Nat. Hazards Earth Syst. Sci., 13, 53–64, 2013
www.nat-hazards-earth-syst-sci.net/13/53/2013/
doi:10.5194/nhess-13-53-2013

Natural Hazards and Earth System Sciences

Multi-variate flood damage assessment: a tree-based data-mining approach

B. Merz[1], **H. Kreibich**[1], **and U. Lall**[2]

[1]GFZ German Research Centre for Geosciences, Section 5.4, 14473 Potsdam, Germany
[2]Columbia University, Department of Earth & Environmental Engineering, New York, NY 10027, USA

Correspondence to: B. Merz (bmerz@gfz-potsdam.de)

Received: 10 August 2012 – Revised: 3 December 2012 – Accepted: 6 December 2012 – Published: 11 January 2013

Abstract. The usual approach for flood damage assessment consists of stage-damage functions which relate the relative or absolute damage for a certain class of objects to the inundation depth. Other characteristics of the flooding situation and of the flooded object are rarely taken into account, although flood damage is influenced by a variety of factors. We apply a group of data-mining techniques, known as tree-structured models, to flood damage assessment. A very comprehensive data set of more than 1000 records of direct building damage of private households in Germany is used. Each record contains details about a large variety of potential damage-influencing characteristics, such as hydrological and hydraulic aspects of the flooding situation, early warning and emergency measures undertaken, state of precaution of the household, building characteristics and socio-economic status of the household. Regression trees and bagging decision trees are used to select the more important damage-influencing variables and to derive multi-variate flood damage models. It is shown that these models outperform existing models, and that tree-structured models are a promising alternative to traditional damage models.

1 Introduction

Flood risk management has to be built upon a sound assessment of flood hazard and flood vulnerability which includes the estimation of damage and effectiveness of different mitigation measures. Compared to the wealth of methods and available information on flood hazard, reliable flood damage data are scarce (Handmer et al., 2005; Gall et al., 2009), understanding of the damaging processes is weak (Bubeck and Kreibich, 2011) and damage estimation methods are crude (Merz et al., 2010).

In this paper, we analyze direct damage to residential buildings. Direct flood damage depends on many factors, in particular water depth, but also early warning, flood experience and precautionary measures. These factors may not be independent from each other. For example, the damage-reducing effect of early warning depends on the preparedness of the affected people which in turn may depend on their flood experience. Further, flood experience is expected to influence also the state of precautionary measures of the flooded household. Other factors are, for instance, flow velocity, duration of inundation, contamination of flood water, and the quality of external response in a flood situation. The single and joint effects of these parameters on the damage are largely unknown and widely neglected in damage assessment. Exceptions are, for instance, Wind et al. (1999) who investigate the influence of flood warning time and flood experience on damage at the municipality level. Penning-Rowsell and Green (2000) present an equation to estimate flood damage avoided because of warning with the following parameters: "proportion of the population at risk which is warned with sufficient lead time to take action", "proportion of residents available to respond to a warning", "proportion of residents able to respond to a warning" and "proportion of households who respond effectively"; Parker et al. (2007) use this approach and evaluate with a broader perspective the factors influencing the benefits of flood warning including intangible benefits to public health, safety and security.

Traditional flood damage models are stage-damage functions that are solely based on the type or use of an element at risk and the water depth (Merz et al., 2010). Recently some multi-parameter models have been developed: a conceptual

model only suggesting which parameters should be considered in flood damage estimation without quantifying their effect on the damage has been developed in the UK (Nicholas et al., 2001). Zhai et al. (2005) developed a multi-variate regression model with inundation depth, house ownership, house structure, length of residence and household income to estimate losses in private households in Japan. To our knowledge, this model has not been validated or compared with other models, so its uncertainty is unknown.

