Schröter, K., Kreibich, H., Vogel, K., Riggelsen, C., Scherbaum, F., Merz, B. (2014): How useful are complex flood damage models? - *Water Resources Research*, *50*, 4, p. 3378-3395.

DOI: http://doi.org/10.1002/2013WR014396

AGU PUBLICATIONS

Water Resources Research

RESEARCH ARTICLE

10.1002/2013WR014396

Key Points:

- Increased complexity improves the predictive capability of flood damage models
- Model approach seems more important than using additional variables
- Bayesian network-based predictions show superior precision and reliability

Correspondence to:
K. Schröter,
kai.schroeter@gfz-potsdam.de

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

Received 9 JULY 2013
Accepted 27 MAR 2014
Accepted article online 1 APR 2014
Published online 22 APR 2014

How useful are complex flood damage models?

Kai Schröter[1], Heidi Kreibich[1], Kristin Vogel[2], Carsten Riggelsen[2], Frank Scherbaum[2], and Bruno Merz[1]

[1]Helmholtz Centre Potsdam, GFZ German Research Centre for Geosciences, Section Hydrology, Potsdam, Germany,
[2]University of Potsdam, Institute of Earth and Environmetal Science, Potsdam, Germany

Abstract We investigate the usefulness of complex flood damage models for predicting relative damage to residential buildings in a spatial and temporal transfer context. We apply eight different flood damage models to predict relative building damage for five historic flood events in two different regions of Germany. Model complexity is measured in terms of the number of explanatory variables which varies from 1 variable up to 10 variables which are singled out from 28 candidate variables. Model validation is based on empirical damage data, whereas observation uncertainty is taken into consideration. The comparison of model predictive performance shows that additional explanatory variables besides the water depth improve the predictive capability in a spatial and temporal transfer context, i.e., when the models are transferred to different regions and different flood events. Concerning the trade-off between predictive capability and reliability the model structure seem more important than the number of explanatory variables. Among the models considered, the reliability of Bayesian network-based predictions in space-time transfer is larger than for the remaining models, and the uncertainties associated with damage predictions are reflected more completely.

1. Introduction

Flood losses have increased worldwide during the last decades [*Barredo*, 2009; *Kron et al*., 2012; *UNISDR*, 2011]. At the same time, the perception that floods are recurrent natural phenomena and the recognition that both the hazard and the vulnerability ultimately control flood losses have pushed the implementation of risk oriented approaches to flood design and flood risk management [*Merz et al*., 2004; *EU*, 2007].

Hence, flood damage assessments are of growing importance since damage has to be estimated in any deliberation of cost-effectiveness of flood mitigation measures, analyses of vulnerability and resilience, land use planning, flood risk mapping, comparative risk analyses, and financial appraisal [*Merz et al*., 2010]. For these tasks, reliable models to estimate flood damage are an essential component [*Dutta et al*., 2003; *Kang et al*., 2005; *Thieken et al*., 2005].

Flood-damaging processes are complex: they are influenced by the interplay of various hydrological, hydraulic, and socioeconomic factors [e.g., *Kelman and Spence*, 2004; *Thieken et al*., 2005; *Schwarz and Maiwald*, 2007; *Kreibich et al*., 2009]. In contrast to this complexity, common damage estimation methods are simple [*Merz et al*., 2010]. Traditional damage models are based on the type and/or use of the element at risk and the water depth as the exclusive determining factors for the estimation of damage [*NRC*, 2000; *Green*, 2003]. Essentially, this is due to limited data and knowledge about the single and joint effects of other damage-influencing factors. Important challenges remain to advance the understanding of the damaging process [*Bubeck and Kreibich*, 2011], to deepen its theoretical foundations [*Wind et al*., 1999], and to develop reliable damage models [*Merz et al*., 2013]. Flood damage modeling is subject to considerable uncertainty [*Merz and Thieken*, 2005; *de Moel and Aerts*, 2011]. This uncertainty stems from various sources including incomplete knowledge about the damaging process, which crystallizes, for instance, in generalizations concerning the damage-influencing factors and aggregated input data. Further, numerous quantities involved in the damage process are inherently variable as, for example, the flow velocity or inundation duration.

There are several examples of model developments aimed at a more comprehensive consideration of damage-influencing variables. *Wind et al*. [1999] account for flood warning time and flood experience in flood damage estimation. *Zhai et al*. [2005] include house type, length of residence, and household income in a probabilistic damage model. *Thieken et al*. [2008] set up the rule-based model FLEMOps+ using inundation depth, building type and quality, contamination, and precaution as explanatory variables for flood damage. *Elmer et al*. [2010] includes flood frequency as an additional variable to FLEMOps+r. *Merz et al*. [2013] derive a multivariate model using tree-based methods considering 28 potential explanatory variables describing the flooding situation, early warning and emergency measures, precaution, building characteristics, and the socioeconomic characteristics. Using the same set of candidate variables, *Vogel et al*. [2012] take a data-driven Bayesian network perspective for the development of a probabilistic damage model.

One important step in model development is model validation. The purpose of validation is to evaluate model-generated and real system data and thus to prove the suitability of the model to describe real system behavior. The level of model validation has to reflect the intended purpose of the model application [*Power*, 1993]. The validation of a model to predict flood damage has to evaluate the performance of the model not only in replicative applications but more importantly in predictive applications. In this context, replicative applications refer to a comparison of model results to observed damages which have been used to develop and/or fit the model. Predictive applications correspond to estimating damage and comparing model outcomes to damage data which have not been included in the model development.

In general, model validation is scarcely performed in loss modeling. This might be due to limited or missing data: damage data are rarely gathered, repair cost estimates are uncertain, and data are not updated systematically [*Downton and Pielke*, 2005]. Some damage model validation studies are available: *Penning-Rowsell and Green* [2000] compared synthetic damage functions of *Penning-Rowsell and Chatterton* [1977] against postflood surveys derived by damage adjusters. This validation resulted in a general agreement between surveys and synthetic results. *Ding et al*. [2008] report a good agreement between damage estimates using the HAZUS-MH "level 2" flood damage model [*Scawthorn et al*., 2006] and the outcomes of an alternative detailed approach to flood damage estimation; hence, the validation is not based on observations in this case. *Elmer et al*. [2010] successfully tested the model FLEMOps+r against other models using a leave-one-out cross-validation method. Similar to *Kreibich and Thieken* [2008], who used a split sampling technique to analyze the performance of 12 different versions of stage-damage functions and two different versions of rule-based damage models predicting groundwater flood damage, *Thieken et al*. [2008] compared damage estimates of FLEMOps+ with observed repair costs for the August 2002 flood for several municipalities in Saxony (Germany) and showed that the model delivers very good damage estimates. However, the model has been developed using damage data for the same flood in the same region. Testing the model predictive capability by applying it to the 1993 flood in a different region in Germany (Baden-Wuerttemberg) showed much larger deviations. *Jongman et al*. [2012] compared seven damage models on a predictive validation level. They reported considerable differences in model predictive capability across regions. Recently, *Cammerer et al*. [2013] investigated the transferability of flood damage models to other geographical regions by comparing model results to official damage data. The study confirms that flood damage models which have been derived on data from geographical regions with comparable building and flood event characteristics perform better than those based on more heterogeneous data sets encompassing different regions and floods.

These findings suggest that the predictive capability of flood damage models is rather weak, especially when a temporal and spatial transfer is involved, i.e., the damage models are applied to different flood events and/or in different regions than those which have been used to derive the model. Against this background, we investigate whether and to which extent complex flood damage models are useful to improve the predictive performance in terms of variation, precision, and reliability. In this context, model complexity is basically related to the ability of the model to capture and to reproduce complex processes and thus depends on various factors such as the number and type of explanatory variables included, the interactions between those variables described by the model, and the functional form of those interactions. The functional form is ultimately defined by the model structure and varies from predefined functional relations, e.g., traditional stage-damage functions, to probabilistic dependencies derived from observations without any prior assumption concerning the functional form, e.g., Bayesian networks. In view of these conceptual differences, a general and consistent measure of model complexity can hardly be defined. Measures such as the Akaike or Bayesian information criterion (AIC or BIC) are of little value in this regard, since they are not consistent across the different model approaches. Therefore, we use the number of explanatory variables (predictors) included in the model as an indicator for model complexity.

Table 1. Characteristics of Flood Damage Models

	Damage Model	Knowledge Basis for Model Derivation	Number of Explanatory Variables	Explanatory Variables[a]	Input Requirements	Outcome	Reference
1	Square-root function	Expert knowledge	1	wst	Requires complete observations	Point estimate for rloss	*Buck and Merkel* [2009]
2	FLEMOps+r	Expert knowledge	6	wst, bv, bq, con, pre, rp	Requires complete observations	Point estimate for rloss	*Elmer et al.* [2010]
3	RT	Data mining	5[b]	wst, bv, age, con, rp	Approximate prediction for incomplete observations	Point estimate for rloss	*Merz et al.* [2013]
4	RTp	Data mining	2[b]	wst, bv	Approximate prediction for incomplete observations	Point estimate for rloss	*Merz et al.* [2013]
5	BNd29	Data mining	6[c]	wst, d, con, v, pre, epre	Able to predict with incomplete observations	Distribution of rloss	*Vogel et al.* [2012]
6	BNd11	Data mining	3[c]	wst, con, pre	Able to predict with incomplete observations	Distribution of rloss	*Vogel et al.* [2012]
7	BNe28	Data mining and expert knowledge	10[c]	wst, con, d, v, rp, bq, bv, bt, em, pre	Able to predict with incomplete observations	Distribution of rloss	
8	BNe10	Data mining and expert knowledge	8[c]	wst, con, d, rp, bv, bt, bq, pre	Able to predict with incomplete observations	Distribution of rloss	

[a]See Table 3 for detailed explanations.
[b]Considering the model structure for the "no uncertainty" scenario.
[c]Considering the variables on the Markov Blanket only.

