterman, R. M. and Anderson, J. L.: Decision analysis: a method for taking uncertainties into account in risk-based decision making, Hum. Ecol. Risk Assess., 5, 231–244, https://doi.org/10.1080/10807039991289383, 1999.

Pistrika, A. K. and Jonkman, S. N.: Damage to residential buildings due to flooding of New Orleans after hurricane Katrina, Nat. Hazards, 54, 413–434, https://doi.org/10.1007/s11069-009-9476-y, 2010.

Pregnolato, M., Galasso, C., and Parisi, F.: A compendium of existing vulnerability and fragility relationships for flood: preliminary results, in: 12th International Conference on Applications of Statistics and Probability in Civil Engineering, ICASP12, Vancouver, Canada, 12–15 July, 2015, 1–8, https://doi.org/10.14288/1.0076226, 2015.

Reese, S., Markau, H.-J., and Sterr, H.: MERK – Mikroskalige Evaluation der Risiken in überflutungsgefährdeten Küstenniederungen, Kiel, Forschungsprojekt im Auftrag des Bundesministeriums für Bildung und Forschung und des Ministeriums für ländliche Räume, Landesplanung, Landwirtschaft und Tourismus des Landes Schleswig-Holstein, 2003.

Regione del Veneto: 31 ottobre–2 novembre 2010: l'alluvione dei Santi, in: Rapporto Statistico 2011, 410–425, 2011a.

Regione del Veneto: Veneto, La grande alluvione, 2011b.

Reichler, T. and Kim, J.: How well do coupled models simulate today's climate?, B. Am. Meteorol. Soc., 89, 303–311, https://doi.org/10.1175/BAMS-89-3-303, 2008.

Riha, J. and Marcikova, M.: Classification and estimation of flood losses, in: International Symposium on Water Management and Hydraulic Engineering, Ohrid, Macedonia, 1–5 September 2009, 863–872, 2009.

Rossetto, T., Ayala, D. D., Ioannou, I., and Meslem, A.: Evaluation of existing fragility curves, in: SYNER-G: Typology Definition

and Fragility Functions for Physical Elements at Seismic Risk, edited by: Pitilakis, K., Crowley, H., and Kaynia, A. M., Springer Netherlands, https://doi.org/10.1007/978-94-007-7872-6_3, 47–93, 2014.

Scawthorn, C., Flores, P., Blais, N., Seligson, H., Tate, E., Chang, S., Mifflin, E., Thomas, W., Murphy, J., Jones, C., and Lawrence, M.: HAZUS-MH flood loss estimation methodology. II. Damage and loss assessment, Nat. Hazards Rev., 7, 72–81, https://doi.org/10.1061/(ASCE)1527-6988(2006)7:2(72), 2006.

Scherbaum, F. and Kuehn, N. M.: Logic tree branch weights and probabilities: summing up to one is not enough, Earthq. Spectra, 27, 1237–1251, https://doi.org/10.1193/1.3652744, 2011.

Schröter, K., Kreibich, H., Vogel, K., Riggelsen, C., Scherbaum, F., and Merz, B.: How useful are complex flood damage models?, Water Resour. Res., 50, 3378–3395, https://doi.org/10.1002/2013WR014396, 2014.

Scorzini, A. R. and Frank, E.: Flood damage curves: new insights from the 2010 flood in Veneto, Italy, J. Flood Risk Manag., 1–12, https://doi.org/10.1111/jfr3.12163, 2015.

Smith, D.: Flood damage estimation – a review of urban stage damage curves and loss functions, Water SA, 20, 231–238, 1994.

Spillatura, A.: Overview and Harmonization of Existing Vulnerability Functions for Italy, Istituto Universitario di Studi Superiori di Pavia, Pavia, Italy, 2014.

Spillatura, A., Fiorini, E., Bazzurro, P., and Pennucci, D.: Harmonization of vulnerability curves for masonry buildings, 2nd European Conference on Earthquake Engineering and Seismology (2ECEES), Istanbul, 25–29 August, 2014.

Talagrand, O., Vautard, R., and Strauss, B.: Evaluation of probabilistic prediction systems, in: Proceedings of ECMWF Workshop on Predictability, Reading, England, 20–22 October 1997, 1–25, 1997.