After recent floods in Germany, we have made significant efforts to investigate the damaging processes, to identify the important damage-determining parameters, and to develop damage models. An extensive set of detailed, object-specific flood damage data was collected by computer-aided telephone interviews with 2158 households and 642 companies after the 2002 and 2005/2006 floods in the Elbe and Danube catchments (Thieken et al., 2007; Kreibich et al., 2007, 2011). Thieken et al. (2005) investigated single and joint effects of impact factors (i.e. flood characteristics like inundation depth) and resistance factors (i.e. characteristics of exposed elements to resist a flood, like type or structure of a building) on flood damage to private households. The investigation revealed that flood impact variables, particularly water depth, flood duration and contamination are the most influential factors for building and for content damage, followed by items quantifying the size and the value of the affected building/flat. Kreibich et al. (2005, 2007) quantified the positive damage-reducing effects of different precautionary measures for residential buildings and companies. The general consideration of flow velocity in flood damage modeling, particularly for estimating monetary loss, was not supported (Kreibich et al., 2009). Based on these results, multi-parameter flood loss estimation models for private households and companies (FLEMOps, FLEMOcs) have been developed, applied and validated at the micro- and meso-scale (Büchele et al., 2006; Thieken et al., 2008; Kreibich and Thieken, 2008; Kreibich et al., 2010; Elmer et al., 2010). For instance, FLEMOps calculates the damage ratio for private households using five different classes of inundation depth, three individual building types, two classes of building quality, there classes of contamination and three classes of private precaution (Thieken et al., 2008). Elmer et al. (2010) identified the return period of the inundation at the affected residential building as an important damage determinant and included the return period (divided into three classes) as additional parameter (FLEMOps+r).

These analyses suggest that multi-variate models that take several damage-influencing parameters into account can improve flood damage modeling. There is a need for multivariate statistical analyses of comprehensive flood damage data to quantify the interaction and influence of various factors and to further develop reliable damage models. Against this background, we test in this paper, if and to which extent tree-based methods can contribute to a better understanding of damage processes and better flood damage estimation. Well-known variants of tree-based models are classification and regression trees (CART) for categorical predictor variables (classification tree) or predicting continuous dependent variables (regression tree), respectively (Breiman et al., 1984).

A traditional approach is likely to consider a generalized linear modeling framework in which interactions across variables are considered at best through product terms. However, such interactions may only be important if certain thresholds (typically unknown a priori) are crossed. The regression tree models applied in this paper attempt to identify statistically meaningful thresholds, and their ordering that best explains the variance in flood damage. Such an approach can lead to a proliferation of parameters and choice across competing model structures. The selection of a single "best" model in this context is difficult to justify. Consequently, "bagging" approaches (bootstrap aggregation, see Breiman, 1996) have been developed to consider an ensemble of such models, and their pooling to reduce uncertainty in prediction. Both ideas are explored here.

Data mining or machine learning aims at discovering patterns, classifying data or understanding relationships in usually large data sets. We apply decision-tree learning whereas tree-based structures are derived from the data. The central idea of tree-based models is to recursively split the data space into sub-spaces according to the behavior of a response variable. The succession of binary splits leads to a set of tree branches subdividing the data space into disjoint partitions of the response variable (leaves). The splits are made in such a way that the homogeneity or purity of the response variable in the leaves is maximized.

An advantage for many problems is the non-parametric and non-linear nature of tree-based models. There are no assumptions concerning the relationship between predictors and response variable, and non-linear and non-monotonic dependencies can be represented by a tree. A further advantage is their ability to assess the local behavior of the response variable. Parametric regression needs to identify relationships which hold globally, across the complete data space. When the data consist of many features which interact in complex ways, a single global model may not be found or it may be very complex. Tree-based models take an alternative route, by dividing the data space successively in subdivisions until these subdivisions are so tame that a simple model can be fit to them. Tree-based methods are particularly well suited when there is little knowledge about how the predictor variables and the response variable relate to each other. However, they need large data sets in order to detect complex patterns and relationships.

Tree-based data analysis and modeling is finding increased attention in hydrology and water resources research. It has been used in rainfall–runoff analysis, namely to detect thresholds in hydro-meteorological variables corresponding to switching conditions between catchment response types (Ali et al., 2010), and to identify topographic controls on

overland flow generation (Loos and Elsenbeer, 2011). In these cases, tree-based models have been applied to multivariate data sets where interactions between parameters and threshold processes played a role. Similarly, Carlisle et al. (2010) applied tree-based models to predict the value for 13 metrics of the magnitude, frequency, duration, timing and rate of change of streamflow given watershed characteristics, and Mototch et al. (2005) investigated the relationship between the spatial distribution of snow water equivalent and landscape properties in an alpine catchment. Grunwald et al. (2009) modeled successfully the relationship between phosphor load and environmental predictor variables for ten farms. Solomatine and Dulal (2003) and Iorgulescu and Beven (2004) used tree-based models to simulate the rainfall–runoff behavior. Pappenberger et al. (2006) incorporated multiple regression trees in a method for sensitivity analysis. Recently, tree-based models have been used for forecasting seasonal streamflow and for evaluating large-scale climate indices for their potential as streamflow predictor (Wei et al., 2011).