To test the hypothesis that increasing complexity of flood damage models improves the predictive capability in a spatial and temporal transfer context, we apply eight flood damage models for three flood events in two different regions in Germany. This enables us to distinguish between different model usages, namely, local, cross regional, and/or temporal transfer applications. We focus on the estimation of direct damage to residential buildings. Flood damage estimation is carried out on the scale of the individual buildings in terms of relative damage (*rloss*), whereas relative building damage is defined as the ratio of the actual building loss and its total replacement value [*Elmer et al*., 2010]. The predictive performance of the models is assessed by comparing modeled to observed relative building damage.

2. Setup of Validation Exercise

The damage models compared are a stage-damage function, using only water depth as explanatory variable, FLEMOps+r, a rule-based model using six variables, and data mining approaches, namely regression trees and Bayesian networks, using up to 28 explanatory variables. Table 1 summarizes key qualities of these damage models. A detailed description will be given in section 3.

The comparison of model performance is based on empirical damage data. Flood damage records are available from computer-aided telephone interviews that were compiled after the floods in 2002, 2005, and 2006, respectively, in the Elbe and Danube catchments in Germany [*Thieken et al*., 2007; *Kreibich and Thieken*, 2009; *Kreibich et al*., 2011]. The considered flood events and data sets are portrayed in sections 2.1 and 2.2. The location of the communities in which interviews have been undertaken is shown in Figure 1. For the most part, varying localities were affected by the different flood events except for some communities mainly in the region of Dresden (Elbe catchment).

This database is split up according to different floods (2002, 2005, and 2006) and to different regions (Elbe and Danube catchments), see Table 2. Furthermore, the subsample for the 2002 flood in the Elbe catchment is partitioned randomly into two parts: two third of the data set is used to derive the damage models (Elbe 2002 id), and the remaining third (Elbe 2002 pr) is used for local validation.

The derivation of the models includes all steps to set up the model for the estimation of *rloss*. Depending on the model approach, this involves the selection of the model structure (i.e., functional forms and functional relations), the estimation of model parameters, scaling factors, parameter distributions, and their discretization.

Next, the models are applied to predict *rloss* for the flood events (Elbe 2002 pr, Danube 2002, Danube 2005, and Elbe 2006). For each event, the models are validated by comparing model results to observed damage data using the criteria detailed in section 2.4. The subdivision of damage data according to different events

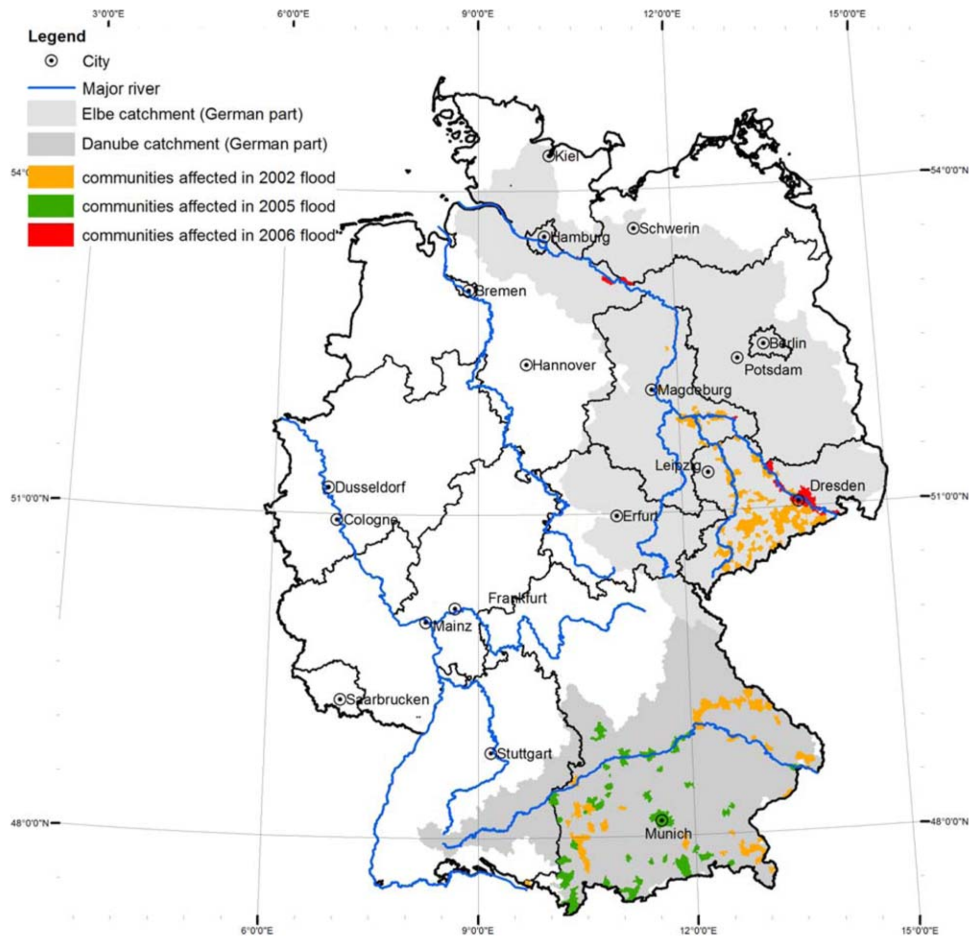

Figure 1. Location of communities in which interviews had been undertaken after the 2002, 2005, and 2006 flood events in the Elbe and Danube catchments (Germany), data sources—rivers and administrative borders: DLM1000 of BKG (2001); cities: Esri, DeLorme Publishing Company, Inc.; catchment areas: CCM2 data of JRC [*Vogt et al.*, 2007].

and regions allows for the evaluation of model predictive capability in spatial and temporal transfer applications.

In view of the uncertainty present in flood damage modeling, we are particularly addressing the uncertainty associated with the observations of the explanatory variables, which have been acquired via surveys with flood-affected private households, e.g., uncertainties due to imprecise observations, bad memory, reluctant answers, and misunderstandings. To ensure the robustness of the results against disturbances induced by observation uncertainty, we explicitly consider this uncertainty. In this way, we substantiate that the differences in the predictive performance of the various damage models are actually a consequence of different model complexity and not the outcome of uncertainties in the underlying database. The procedure to consider observation uncertainty is outlined in section 2.3.

Table 2. Usage of Data Subsamples From Different Flood Events and River Basins

Subsample	Year	Location	Interviews Completed	Rloss Given	Usage
Elbe 2002 id	2002	Elbe	850	426	Model derivation
Elbe 2002 pr	2002	Elbe	398	235	Local validation
Danube 2002	2002	Danube	449	286	Cross-regional validation
Danube 2005	2005	Danube	275	116	Cross-regional and Temporal validation
Elbe 2006	2006	Elbe	126	46	Temporal validation

2.1. Flood Events

The extreme flood in central Europe in August 2002 was caused by intense long-lasting precipitation covering large areas in Austria, Slovakia, the Czech-Republic, and Germany. As a result, flash floods were observed in the headwaters of the Elbe tributaries in the Ore Mountains as well as in some alpine tributaries of the Danube and in the Bohemian Forest. Further downstream, floods came along with unprecedented water levels, flooding of polders, and vast inundated areas as a result of levee overtopping and numerous levee breaches, in particular, in the Elbe catchment. Twenty-two people were killed in Germany during this flood. Infrastructures and buildings suffered substantial damage. As shown in Figure 1, mainly the tributaries in the Ore Mountains (Saxony, Elbe catchment), the river Regen in Bavaria, and several southern tributaries of the Danube were affected. Total damage amounted to 9.9 Bn € in the German part of the Elbe catchment and 0.2 Bn € in the German part of the Danube catchment [*IKSE*, 2004].

In August 2005, another considerable flood affected the Alpine region, particularly Switzerland but also the German part of the Danube catchment. The alpine foothills were affected by flash floods characterized by a rapid increase of discharges and water levels. Inundations occurred along the Danube and its southern tributaries being the main areas reporting damage in Germany (see Figure 1). Flood protection measures and effective operation of dams reduced the flood impact. The total economic damage is estimated at about 190 M € [*LfU*, 2006].

In the Elbe catchment, another flood event followed in April 2006, which was caused by the combination of widespread heavy rainfall and snowmelt in the upper catchment. In Dresden, the maximum water level was clearly below the 2002 flood peak. In contrast, the flood situation downstream of the Havel confluence was comparable or even worse than during the 2002 flood. Several towns in the Saxon Elbe valley and in the lower reaches of the Elbe were inundated (Figure 1). Several hundreds of people were evacuated. The resulting damage in Germany was estimated to be 75 M € [*Munich Re*, 2009].

2.2. Empirical Damage Data Collection

After the floods of 2002, 2005, and 2006 in the Elbe and Danube catchments in Germany, damage data have been collected from affected households via computer-aided telephone interviews. In total, data from 2098 interviews are available, see Table 2. The survey for the 2002 flood resulted in 1248 completed interviews in the Elbe catchment and 449 completed interviews in the Danube catchment. The survey for the 2005 flood in the Danube basin provided 275 interviews. For the 2006 flood in the Elbe catchment, 126 interviews were completed. However, the loss ratio for the damaged buildings (*rloss*) could not be provided for all interviews. As *rloss* is the variable to be predicted by the model, the data set is limited to 1109 damage cases.

The households interviewed were randomly sampled from lists which have been compiled of all streets affected by flooding with the help of information from local authorities, flood reports, or press releases as well as with the help of flood footprints derived from satellite radar data (DLR, Centre for Satellite Based Crisis information, www.zki.dlr.de). The raw data were supplemented by estimates of return period, building values, loss ratio, i.e., the relation between the actual building damage and replacement costs, and indicators for flow velocity, contamination, flood warning, emergency measures, precautionary measures, flood experience, and socioeconomic variables [*Thieken et al*., 2005; *Elmer et al*., 2010].