Tanoue, M., Hirabayashi, Y., and Ikeuchi, H.: Global-scale river flood vulnerability in the last 50 years, Sci. Rep.-UK, 6, 36021, https://doi.org/10.1038/srep36021, 2016.

Thieken, A. H., Müller, M., Kreibich, H., and Merz, B.: Flood damage and influencing factors: new insights from the August 2002 flood in Germany, Water Resour. Res., 41, 1–16, https://doi.org/10.1029/2005WR004177, 2005.

Thieken, A. H., Olschewski, A., Kreibich, H., Kobsch, S., and Merz, B.: Development and evaluation of FLEMOps – a new Flood Loss Estimation MOdel for the private sector, in: Flood Recovery, Innovation and Response I, WIT Press, https://doi.org/10.2495/FRIAR080301, 315–324, 2008.

Tóth, S., Kovács, S., and Kummer, L.: Vulnerability Analysis in the Körös-Corner Flood Area Along the Middle-Tisza River – Pilot Study Application of General Vulnerability Analysis Techniques, FLOODsite Project Deliverable D22.3, 2008.

Ulbrich, U., Brücher, T., Fink, A. H., Leckebusch, G. C., Krüger, A., and Pinto, J. G.: The central European floods of August 2002: Part 1 – Rainfall periods and flood development, Weather, 58, 371–377, https://doi.org/10.1256/wea.61.03A, 2003.

USACE: Guidelines for Risk and Uncertainty Analysis in Water Resources Planning, Vol. I, USACE, Fort Belvoir, VA, USA, 1992.

Vanneuville, W., Maddens, R., Collard, C., Bogaert, P., De Maeyer, P., and Antrop, M.: Impact op mens en economie t.g.v. overstromingen bekeken in het licht van wijzigende hydraulische condities, omgevingsfactoren en klimatologische omstandigheden, studie uitgevoerd in opdracht van de Vlaamse Milieumaatschappij, MIRA, MIRA/2006/02, UGent, 2006.

Vogel, K., Riggelsen, C., Merz, B., Kreibich, H., and Scherbaum, F.: Flood damage and influencing factors: a Bayesian network perspective, in: Proceedings of the 6th European Workshop on Probabilistic Graphical Models (PGM 2012), Granada, Spain, edited by: Cano, A., Gómez-Olmedo, M. G., and Nielsen, T. D., 19–21 September 2012, 314–354, 2012.

Vojinovic, Z., Ediriweera, J. C. W., and Fikri, A. K.: An approach to the model-based spatial assessment of damages caused by urban floods, in: 11th International Conference on Urban Drainage, Edinburgh, Scotland, UK, 31 August–5 September, 2008.

Wagenaar, D. J., de Bruijn, K. M., Bouwer, L. M., and de Moel, H.: Uncertainty in flood damage estimates and its potential effect on investment decisions, Nat. Hazards Earth Syst. Sci., 16, 1–14, https://doi.org/10.5194/nhess-16-1-2016, 2016.

Weigel, A. P.: Ensemble forecasts, in: Forecast Verification: A Practitioner's Guide in Atmospheric Science, edited by: Jolliffe, I. T. and Stephenson, D. B., John Wiley & Sons, Ltd, Chichester, UK, https://doi.org/10.1002/9781119960003.ch8, 141–166, 2012.

Weigel, A. P., Liniger, M. A., and Appenzeller, C.: Can multi-model combination really enhance the prediction skill of probabilistic ensemble forecasts?, Q. J. Roy. Meteor. Soc., 134, 241–260, 2008.

Wünsch, A., Herrmann, U., Kreibich, H., and Thieken, A. H.: The role of disaggregation of asset values in flood loss estimation: a comparison of different modeling approaches at the Mulde River, Germany, Environ. Manage., 44, 524–541, https://doi.org/10.1007/s00267-009-9335-3, 2009.

Yazdi, J. and Neyshabouri, S. A. A.: Optimal design of flood-control multi-reservoir system on a watershed scale, Nat. Hazards, 63, 629–646, https://doi.org/10.1007/s11069-012-0169-6, 2012.