We hypothesize that tree-based models are an effective alternative to traditional flood damage analysis and modeling approaches where nonlinearity and parameter interactions play a role. Hence, the purpose of the paper is twofold: (1) deriving important damage-influencing variables and relationships between parameters by applying tree-based data mining methods to a comprehensive flood damage data set, and (2) establishing tree-based damage prediction models and comparing their performance to established models. The analysis is limited to direct building damage of private households.

2 Data

We use the flood damage data set that has been compiled after the flood of 2002 and the floods of 2005 and 2006 in the Elbe and Danube catchments in Germany (Thieken et al., 2007; Kreibich et al., 2011). This data set is based on telephone interviews with flood affected private households. Lists were compiled of all streets affected with the help of information from local authorities, flood reports or press releases as well as with the help of flood masks derived from radar satellite data (DLR, Centre for Satellite Based Crisis information, www.zki.dlr.de). This provided the basis for generating random samples of households. The SOKO institute for social research and communication (www.soko-institut.de) conducted the telephone interviews in April and May 2003. The Explorare Market Research Institute (www.explorare.de) conducted the interviews in November and December 2006. The person in the household with the best knowledge of the flood damage was interviewed. The survey after the 2002 flood resulted in 1697 completed interviews, the survey in 2006 resulted in 461 interviews. The questionnaires addressed the following topics: hydrological and hydraulic aspects of the flooding situation, early warning and emergency measures undertaken, state of precaution of the household, building characteristics, socio-economic status of the household and flood damage to buildings and contents. The questionnaire contained detailed questions addressing not only total damage but also the area affected per story, the damage ratio, the type and amount of the most expensive item damaged, and the type and costs of all building repairs and all expensive domestic appliances affected. This generated the most accurate information possible about the extent of damage, avoiding a strategic response bias. Cross checks of answers also during the interview were undertaken to improve data quality, since it allowed clarification of contradictory answers. Since many people claimed their damages either from government funds or from their insurers, the damage estimates are relatively reliable. This was also confirmed by a comparison of the damage data collected after the 2002 flood with official damage data from the Saxon Bank of Reconstruction which was responsible for administering governmental disaster assistance after the 2002 flood in the federal state of Saxony (Thieken et al., 2005).

Data from 2158 interviews with flood-affected private households are available for this analysis. The raw data were supplemented by estimates of building values, loss ratio L_R, i.e. the relation between the building damage and its value, and indicators for flow velocity, contamination, flood warning, emergency measures, precautionary measures, flood experience and socio-economic variables. For instance, the interviewees were asked if they had undertaken different precautionary measures, i.e. two informational measures like gathering information about precautionary measures and joining neighborhood flood networks, flood insurance and six different building precautionary measures, e.g. flood adapted building use, sealing of the building, purchase of water barriers (open answers and multiple answers were possible). On basis of this information, the precautionary measures indicator was developed, taking into account the type of precaution and how many precautionary measures have been applied (for more details see Thieken et al. (2005)). The loss ratio could be provided for 1103 cases. Since the loss ratio is the variable of interest, the data analysis is limited to these 1103 households.

As input for the data mining analysis, 28 candidate predictors (Table 1) for the predictand "loss ratio of residential building (continuous between $0 =$ no damage and $1 =$ total damage)" were used. The predictors were selected according to previous analyses: four predictors related to the hydrological and hydraulic aspects of the flooding situation at the affected building (Table 4 in Thieken et al., 2005), one predictor is the return period at the affected building (Elmer et al., 2010), ten predictors related to damage reduction particularly to early warning and emergency measures undertaken as well as to state of precaution of the household (Table 5 in Thieken et al., 2005) and 13 predictors related to the