From this extensive data set, 28 candidate variables were preselected to be used in a modeling context for predicting the loss ratio of residential buildings (*rloss*), see Table 3. These candidate variables were selected according to experiences from previous analyses [*Thieken et al*., 2005; *Merz et al*., 2013]: 5 variables are related to the hydrological and hydraulic aspects of the flooding situation at the affected building, 10 variables are related to damage reduction, particularly to early warning and emergency measures undertaken, as well as to the state of precaution of the household, and 13 variables are related to the residential building characteristics and the socioeconomic status of the household.

2.3. Consideration of Observation Uncertainty

The collection of empirical damage data via computer-aided telephone interviews applied crosschecks of answers during the interviews to avoid contradictions and to improve data quality. The information retrieved from the interviews is considered to be of comparatively high quality and to be free of a strategic response bias [*Kreibich et al*., 2005; *Thieken et al*., 2007]. This was confirmed via a comparison of the damage

Table 3. Description of Candidate Variables for Flood Damage Modeling[a]

	Abbreviation	Variable	Scale and Range	Uncertainty Model
Flooding Situation				
1	wst	Water depth[b]	c: 248 cm below ground to 670 cm above ground	∼N(m,sd)
2	d	Inundation duration	c: 1–1,440 h	∼N(m,sd)
3	v	Flow velocity indicator	o: 0 = still to 3 = high velocity	C(Px)
4	con	Contamination indicator	o: 0 = no contamination to 6 = heavy contamination	C(Px)
5	rp	Return period	c: 1–848 years	C(Px)
Early Warning and Emergency Measures				
6	wt	Early warning lead time	c: 0–336 h	∼N(m,sd)
7	wq	Quality of warning	o: 1 = receiver of warning knew exactly what to do to 6 = receiver of warning had no idea what to do	C(Px)
8	ws	Indicator of flood warning source	o: 0 = no warning to 4 = official warning through authorities	C(Px)
9	wi	Indicator of flood warning information	o: 0 = no helpful information to 11 = many helpful information	C(Px)
10	wte	Lead time period elapsed without using it for emergency measures	c: 0–335 h	∼N(m,sd)
11	em	Emergency measures indicator	o: 1 = no measures undertaken to 17 = many measures undertaken	C(Px)
Precaution				
12	pre	Precautionary measures indicator	o: 0 = no measures undertaken to 38 = many, efficient measures undertaken	C(Px)
13	epre	Perception of efficiency of private precaution	o: 1 = very efficient to 6 = not efficient at all	C(Px)
14	fe	Flood experience indicator	o: 0 = no experience to 9 = recent flood experience	C(Px)
15	kh	Knowledge of flood hazard	n (yes/no)	C(Px)
Building Characteristics				
16	bt	Building type	n (1 = multifamily house, 2 = semidetached house, 3 = one-family house)	C(Px)
17	nfb	Number of flats in building	c: 1–45 flats	∼N(m,sd)
18	fsb	Floor space of building	c: 45–18,000 m^2	∼N(m,sd)
19	bq	Building quality	o: 1 = very good to 6 = very bad	C(Px)
20	bv	Building value	c: 92,244–3,718,677 €	∼N(m,sd)
Socioeconomic Status				
21	age	Age of the interviewed person	c: 16–95 years	
22	hs	Household size, i.e., number of persons	c: 1–20 people	
23	chi	Number of children (<14 years) in household	c: 0–6	
24	eld	Number of elderly people (>65 years) in household	c: 0–4	
25	own	Ownership structure	n (1 = tenant; 2 = owner of flat; 3 = owner of building)	
26	inc	Monthly net income in classes	o: 11 = below 500 € to 16 = 3,000 € and more	C(Px)
27	socp	Socioeconomic status according to *Plapp* [2003]	o: 3 = very low socioeconomic status to 13 = very high socioeconomic status	C(Px)
28	socs	Socioeconomic status according to *Schnell et al.* [1999]	o: 9 = very low socioeconomic status to 60 = very high socioeconomic status	C(Px)
Flood Damage				
29	rloss	Loss ratio of residential building	c: 0 = no damage to 1 = total damage	∼N(m,sd)

[a]c: continuous; o: ordinal; n: nominal.
[b]The depth of the basement of a building is assumed to be 250 cm below ground level.

data collected after the 2002 flood with official damage data in the federal state of Saxony [*Thieken et al*., 2005]. Nonetheless, to ensure the robustness of the findings against disturbances resulting from observation uncertainty, we assume that the observations, i.e., the interview responses, are uncertain. Uncertainty is understood as the deviation of the interview responses to the real situation. For the consideration of the observation uncertainty within the evaluation of the predictive performance, we need to discern the fundamentally different modeling approaches considered in this study. Bayesian networks are probabilistic models, which as a basic principle treat all quantities involved as random variables, and do not distinguish between explanatory and response variables. This enables to capture the joint probability distribution of all variables. Consequently, the Bayesian network models the probabilistic dependency among the variables. Thereby, it is assumed that observations are uncertain. On that note, the Bayesian network inherently captures the uncertainty that is related to the observations of the individual variables and thus implicitly includes this uncertainty in the joint probability distribution modeled. Accordingly, the joint probability distribution reflects both the probabilistic dependence of the variables and the observation uncertainty associated with the variables. An explicit consideration of observation uncertainty is not indicated since it would be double accounted. However, Bayesian networks do not allow to distinguish between observation uncertainty and other uncertainty sources uncertainty, i.e., it cannot be determined to which extent the uncertainty in the probabilistic dependence is actually due to uncertain observations.

In contrast, stage-damage functions, FLEMOps+r, and regression trees assume a deterministic relation between explanatory and predicted variables, providing point estimates for specific observations of explanatory variables. For the derivation of these models, methods of regression analysis are applied. Usually in regression analysis, the observations are assumed to be error free. In this framework, a feasible approach to account for the observation uncertainty is to explicitly describe this uncertainty and to feed it into the analysis [*Saltelli et al*., 2000]. This includes the following steps: (i) define uncertainty models which quantify the uncertainty associated with the observations of each explanatory variable, (ii) generate a large number of variations of the observed data set, Elbe 2002 id, by random sampling from the different uncertainty models using Monte Carlo techniques, (iii) derive a sample of models using each realization of the data set variations generated in step (ii), (iv) predict flood damage for the other data subsamples (Elbe 2002 pr, Danube 2002, Danube 2005, and Elbe 2006) using each sample member of the models derived in step iii. As a result, for each damage model and data subsample we obtain a distribution of model predictions which represents the predictive uncertainty of the models given the assumed uncertainty in data observations.

The definition of uncertainty models for the 28 explanatory variables is difficult, since knowledge about the degree and characteristics of observation uncertainty is hardly available. As an exception, the water depths (*wst*) obtained from the telephone interviews after the Elbe 2002 flood have been compared with a sample of water marks at 409 buildings in the community of Eilenburg (Saxony) [*Poser and Dransch*, 2010]. A bias of 0.37 m and a root mean square error of 0.76 m are reported indicating the order of magnitude of the deviation (in this case ca. 30%) and thus the degree of uncertainty concerning *wst* based on telephone interviews. However, the results from this localized comparison can hardly be generalized.

In this light, we need to resort to several assumptions concerning the degree and the characteristics of observation uncertainty. First, the explanatory variables are distinguished concerning their underlying scale: continuous, ordinal, or nominal scale. For the continuous variables (e.g., *wst*, *d*), we describe the uncertainty in the measurement process using an uncertainty model which adds Gaussian noise to the observed values. In this model, the mean (m) defines the systematic deviation from the unobservable true value and the standard deviation (sd) controls the magnitude of the deviations. We parameterize the uncertainty model in such a way that m = 0 and sd corresponds to a constant percentage of the individual observations.

The variables based on ordinal or nominal scales describe differences in the observations using a spectrum of values (e.g., *con*: 0 = no contamination, 6 = heavy contamination) or qualitative classifications (e.g., *kh*: yes, no). Hence, for these variables observation uncertainty refers to the attribution to a specific category (C). We describe this uncertainty by setting a probability level (Px) for an erroneous categorization of an observation. When an observation is wrongly classified, the observation is allocated to a wrong category assuming uniform probability for all possible categories. For the variables *age*, *hs*, *chi*, *eld*, and *own*, we assume that the uncertainty of the measurement process is negligible. The uncertainty models for Gaussian noise (∼N(m,sd)) and categorical uncertainty (C(Px)) applied to the explanatory variables are listed in Table 3.

As there is hardly any evidence to quantitatively frame the observation uncertainty and the cross-correlation structure of the explanatory variables, we investigate the potential implications for the predictive performance of the deterministic damage models in terms of different uncertainty scenarios comprising (i) no uncertainty, (ii) small uncertainty, and (iii) large uncertainty. Within these scenarios we assume that the observations of the different variables are independent from each other. Further, we control the degree of uncertainty by varying both the magnitude of the standard deviation for continuous variables and the probability level for erroneous observations for ordinal and nominal variables. The first scenario (no uncertainty) corresponds to a deterministic regression approach, i.e., error-free input data which yields point estimates for *rloss*. Within the small uncertainty scenario, we set the standard deviation to 5% of the observed value and the probability level for an observation to be erroneous to 5%. Large uncertainty corresponds to a standard deviation of 20% of the observed value and a probability level for erroneous observations of 20%. For the latter two uncertainty scenarios, we generate a sample size of 5000 realizations for each observation of explanatory variables. This sample size has been checked for stable convergence using the median and the quantile range at the 90% level as diagnostics. Note we have also conducted the analyses for an additional scenario of "very large uncertainty" defined by a standard deviation of 50% and a probability level for erroneous observations of 50%. The outcomes for this scenario are very similar to the "large uncertainty" scenario and, therefore, are not discussed in detail.