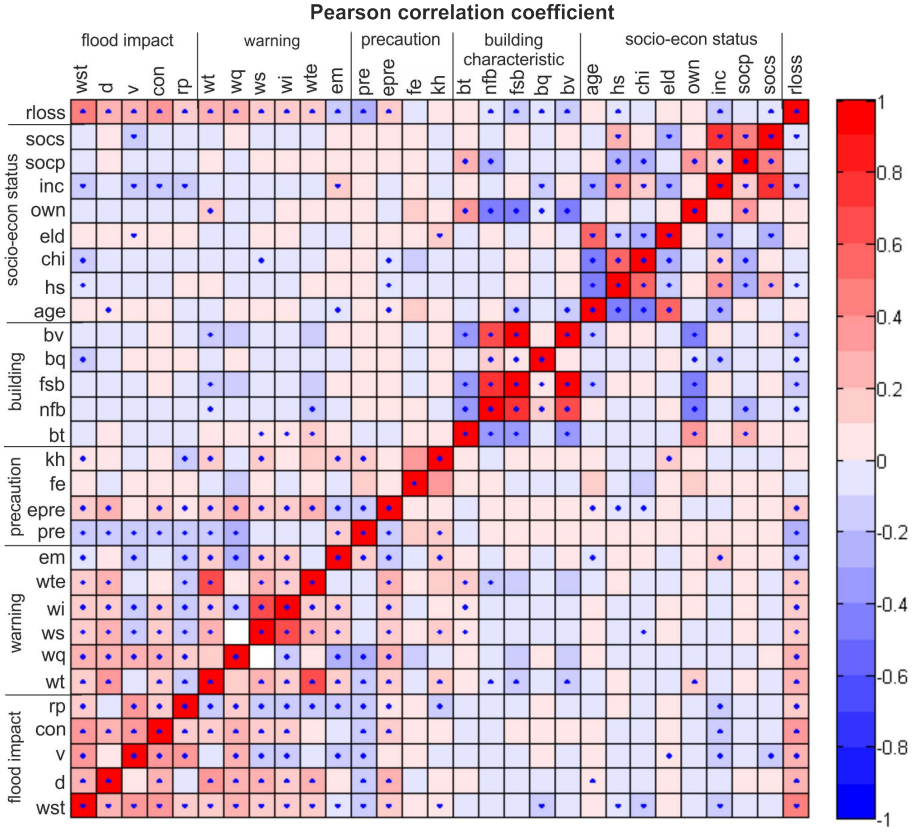

Fig. 1. Pearson correlation of the 29 variables (28 candidate predictors, see Table 1, and loss ratio). Significant correlation (1 % significance level) is marked by a dot.

residential building characteristics and socio-economic status of the household (Table 6 in Thieken et al., 2005).

Figure 1 shows the Pearson correlation coefficient of the 28 candidate predictors and the predictand loss ratio L_R. The upper row contains the correlation between the candidate predictors and loss ratio. Although there are many variables that are significantly correlated to loss ratio, correlation coefficients are usually low. Water depth has the highest absolute correlation (0.50) to loss ratio, followed by contamination (0.37), duration (0.26), warning quality (0.24), precautionary measures indicator (-0.23), warning time (0.23), and flow velocity (0.21). In addition, the Spearman rank correlation coefficient has been calculated (not shown). In many cases both correlation coefficients are very similar, indicating that monotonic non-linearity is weak. There are a number of cases where both correlation coefficients differ clearly. Hence, there is evidence for either strong non-linearity or for serious outliers in the data. Figure 2 shows the scatter plot between water depth and loss ratio. It is obvious that water depth explains only a small part of the total variability.

3 Methods

We apply two variants of tree-based models: regression trees and bagging decision trees. They are used to determine the important damage-influencing parameters from a large database, to understand interactions between predictor variables, and to estimate the direct building damage. We screen the list of 28 candidate predictors and determine their relevance for the predictand L_R (relative loss). To test the ability of these tree-based models for predicting damage, we compare their prediction power with established damage models. Hence, we establish a model which predicts the relative loss as function of the relevant damage-influencing variables $L_R = f_T(x_1, x_2, ..., x_k)$ with the relevant predictors $x_1, x_2, ..., x_k$, and the tree-based structure $f_T(...)$. The tree-based analyses are performed with the Matlab Statistics Toolbox whose algorithms are based on Breiman et al. (1984).

3.1 Regression trees (RT)

Regression trees are tree-building algorithms for predicting continuous dependent variables. They recursively sub-divide the predictor data space into smaller regions in order to approximate a nonlinear regression structure. At each split the

Table 1. Description of the 28 candidate predictors (C: continuous, O: ordinal, N: nominal).

	#	Predictors		Type and range	Amount of data*
hydrologic, hydraulic aspects	1	wst	Water depth	C: 248 cm below ground to 670 cm above ground	2108
	2	d	Inundation duration	C: 1 to 1440 h	2094
	3	v	Flow velocity indicator	O: 0 = still to 3 = high velocity	2120
	4	con	Contamination indicator	O: 0 = no contamination to 6 = heavy contamination	2122
	5	rp	Return period	C: 1 to 848 yr	2158
early warning and emergency measures	6	wt	Early warning lead time	C: 0 to 336 h	1364
	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	955
	8	ws	Indicator of flood warning source	O: 0 = no warning to 4 = official warning through authorities	1675
	9	wi	Indicator of flood warning information	O: 0 = no helpful information to 11 = many helpful information	1631
	10	wte	Lead time period elapsed without using it for emergency measures	C: 0 to 335 h	842
	11	em	Emergency measures indicator	O: 1 = no measures undertaken to 17 = many measures undertaken	2158
precaution, experience	12	pre	Precautionary measures indicator	O: 0 = no measures undertaken to 38 = many, efficient measures undertaken	2158
	13	epre	Perception of efficiency of private precaution	O: 1 = very efficient to 6 = not efficient at all	2043
	14	fe	Flood experience indicator	O: 0 = no experience to 9 = recent flood experience	619
	15	kh	Knowledge of flood hazard	N (yes/no)	1472
building characteristics	16	bt	Building type	N (1 = multifamily house, 2 = semi-detached house, 3 = one-family house)	1816
	17	nfb	Number of flats in building	C: 1 to 45 flats	1726
	18	fsb	Floor space of building	C: 45 to 18 000 m^2	1496
	19	bq	Building quality	O: 1 = very good to 6 = very bad	1758
	20	bv	Building value	C: 92 244 to 3 718 677 €	1419
socio-economic status	21	age	Age of the interviewed person	C: 16 to 95 yr	2097
	22	hs	Household size, i.e. number of persons	C: 1 to 20 people	2125
	23	chi	Number of children (< 14 yr) in household	C: 0 to 6	1877
	24	eld	Number of elderly persons (> 65 yr) in household	C: 0 to 4	1983
	25	own	Ownership structure	N (1 = tenant; 2 = owner of flat; 3 = owner of building)	2158
	26	inc	Monthly net income in classes	O: 11 = below 500 € to 16 = 3000 € and more	1666
	27	socp	Socioeconomic status according to Plapp (2003)	O: 3 = very low socioeconomic status to 13 = very high socioeconomic status	1469
	28	socs	Socioeconomic status according to Schnell et al. (1999)	O: 9 = very low socioeconomic status to 60 = very high socioeconomic status	1308

*Since not all people were willing to answer all questions, not all information is available for each interview.

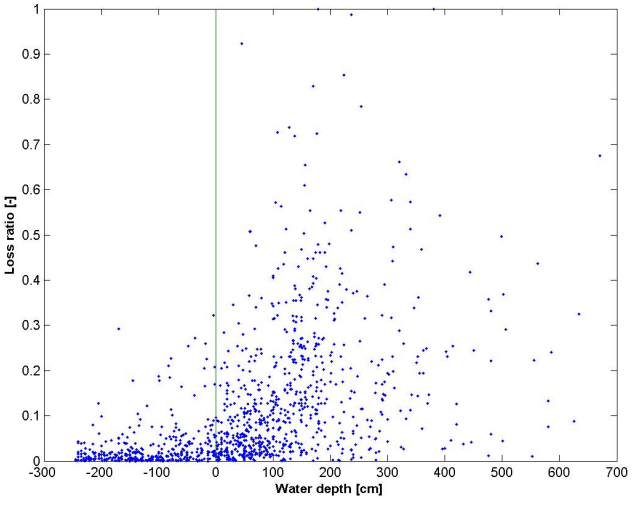

Fig. 2. Scatter plot showing the relation between loss ratio and water depth. Negative water depths indicate basement flooding.

data set is partitioned into two sub-spaces in such a way that the improvement in predictive accuracy is maximized. The algorithm searches over all possible split values of all predictor variables to identify the split which minimizes an error criterion. The variance of the response variable L_R (relative loss) is used as the criterion, however, other splitting criteria are possible (Breiman et al., 1984; Torgo, 1999). The repeated binary partitioning leads to a tree structure, from the root node to the terminal nodes (or leaves). Each terminal node of the tree represents a cell of the partition. The interior nodes (or splits) are labeled with questions, and the binary branches are labeled with the answers (Fig. 3). To obtain a prediction using the regression tree, a sequence of questions is asked which starts at the root node and ends at a terminal node. The prediction for an input x_1, x_2, ..., x_k is the average of the response variable of all the samples of the training data set that belong to this terminal node. More complicated tree-based models can be developed, e.g. fitting a local regression model to the data of each terminal node.