Table 4. Model Evaluation Criteria[a]

Criterion	LB	UB	OPT
$mbe=\frac{1}{n}\sum_{i=1}^{n}(Q_{50i}-O_i)$	$-$inf	inf	0
$mae=\frac{1}{n}\sum_{i=1}^{n}\lvert Q_{50i}-O_i\rvert$	0	inf	0
$QR_{90}=\frac{1}{n}\sum_{i=1}^{n}(Q_{95i}-Q_{05i})/Q_{50i}$	0	inf	0
$HR=\frac{1}{n}\sum_{i=1}^{n}h_i;\ h_i=\begin{cases}1, if O_i\in[Q_{95i},Q_{05i}]\\ 0, otherwise\end{cases}$	0	1	0.9[b]

[a]LB: lower bound; UB: upper bound; OPT: perfect prediction.
[b]Depending on the nominal coverage rate applied, e.g., 0.9 for the 95–5 quantile range.

2.4. Evaluation Criteria

The predictive capability of the various damage models is evaluated with regard to precision, variation, and reliability. For the scenarios of small and large uncertainty, we use the median (Q_{50}) as a summary statistic of the distribution of model predictions. The Bayesian networks naturally provide the conditional probabilities of the predicted variable *rloss* and thus also capture the prediction uncertainty. For the model comparison also uses the median of *rloss* distribution.

The model precision is evaluated in terms of the mean bias (*mbe*) and the mean absolute error (*mae*) as listed in Table 4. Both criteria evaluate the model residuals, i.e., the differences between predictions (Q_{50}) and observations (O) of *rloss*. The mean bias error (*mbe*) provides information about a systematic deviation, i.e., an average overprediction or under prediction. The mean absolute error (*mae*) describes the average magnitude of the residuals and allows for a dimensioned comparison of average model precision [*Willmott and Matsuura*, 2005].

The variation of the model predictions is quantified using the quantile range at the 90% level (QR_{90}), see Table 4. A smaller value for QR_{90} corresponds to a smaller spread in the predictions, i.e., the model predictions are less uncertain.

Concerning the model reliability we evaluate whether the predictive distribution actually covers the observed values of *rloss*. For this purpose, we compute the hit rate (HR), which represents the proportion of the number of observations that fall within the 95 and 5 quantile predictive interval and the total number of observations available, see Table 4. The 95–5 quantile range corresponds to a nominal coverage of 0.9, and thus a $HR=0.9$ indicates that the coverage of model predictions is equal to the nominal coverage representing a perfect reliability of model prediction on this level [*Thordarson et al*., 2012].

We derive the models and calculate the model predictions for the log transformed variable *rloss* (log-*rloss*). This transformation reduces the influence of the few very high loss ratios present in the data sample.

3. Damage Models

We compare eight flood damage models of different complexity, see Table 1. In particular, we examine a "traditional" depth-damage function based on a root function [*Buck and Merkel*, 1999], the rule-based model FLEMOps+r [*Elmer et al*., 2010], two variants of the regression tree model approach using different numbers of explanatory variables proposed by *Merz et al*. [2013], and two variants of Bayesian network approaches: first, completely data based [*Vogel et al*., 2012] and, second, using both expert knowledge and data to derive the network structure. Further, the Bayesian network approach is applied taking into consideration two different numbers of explanatory variables.

3.1. Depth-Damage Function

Damage functions are a central concept of damage estimation. They relate the damage for the respective element at risk to characteristics of the inundation [*Merz et al*., 2010]. Most often, damage is estimated for the type or use of the element at risk and the inundation depth [*Wind et al*., 1999; *NRC*, 2000]. Such depth-damage curves were for the first time proposed in the USA [*White*, 1945] and since then have been applied in numerous models and case studies. Depth-damage functions remain the standard approach to assessing urban flood damage [*Smith*, 1994; *Merz et al*., 2010]. For Germany, stage-damage curves as separate square-root functions for water depths in the basement and above ground floor were suggested by *Buck and Merkel* [1999].

In this study, we apply a combination of root functions to predict *rloss* in the basement and in the building above ground level separately. The model structure is defined as given in equation (1) for the basement and equation (2) above the ground level

$$rloss = a_b + b_b\sqrt{w_b} \quad ; w_b = w + 250 \quad \{w | w < 0\} \tag{1}$$

$$rloss = a + b\sqrt{w} \quad \{w | w \geq 0\} \tag{2}$$

With w (water depth in relation to the ground level), w_b (water depth above basement level), and a, b, a_b, b_b parameters of the respective root functions. A general basement height of 250 cm is assumed.

The derivation of the model consists in estimating the parameters a, b, and a_b, b_b. This is done in a regression approach by minimizing the sum of squared residuals, i.e., the differences between modeled and observed damage (*rloss*) in the subsample Elbe 2002 id. These best estimates of model parameters are used to predict flood damage (*rloss*) for observed water levels during the other events.

3.2. FLEMOps+r

The Flood Loss Estimation Model for the private sector (FLEMOps) uses a rule-based multifactorial approach to estimate direct tangible damage to residential buildings. It has been developed at the German Research Centre for Geosciences, primarily for scientific flood risk analyses from the local to national scale [e.g., *Apel et al*., 2009; *Vorogushyn et al*., 2012].

Since the initial version proposed by *Thieken et al*. [2008], the model has undergone several enhancements including an increasing number of explanatory variables. We apply the most recent version FLEMOps+r [*Elmer et al*., 2010]. This version incorporates six explanatory variables. Flood damage is calculated using five different classes of water depth, three classes of contamination, and three classes of flood frequency, three individual building types, two classes of building quality, and three classes of private precaution. The class limits for inundation depth, building types, and building quality have been defined in a way to appropriately reflect the range and variability included in available observations and other basic data as defined in *Büchele et al*. [2006] and *Thieken et al*. [2008].

The derivation of the model within this study comprises two steps. First, for each class combination of inundation depth, building types and building quality mean damage values are determined from observed damage records (Elbe 2002 id). Second, a set of scaling factors are derived for each class combination to reflect the impact of contamination, precaution, and flood frequency. Again, these scaling factors are derived from observed damage records (Elbe 2002 id).

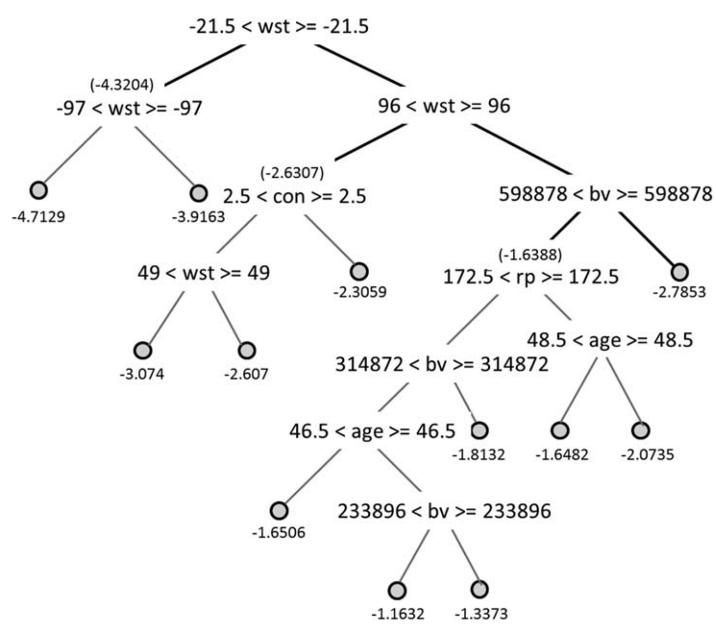

Figure 2. Regression tree RT with 12 leaves considering five variables for estimating log(*rloss*) and pruned regression tree (RTp, indicated by bold joins) with 4 leaves considering two variables for estimating log(*rloss*) grown within the "no uncertainty" scenario.

3.3. Regression Trees

Following the approach by *Merz et al*. [2013], a regression tree has been grown based on 28 candidate explanatory variables (Table 3) within the subsample Elbe 2002 id. For the scenario of "no uncertainty," this results in a tree with 12 terminal nodes (Figure 2) considering five variables. The most important variable is water depth, followed by building value, age, contamination, and return period.

Using regression trees, overfitting needs careful attention. Hence, the large tree RT is cut back to obtain a simpler tree, which, however, should have a predictive error comparable to the most accurate large tree. For this purpose, branches which give less improvement in error cost have been pruned from RT.

RTp is the tree which results in the lowest cost. For the "no uncertainty" scenario, it consists of four leaves and considers only water depth and building value to predict *rloss*. The trees derived from uncertain observations within the frame of the "small uncertainty" and "large uncertainty" scenarios differ from this solution, since the subdivision of the explanatory variable space strives for the minimization of the error for each realization of the data set variations.

Tree-based models are a simple means to multivariate damage modeling, since they permit inclusion of both continuous and categorical variables and they allow for nonlinearities and predictor interactions [*Merz et al*., 2013]. Regression trees can handle incomplete data. In this case, predictions are made by considering only the leaves that can be reached given the available data.

3.4. Data-Based Bayesian Networks

The Bayesian network approach relies on a probabilistic formalism which aims to describe the joint distribution of all variables involved in the system. The number of parameters that is needed to describe the distribution is reduced by decomposing the joint probability into a product of conditional probabilities according to a directed acyclic graph (DAG) capturing probabilistic independencies between the variables. The joint probability of a Bayesian network is given in equation (3)

$$P(A_i, \ldots, A_n) = \prod_{i=1}^{n} P(A_i | parents(A_i)) \tag{3}$$

Where A_i are the variables and *parents* (A_i) denotes the set of parent nodes of the node *Ai* as defined by the DAG.

Using a completely data-driven approach, the graph structure as well as the parameters of the conditional distributions can be learned from data, such that the learning algorithm respects various model complexity issues relating to sample size, overfitting, etc. For a detailed description of the learning procedure we refer to *Vogel et al*. [2012, 2013].

In contrast to other models, the Bayesian network allows use of all 850 records of the Elbe 2002 id subsample for the model derivation, thus exploiting information present in partially observed records. This also includes those cases where *rloss* is unobserved. A Bayesian network treats all quantities involved as random variables and does not distinguish between explanatory and response variables. This enables to capture the joint probability distribution of all variables and to infer in any direction as new evidence, i.e., observations, become available. The network learned for 29 variables (including *rloss*) is shown in Figure 3 (left). Its graph structure gives insight into the (in-) dependency structure of the involved variables (note that this is different from causality).

The gray shaded variables in Figure 3 form the so-called Markov Blanket of *rloss*, which is the minimal set of variables having influence on *rloss*. This means that in this specific case of Bayesian network learning the estimation of *rloss* depends only on six variables: water depth (*wst*), contamination indicator (*con*), inundation duration (*d*), flow velocity indicator (*v*), precautionary measure indicator (*pre*), and perception of efficiency of private precaution (*epre*), and that all other variables can be ignored provided that the Markov Blanket is fully observed. However, if observations of some of these variables are unknown or missing, observations on variables from outside the Markov Blanket provide indirect knowledge "flowing" toward *rloss*, thus helping to improve the prediction thereof.

In Bayesian network learning, in general, we strive to approximate the joint distribution of all variables. Hence, we consider all variables equally important. However, in this study we are particularly interested in the variable *rloss*. Therefore, in another attempt of learning a Bayesian network, we restrict attention to those variable assumed to be highly relevant for the prediction of this target. The selection of these variables is based on available knowledge and experience from previous studies [e.g., *Thieken et al*., 2005; *Merz et al*., 2010]. For *rloss* this amounts to a subset of 10 variables from the original set of 28 explanatory variables. The resulting network is shown in Figure 3 (right) where the Markov Blanket of *rloss* shows that only contamination (*con*), water depth (*wst*), and precaution (*pre*) have direct predictive relevance.

Predictions of any of the variables represented in the Bayesian networks are achieved by inferring the respective conditional probabilities given the observations of other variables. For instance, the prediction of

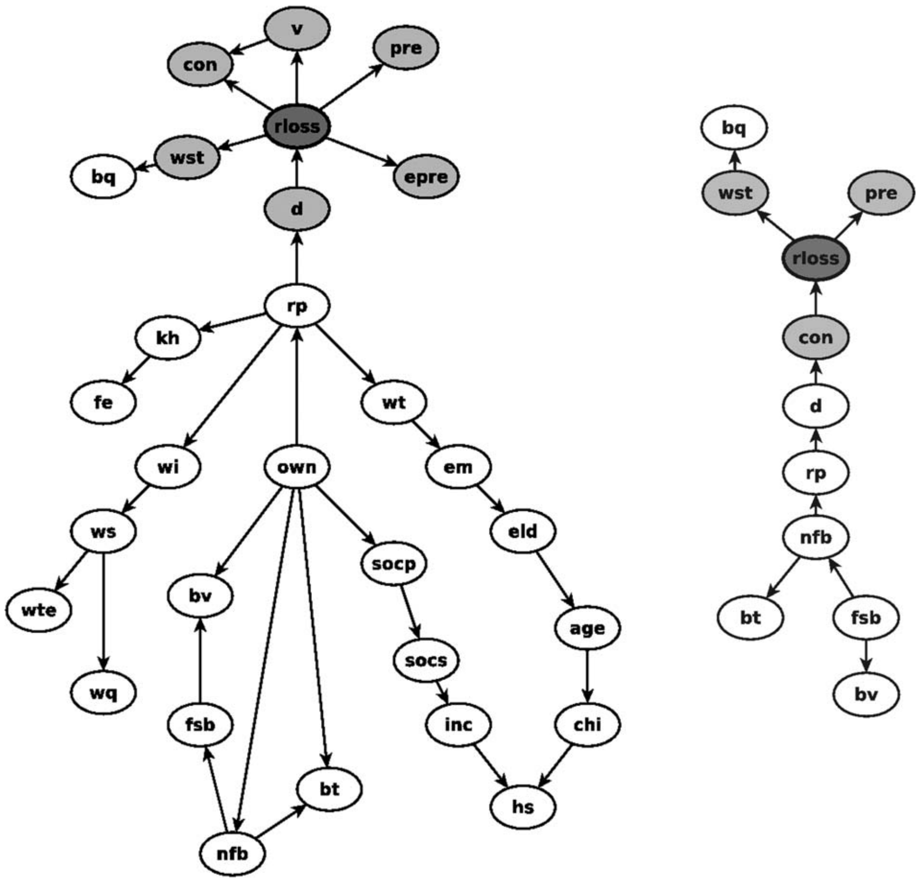

Figure 3. DAG of (left) data-based Bayesian network for 29 explanatory variables including *rloss* and (right) data-based Bayesian network for 11 explanatory variables including *rloss*.

rloss given observations of water depth, return period, and precautionary measures corresponds to the operation of forward inference and is accomplished by marginalization of the conditional probability for *rloss*. In this context, we stress the fundamental property of Bayesian networks providing conditional probabilities of the target variable and thus inherently capturing the uncertainty of the prediction. This is in contrast to the other models examined, which, without the additional efforts for uncertainty analysis, offer only a single deterministic point estimate.

3.5. Expert Bayesian Networks

The construction of a Bayesian network can also incorporate domain or expert knowledge. This knowledge may be included in the definition of the network structure, the direction of the arcs, and the distribution of the parameters. This is of interest because a totally data-driven approach for Bayesian network learning can result in models that capture unwanted artifacts of the data. Especially, when the data set used for model derivation is sparse, those artifacts may overrule physical/causal relationships. The inclusion of expert knowledge into Bayesian network construction might reduce the effect of data anomalies.

For the construction of expert Bayesian networks, we define the graph skeleton based on domain knowledge and learn only the arc directions, the discretization of the ordinal variables and the parameters of the conditional probabilities from the data.

For the definition of the network skeleton, we adopt a causal mapping approach as proposed by *Nadkarni and Shenoy* [2001]. This procedure involves in a first step the derivation of a causal map for the variables of interest. The causal map depicts the cause-effect relations among these variables according to expert knowledge in terms of a directed graph [*Nadkarni and Shenoy*, 2004]. Next, the causal map is modified using the idiomatic introduced by *Fenton and Neil* [2012] in order to construct a Bayesian causal map which

10.1002/2013WR014396

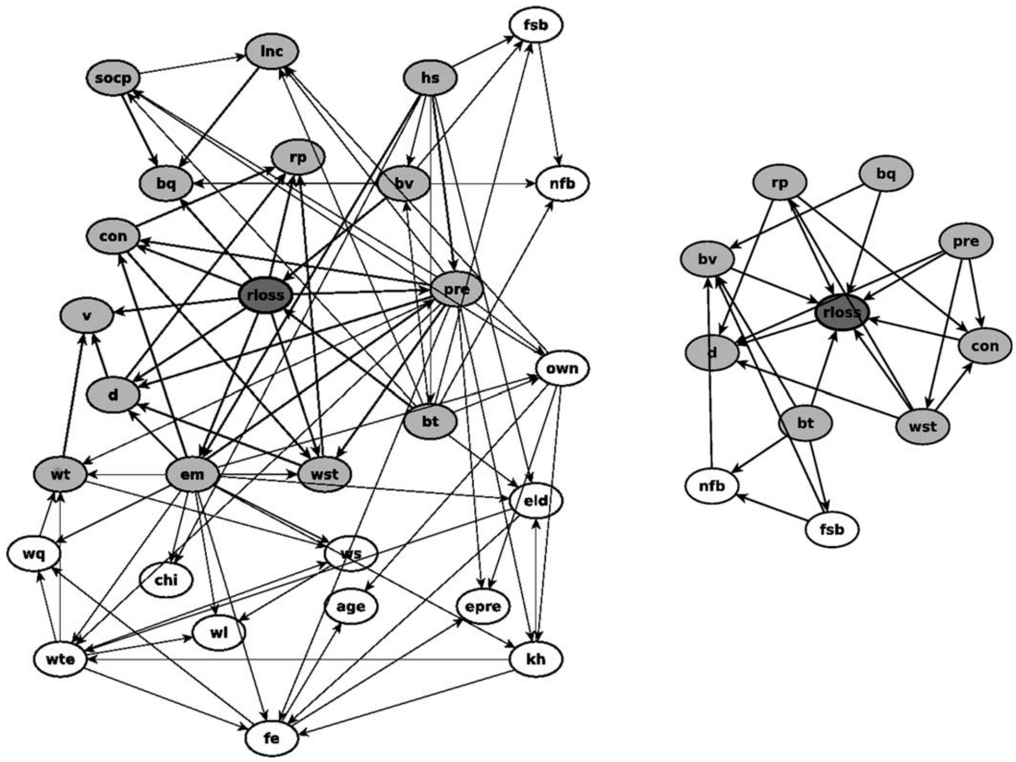

Figure 4. DAG of (left) expert Bayesian network for 28 explanatory variables and (right) expert Bayesian network for 10 explanatory variables constructed with the specific goal to predict *rloss*.

satisfies the requirement of conditional independence among variables and acyclic structure of the graph in order to represent a proper Bayesian network.

The causal map has been derived by using an adjacency matrix which is defined by the 29 candidate variables (including *rloss*) listed in Table 3. This matrix was independently completed by three flood damage experts. In this matrix, the experts indicated whether there is a causal relation between any two variables and defined the direction of these relations. On this basis, a causal map was derived as a directed graph by superimposing the relations identified from the different experts.

Next, within a discussion among the experts any inconsistency in this draft directed graph was reviewed and modified to be compatible with a Bayesian network. This is to ensure that the presence of a link between variables represents dependence and that the lack of a link represents independence between these variables. Further, the presence of direct and indirect relations between subsets of variables was scrutinized and any circular relations were eliminated.

Two alternative versions of expert Bayesian networks are derived: first, accounting for the complete set of candidate explanatory variables and, second, for a subset of 11 of these variables which have been identified to be most informative to predict *rloss* on the basis of existing knowledge and damage modeling experience [*Thieken et al*., 2005; *Kreibich et al*., 2009; *Merz et al*., 2010; *Elmer et al*., 2010]. The resulting networks are shown in Figure 4.

The definition of the network skeleton based on expert knowledge is a nontrivial task. Wrong independence assumptions cannot be corrected by the data and should be avoided. A dense network structure leads to a large number of combinations in the node probability tables. The automatic regularization of network complexity which is ensured in a data-driven approach by means of a structure fitness score [*Riggelsen*, 2008; *Vogel et al*., 2013] is thereby impaired by the network structure imposed. Therefore, the Markov Blanket of *rloss* in an expert Bayesian network will be much larger than in a fully data-driven well-regularized learned Bayesian network. This can be realized from the comparison of the data-based DAGs shown in Figure 3 and the expert DAGs shown in Figure 4.

10.1002/2013WR014396

Table 5. Precision of Damage Model Predictions for "No Uncertainty" Scenario[a]

		Model							
Subsample	Criterion	sdf	FLEMOps+r	RTp	RT	BNe10	BNe28	BNd11	BNd29
Model Derivation									
Elbe 2002 id	mbe	**0.00**	0.02	**0.00**	**0.00**	0.04	0.07	0.08	0.07
	mae	0.76	0.77	0.72	0.67	**0.60**	0.69	0.73	0.71
	ED	0.76	0.77	0.72	0.67	**0.60**	0.69	0.73	0.71
Local Validation									
Elbe 2002 pr	mbe	0.18	0.05	0.15	0.08	**0.03**	0.05	0.18	**0.03**
	mae	0.96	**0.90**	0.94	0.95	1.05	1.00	0.91	0.93
	ED	0.98	**0.90**	0.95	0.95	1.05	1.00	0.93	0.93
Cross-Regional and Temporal Validation									
Danube 2002	mbe	**0.48**	0.72	0.77	0.65	0.75	0.77	0.76	0.51
	mae	1.29	1.29	1.26	1.21	1.47	1.36	1.23	**1.19**
	ED	1.38	1.48	1.48	1.37	1.65	1.56	1.45	**1.29**
Danube 2005	mbe	0.96	1.03	1.08	0.99	0.75	0.77	0.91	**0.42**
	mae	1.71	1.69	**1.51**	**1.51**	1.70	1.60	1.55	1.54
	ED	1.96	1.98	1.86	1.81	1.86	1.78	1.80	**1.60**
Elbe 2006	mbe	1.13	0.92	1.26	1.20	**0.10**	0.57	0.79	0.55
	mae	1.75	1.55	1.55	1.49	1.54	1.35	**1.31**	1.35
	ED	2.08	1.80	2.00	1.91	1.54	1.47	1.53	**1.46**

[a]ED: Euclidian distance to the perfect prediction in the two-dimensional space defined by mbe and mae. Best scores for each criterion and data subsample are marked with bold numbers.

Predictions within the expert Bayesian network follow the same procedure as within the data-based networks.

4. Results and Discussion

We test the hypothesis that increasing complexity improves the predictive capability of flood damage models by comparing the performance of the different models for the data subsamples defined in Table 2 with regard to precision, variation, and reliability.

In the first instance, we examine the scores for the evaluation criteria *mbe* and *mae* within the "no uncertainty" scenario. The scores which are achieved by the models for each subsample are compiled in Table 5 with the best score for each criterion marked bold. Obviously, different models perform best with regard to *mbe* or *mae* within the different subsamples. Noticeably, the magnitude of *mbe* and *mae* increases from the Elbe 2002 id subsample to the Elbe 2002 pr, Danube 2002, Elbe 2006, and Danube 2005 subsamples reflecting the increasing difficulty to predict *rloss* in spatial and temporal transfer applications. At the same time, the differences between the *mae* and *mbe* scores achieved by the different model approaches become more pronounced. While the variations of model performance scores are small for the Elbe 2002 id subsample, they are clearly larger for cross-regional and temporal validation exercises based on the subsamples Danube 2002, Elbe 2006, and Danube 2005. This suggests that the model approach makes a difference in terms of the predictive precision, in particular, in a spatial and temporal transfer context, but is there a relation between model predictive capability and model complexity?

To investigate this question, we use the Euclidean Distance (*ED*) of *mbe* and *mae* to the point of optimum model performance as a multicriteria measure of model predictive precision. *ED* values for the different models and data subsamples are listed in Table 5. We relate *ED* to the number of explanatory variables included by the different models (cf. Table 1). For each model, we calculate the average *ED* out of the results obtained for the subsamples Danube 2002, Danube 2005, and Elbe 2006 which involve a temporal or spatial transfer. The results are plotted in Figure 5 against the number of explanatory variables. This graph illustrates a trade-off between model complexity and predictive performance. However, this relationship is not monotonic, but interfered by differences in the performance of different model approaches. In this regard, for instance, the data-based Bayesian networks utilize fewer explanatory variables to predict *rloss* than the expert Bayesian networks but still, on average, provide predictions with higher precision. Comparing alternatives of different complexity within similar modeling approaches, e.g., the pruned regression tree and the complete regression tree, the data based or expert Bayesian networks based on a reduced number of

10.1002/2013WR014396

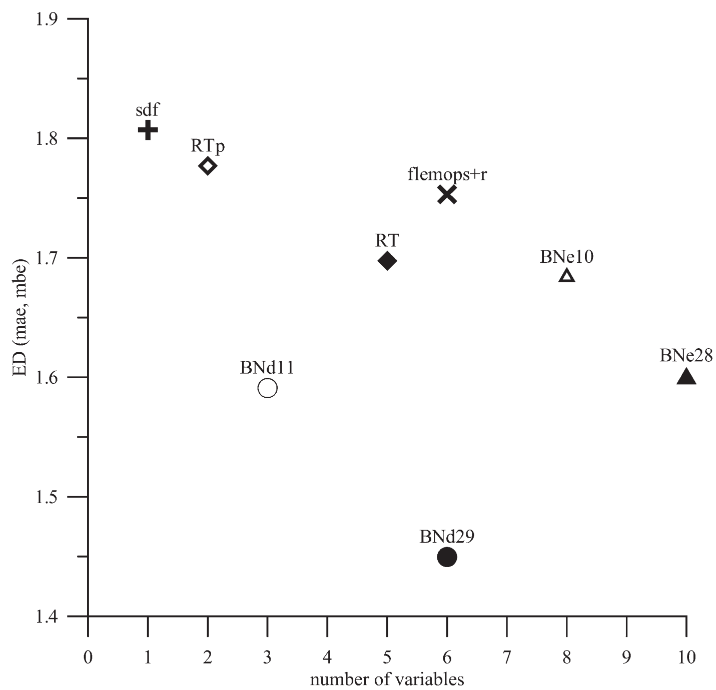

Figure 5. Average model predictive precision ED (*mbe* and *mae*) in cross-regional and temporal validations against model complexity (number of explanatory variables used to predict *rloss*) for "no uncertainty" scenario.

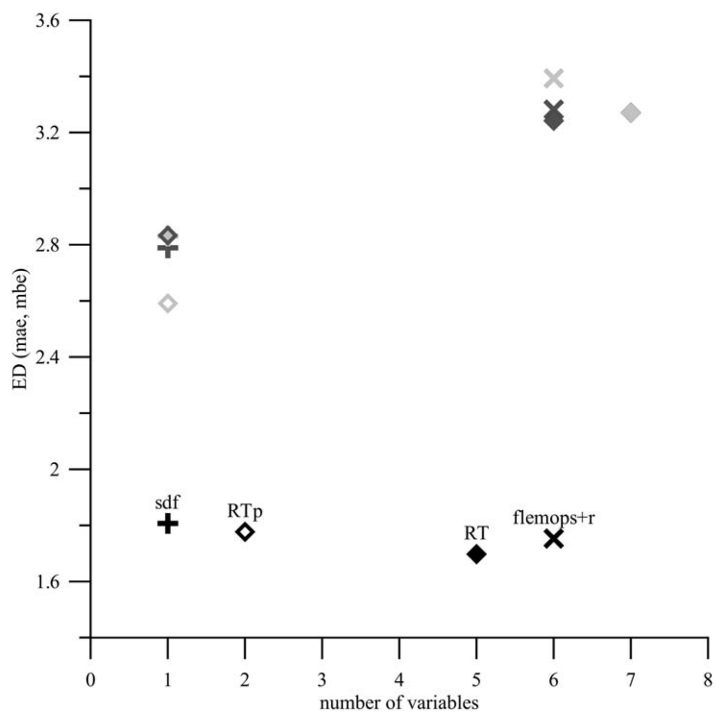

Figure 6. Average model predictive precision ED (*mbe* and *mae*) of deterministic models in cross-regional and temporal validations against model complexity (number of explanatory variables used to predict *rloss*) within the uncertainty scenarios. Model approaches are described by different symbols; uncertainty scenarios are represented by different colors: black (no uncertainty), gray (small uncertainty), light gray (large uncertainty).

explanatory variables and the complete set of variables, we recognize that the more complex variant performs better in any case.

However, the results obtained for the deterministic models are based on the assumption that the explanatory variables are observed without uncertainty. The implications of potential observations uncertainty on the predictive performance are shown in Figure 6. Essentially, two effects are apparent. First, the noise added to the observations used for model derivation propagates to the model predictions and impairs the precision of the models. Unsurprisingly, the more variables are used to predict *rloss*, the larger the decline of predictive precision, since with additional variables additional sources of uncertainty take effect. Accordingly, RTp and the sdf model achieve best predictive precision given uncertain observations. Second, observation uncertainty influences the structure of the regression trees derived from the data and hence the complexity of the resulting model. Mostly, RTp models trained with uncertain observations include only one variable defining a single branch. In contrast, RT models tend to include more variables in the tree structure than without considering observation uncertainty (within the sample of models, on average six variables ($\pm$4 based on IQR) are assuming "small uncertainty," and seven variables ($\pm$5 based on IQR) are used assuming "large uncertainty"). In this context, the regression trees derived within "no uncertainty" scenario have to be thought of as a single realization of the sample of models considered within the uncertainty scenarios.

Table 6. Variation and Reliability of Predictive Distributions of Deterministic Models (sdf, flemops+r, RTp, RT) Within Small and Large Uncertainty Scenarios (su, lu) and Joint Probability Distributions of Bayesian Network (BN) Variants

		Model											
		sdf		FLEMOps+r		RTp		RT					
Subsample	Criterion	su	lu	su	lu	su	lu	su	lu	BNe10	BNe28	BNd11	BNd29
Model Derivation													
Elbe 2002 id	QR_{90}	0.04	0.16	0.28	0.13	0.30	0.00	0.72	0.41	1.41	1.48	1.47	1.41
	HR	0.05	0.13	0.16	0.57	0.25	0.00	0.50	0.41	0.96	0.94	0.92	0.92
Local Validation													
Elbe 2002 pr	QR_{90}	0.01	0.03	0.37	0.65	0.32	0.00	0.66	0.63	1.39	1.41	1.38	1.30
	HR	0.00	0.02	0.13	0.44	0.26	0.00	0.46	0.42	0.86	0.87	0.86	0.85
Cross-Regional and Temporal Validation													
Danube 2002	QR_{90}	0.01	0.03	0.35	0.44	0.42	0.00	0.70	0.64	1.20	1.41	1.20	1.10
	HR	0.00	0.01	0.09	0.15	0.16	0.00	0.37	0.33	0.73	0.83	0.83	0.83
Danube 2005	QR_{90}	0.01	0.03	0.33	0.52	0.40	0.00	0.68	0.60	1.25	1.36	1.27	1.10
	HR	0.00	0.01	0.09	0.15	0.16	0.00	0.31	0.28	0.62	0.64	0.69	0.66
Elbe 2006	QR_{90}	0.01	0.03	0.36	0.36	0.33	0.00	0.68	0.56	1.16	1.39	1.28	1.16
	HR	0.00	0.00	0.08	0.08	0.12	0.00	0.28	0.23	0.63	0.67	0.67	0.65

To compare the variation and reliability of the model predictions (using the criteria QR_{90} and HR), we evaluate the predictive distributions of *rloss* for the different damage models, whereat we discern, on the one hand, the predictive distributions obtained for the deterministic models for the small and large uncertainty scenarios and, on the other hand, the joint probability distributions of *rloss* from the Bayesian networks. Table 6 documents QR_{90} and HR scores for the different subsamples and models. For the deterministic models, sdf and FLEMOps+r, QR_{90} increases with the degree of uncertainty added to the observations. This relation is reverse for the RTp and RT models. For RTp this is due to the difficulty to derive meaningful tree structures from increasingly uncertain observations which results in very simple model structures, and in turn zero variability of model predictions with $QR_{90} = 0$ and $HR = 0$. For the BN model variants QR_{90} is clearly larger (by a factor of 2). In this regard, we recall that the joint probability distribution inferred by BN reflects both the probabilistic dependence of the variables and the observation uncertainty associated with the variables. Overall, the variability of model predictions increases with model complexity since with additional variables additional sources of uncertainty take effect. Hence, more complex models tend to provide more variable predictions.

For the prediction interval to be of use it should be reliable. The HR criterion quantifies the proportion of the observations that fall within the prediction interval. The HR scores are listed in Table 6. Obviously, HR increases with QR_{90}. Among the deterministic models RT provides the most reliable predictions. Further, the HR scores provide evidence to which extent the uncertainty associated with the damage estimation is represented by the model. According to this interpretation, the closer HR is to the nominal coverage (0.9 for the 95–5 quantile range), the better the representation of the uncertainty. In this regard, the Bayesian networks apparently embrace the prediction uncertainty more completely than the deterministic models. This is comprehensible since for the derivation of the deterministic models the uncertainty in the observations has been considered, but the uncertainty concerning the probabilistic dependence of variables is not accounted for.

Figure 7 relates the average HR achieved by the different models in cross-regional and temporal validations to the number of explanatory variables. This graph illustrates that the reliability of model predictions seems to depend more on the model approach and the underlying concept to handle predictive uncertainty than on model complexity. In this regard, the largest values for HR are achieved by BNd11 using three explanatory variables to predict *rloss*. Overall, the reliability of BN-based predictions in space-time transfer is larger than for the remaining models, and the uncertainties associated with damage predictions are reflected more completely. For these models the average HR is quite close to the nominal coverage of 0.9 for the 95–5 quantile range, and, hence, the joint probability distribution of *rloss* describes the predictive uncertainty relatively well.

5. Conclusions

In this paper, we investigated the usefulness of complex flood damage models for the improvement of predicting relative damage to residential buildings. The results confirm the hypothesis that increasing

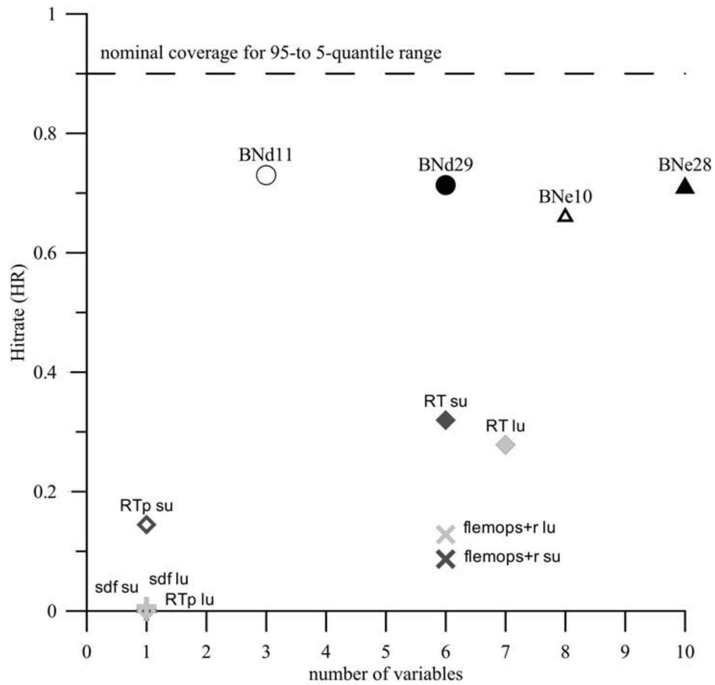

Figure 7. Average predictive reliability HR in cross-regional and temporal validations against model complexity (number of explanatory variables used to predict rloss) for the predictive distributions of deterministic models within uncertainty scenarios (su, lu) and joint probability distribution of Bayesian networks. Model approaches are described by different symbols; uncertainty scenarios applied to deterministic models are represented by different colors: gray (small uncertainty) and light gray (large uncertainty).

complexity of flood damage models improves the capability to predict flood damage. In particular, this applies to model transfer applications to different regions and different flood events. Using additional explanatory variables besides water depth improves the precision of predictions assuming that there is "no uncertainty" in the observed explanatory variables. However, the relation between predictive capability and model complexity is not monotonic but is interfered by differences in the performance of different model approaches. In this regard, using the number of explanatory variables as a proxy for model complexity obviously falls short in capturing the various facets of model complexity, as, for instance, the representation of interactions between explanatory variables and their functional form implemented within the different model approaches. Still, it is shown that the more complex variants of similar model approaches outperform the simpler alternatives.

Uncertainty is of high relevance in flood damage modeling as all models show difficulties in completely explaining the real damage processes given the damage data available. The analysis of different uncertainty scenarios has shown that observation uncertainty can considerably impair the predictive performance of the deterministic models and may impede the derivation of appropriate model structures. In relation to model complexity, on the one hand, the use of additional explanatory variables incorporates additional knowledge, but, on the other hand, as these observations are uncertain, it also introduces additional uncertainty. In the light of the magnitude of model prediction errors, it is mandatory to quantify the uncertainty of model predictions. For a realistic estimation of model predictive uncertainty, not only observation uncertainty but also other uncertainty sources, e.g., model structure uncertainty, have to be taken into account. In this regard, probabilistic model approaches, as, for instance, Bayesian networks, provide a consistent framework to comprehensively consider uncertainty. The results obtained for the hit-rate in relation to the nominal coverage show that the joint probability distribution of damage estimates provided by Bayesian networks represents the prediction uncertainty very well given the damage data available.

Overall, for the improvement of flood damage predictions more complex models including more details about the damaging process are useful. However, the application of these models requires a sufficient amount of data and a detailed and structured acquisition of explanatory variables preferably gathered within the study region and hence representing local characteristics. Despite this, the variability of damage records and related explanatory variables will remain considerable; thus, the uncertainty of flood damage predictions must be analyzed, quantified, and communicated.

References

Apel, H., G. Aronica, H. Kreibich, and A. Thieken (2009), Flood risk analyses—How detailed do we need to be?, *Nat. Hazards*, *49*(1), 79–98, doi:10.1007/s11069-008-9277-8.

Barredo, J. I. (2009), Normalised flood losses in Europe: 1970–2006, *Nat. Hazards Earth Syst. Sci.*, *9*(1), 97–104.

Bubeck, P., and H. Kreibich (2011), *Natural Hazards: Direct Costs and Losses Due to the Disruption of Production Processes*, Ger. Res. Cent. for Geosci. —GFZ, Sect. Hydrol., Potsdam, Germany.

Büchele, B., H. Kreibich, A. Kron, A. Thieken, J. Ihringer, P. Oberle, B. Merz, and F. Nestmann (2006), Flood-risk mapping: Contributions towards an enhanced assessment of extreme events and associated risks, *Nat. Hazards Earth Syst. Sci.*, *6*(4), 485–503, doi:10.5194/nhess-6–485-2006.

Buck, W., and U. Merkel (1999), *Auswertung der HOWAS-Schadendatenbank*, Univ. Karlsruhe, Inst. für Wasserwirt. und Kulturtechnik, Karlsruhe, Germany.

Cammerer, H., A. H. Thieken, and J. Lammel (2013), Adaptability and transferability of flood loss functions in residential areas, *Nat. Hazards Earth Syst. Sci.*, *13*(11), 3063–3081, doi:10.5194/nhess-13-3063-2013.

Ding, A., J. F. White, P. W. Ullman, and A. O. Fashokun (2008), Evaluation of HAZUS-MH flood model with local data and other program, *Nat. Hazards Rev.*, *9*(1), 20–28, doi:10.1061/(ASCE)1527-6988(2008)9:1(20).

de Moel, H., and J. C. J. H. Aerts (2011), Effect of uncertainty in land use, damage models and inundation depth on flood damage estimates, *Nat. Hazards*, *58*(1), 407–425, doi:10.1007/s11069-010-9675-6.

Downton, M. W., and R. A. Pielke (2005), How accurate are disaster loss data? The case of U.S. flood damage, *Nat. Hazards*, *35*(2), 211–228, doi:10.1007/s11069-004-4808-4.

Dutta, D., S. Herath, and K. Musiakec (2003), A mathematical model for flood loss estimation, *J. Hydrol.*, *277*(1–2), 24–49, doi:10.1016/S0022-1694(03)00084-2.

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

EU (2007), Directive on the assessment and management of flood risks, *Off. J. Eur. Union*, Directive 2007/60/EC *L288*.

Fenton, N., and M. Neil (2012), *Risk Assessment and Decision Analysis With Bayesian Networks*, CRC Press Taylor and Francis Group, Boca Raton, FL, USA.

Green, C. H. (2003), *The Handbook of Water Economics: Principles and Practice*, John Wiley, Chichester, England.

IKSE (2004), *Dokumentation des Hochwassers vom August 2002 im Einzugsgebiet der Elbe*, International Commission for the Protection of the Elbe, Magdeburg.

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

Kang, J.-L., M.-D. Su, and L.-F. Chang (2005), Loss functions and framework for regional flood damage estimation in residential area, *J. Mar. Sci. Technol.*, *13*(3), 193–199.

Kelman, I., and R. Spence (2004), An overview of flood actions on buildings, *Eng. Geol.*, *73*(3–4), 297–309, doi:10.1016/j.enggeo.2004.01.010.

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

Kreibich, H., and A. H. Thieken (2008), Assessment of damage caused by high groundwater inundation, *Water Resour. Res.*, *44*, W09409, doi:10.1029/2007WR006621.

Kreibich, H., and A. H. Thieken (2009), Coping with floods in the city of Dresden, Germany, *Nat. Hazards*, *51*(3), 423–436, doi:10.1007/s11069-007-9200-8.

Kreibich, H., K. Piroth, I. Seifert, H. Maiwald, U. Kunert, J. Schwarz, B. Merz, and A. H. Thieken (2009), Is flow velocity a significant parameter in flood damage modelling?, *Nat. Hazards Earth Syst. Sci.*, *9*(5), 1679–1692, doi:10.5194/nhess-9-1679-2009.

Kreibich, H., I. Seifert, A. H. Thieken, E. Lindquist, K. Wagner, and B. Merz (2011), Recent changes in flood preparedness of private households and businesses in Germany, *Reg. Environ. Change*, *11*(1), 59–71, doi:10.1007/s10113-010-0119-3.

Kron, W., M. Steuer, P. Löw, and A. Wirtz (2012), How to deal properly with a natural catastrophe database—Analysis of flood losses, *Nat. Hazards Earth Syst. Sci.*, *12*(3), 535–550, doi:10.5194/nhess-12–535-2012.

LfU (2006), *August-Hochwasser 2005 in Südbayern*, Bayerisches Landesamt für Umwelt, Augsburg.

Merz, B., and A. H. Thieken (2005), Separating natural and epistemic uncertainty in flood frequency analysis, *J. Hydrol.*, *309*, 114–132.

Merz, B., H. Kreibich, A. Thieken, and R. Schmidtke (2004), Estimation uncertainty of direct monetary flood damage to buildings, *Nat. Hazards Earth Syst. Sci.*, *4*(1), 153–163, doi:10.5194/nhess-4-153-2004.

Merz, B., H. Kreibich, R. Schwarze, and A. Thieken (2010), Review article "Assessment of economic flood damage", *Nat. Hazards Earth Syst. Sci.*, *10*(8), 1697–1724, doi:10.5194/nhess-10-1697-2010.

Merz, B., H. Kreibich, and U. Lall (2013), Multi-variate flood damage assessment: A tree-based data-mining approach, *Nat. Hazards Earth Syst. Sci.*, *13*(1), 53–64, doi:10.5194/nhess-13–53-2013.

Munich Re (2009), NatCatSERVICE. [Available at www.munichre.com, accessed 20 Nov 2009.] Munich.

Nadkarni, S., and P. P. Shenoy (2001), A Bayesian network approach to making inferences in causal maps, *Eur. J. Oper. Res.*, *128*(3), 479–498, doi:10.1016/S0377–2217(99)00368-9.

Nadkarni, S., and P. P. Shenoy (2004), A causal mapping approach to constructing Bayesian networks, *Decis. Support Syst.*, *38*(2), 259–281, doi:10.1016/S0167–9236(03)00095-2.

NRC (2000), *Risk Analysis and Uncertainty in Flood Damage Reduction Studies*, Natl. Acad. Press, Washington, D. C.

Penning-Rowsell, E., and J. B. Chatterton (1977), *The Benefits of Flood Alleviation: A Manual of Assessment Techniques*, Saxon House, Farnborough, England.

Penning-Rowsell, E. C., and C. Green (2000), New insights into the appraisal of flood-alleviation benefits: (1) Flood damage and flood loss information, *Water Environ. J.*, *14*(5), 347–353, doi:10.1111/j.1747–6593.2000.tb00272.x.

Poser, K., and D. Dransch (2010), Volunteered geographic information for disaster management with application to rapid flood damage estimation, *Geomatica*, *64*(1), 89–98.

Power, M. (1993), The predictive validation of ecological and environmental models, *Ecol. Modell.*, *68*(1–2), 33–50.

Riggelsen, C. (2008), Learning Bayesian networks: A MAP criterion for joint selection of model structure and parameter, in *Eighth IEEE International Conference on Data Mining*, 2008, ICDM '08, pp. 522–529, Piscataway, New Jersey, USA.

Saltelli, A., K. Chan, and E. M. Scott (2000), *Sensitivity Analysis, Probab. Stat.*, John Wiley, Chichester, U. K.

Scawthorn, C., N. Blais, H. Seligson, E. Tate, E. Mifflin, W. Thomas, J. Murphy, and C. Jones (2006), HAZUS-MH flood loss estimation methodology. I: Overview and flood hazard characterization, *Nat. Hazards Rev.*, *7*(2), 60–71, doi:10.1061/(ASCE)1527–6988(2006)7:2(60).

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

Smith, D. I. (1994), Flood damage estimation—A review of urban stage-damage curves and loss functions, *Water SA*, *20*(3), 231–238.

Thieken, A. H., M. Müller, H. Kreibich, and B. Merz (2005), Flood damage and influencing factors: New insights from the August 2002 flood in Germany, *Water Resour. Res.*, *41*, W12430, doi:10.1029/2005WR004177.

Thieken, A., H. Kreibich, M. Müller, and B. Merz (2007), Coping with floods: Preparedness, response and recovery of flood-affected residents in Germany in 2002, *Hydrol. Sci. J.*, *52*(5), 1016–1037, doi:10.1623/hysj.52.5.1016.

Thieken, A. H., A. Olschewski, H. Kreibich, S. Kobsch, and B. Merz (2008), Development and evaluation of FLEMOps—A new flood loss estimation model for the private sector, in *Flood Recovery, Innovation and Response*, pp. 315–324, WIT Press, London.

Thordarson, F. Ö., A. Breinholt, J. K. Møller, P. S. Mikkelsen, M. Grum, and H. Madsen (2012), Evaluation of probabilistic flow predictions in sewer systems using grey box models and a skill score criterion, *Stochastic Environ. Res. Risk Assess.*, *26*(8), 1151–1162, doi:10.1007/s00477-012-0563-3.

UNISDR (2011), *Global Assessment Report on Disaster Risk Reduction—Revealing Risk, Redefining Development*, United Nations, Geneva, Switzerland.

Vogel, K., C. Rigg, H. Kreibich, B. Merz, and F. Scherbaum (2012), Flood damage and influencing factors: A Bayesian network perspective, paper presented at the 6th European Workshop on Probabilistic Graphical Models (PGM 2012), University of Granada, Granada, Spain.

Vogel, K., C. Riggelsen, F. Scherbaum, K. Schröter, H. Kreibich, and B. Merz (2013), Challenges for Bayesian network learning in a flood damage assessment application, paper presented at 11th International Conference on Structural Safety and Reliability (ICOSSAR 2013), Columbia University, New York.

Vogt, J., et al. (2007), *A Pan-European River and Catchment Database*, Eur. Comm. Luxembourg.

Vorogushyn, S., K.-E. Lindenschmidt, H. Kreibich, H. Apel, and B. Merz (2012), Analysis of a detention basin impact on dike failure probabilities and flood risk for a channel-dike-floodplain system along the river Elbe, Germany, *J. Hydrol.*, *436–437*, 120–131, doi:10.1016/j.jhydrol.2012.03.006.

White, G. I. (1945), Human adjustments to floods, research paper, Dep. of Geogr., Univ. of Chicago, Chiccago, Ill.

Willmott, C. J., and K. Matsuura (2005), Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance, *Clim. Res.*, *30*(1), 79–82, doi:10.3354/cr030079.

Wind, H. G., T. M. Nierop, C. J. de Blois, and J. L. de Kok (1999), Analysis of flood damages from the 1993 and 1995 Meuse Floods, *Water Resour. Res.*, *35*(11), 3459–3465, doi:10.1029/1999WR900192.

Zhai, G., T. Fukuzono, and S. Ikeda (2005), Modeling flood damage: Case of Tokai flood 2000, *J. Am. Water Resour. Assoc.*, *41*(1), 77–92.