One of the issues that need careful attention is overfitting. The recursive splitting of the data into subsets leads eventually to large trees with many leaves whereas the sample size of the leaves is small. Usually, such trees agree well with the training data, however, their prediction ability for independent data is poor. On the one hand, trees should be complex enough to exploit information that increases predictive power and to account for important relations between predictors and response variable. On the other hand, they should be as simple as possible and should ignore random noise that does not increase predictive power. One method to avoid overfitting

is tree pruning: a large tree is cut back to obtain a simpler tree. A regression tree is pruned by firstly pruning branches which give less improvement in error cost. Pruning generates a sequence of sub-trees of different size. The optimal tree is selected from this sequence by assessing the predictive error of each tree, for instance, by selecting the simplest tree with a predictive error comparable to the most accurate one. This process is conceptually similar to the process followed in stepwise regression where forward and backward variable selection or deletion may be pursued. The predictive error of a tree (or cost of a tree) is defined as the sum over all leaves of the estimated probability of a leaf times its average squared error over the observations in that leaf:

$$C_{\mathrm{T}} = \sum_{j=1}^{n} \left[p_j \; \frac{1}{n} \sum_{k}^{n_k} \left(R_{\mathrm{L},\,k}^{\mathrm{OBS}} - R_{\mathrm{L},\,k}^{\mathrm{SIM}} \right)^2 \right] \quad (1)$$

with: $n =$ number of leaves, $p_j =$ probability of leaf j, $n_k =$ number of observations of leaf j, $R_{\mathrm{L},\,k}^{\mathrm{OBS}} =$ observed loss ratio of observation k, $R_{\mathrm{L},\,k}^{\mathrm{SIM}}$ simulated loss ratio of observation k.

The predictive error as a function of the tree size is estimated by 10-fold cross-validation. The data set is randomly split into ten sub-samples. A tree is computed ten times, each time leaving out one of the sub-samples, and using that sub-sample as a test sample for cross-validation. For each tree size the cost is calculated by averaging the results from the ten sub-samples. The cost function shows the predictive error of a tree as function of its number of terminal nodes.

3.2 Bagging decision trees (BT)

Bagging decision trees are an ensemble of many regression trees. As indicated earlier, they attempt to reduce the uncertainty associated with the selection of a single model, by pooling an ensemble of plausible or candidate models. They are derived by generating many bootstrap replicas of the data set and by growing a regression tree on each replica. The response of a bagging decision tree is the average over the responses of all individual regression trees in the ensemble. Bootstrapping makes it robust against changes in data and avoids overfitting.

A bootstrap replica is generated by randomly drawing with replacement n observations, where n is the data set size. On average, 37 % of observations are not considered for building an individual tree. These observations are called out-of-bag observations. The average out-of-bag error is a quality measure of a BT and is defined as the average over predictions from all trees in the ensemble for which this observation is out of bag. Bagging decision trees provide a metric (called feature importance) for determining the relevance of each potential predictor by randomly permuting out-of-bag data across one variable at a time and estimating the increase in the out-of-bag error: the higher the increase, the more important the feature.

3.3 Comparing tree-based models with established damage models

Once a regression or bagging decision tree is grown, it can be used to estimate the loss ratio. We compare the performance of the tree-based models with two established flood damage models. To have a fair comparison, all models are derived from the same data set. We restrict the comparison of model performance to those cases where the necessary parameters for all established models are given.

Firstly, the traditional approach, i.e. stage-damage function, is compared. We fit a root function ($L_{\mathrm{R}} = a_1 + a_2 \sqrt{\mathrm{wst}}$; coefficients a_1, a_2; water depth "wst") to the damage data by the method of least squares. We differentiate between cases where only the basement is flooded, and those where floors above the basement are affected. Each record in the data set is assigned to one of these two cases, and a root function is fit to these two sub-samples, i.e. for basement cases and for higher floor cases, respectively. This approach – the stage-damage functions – considers only water depth as predictor; other variables are not considered when estimating the flood loss ratio.

Further, the performance of the tree-based models is compared to FLEMOps+r (Elmer et al., 2010). This model has been developed using the same data set and it has been shown to provide superior results compared to other approaches currently used in Germany. FLEMOps+r calculates the building loss ratio for private households using five classes of water depth, three intervals of return period, three individual building types, two classes of building quality, three classes of contamination and three classes of private precaution. In essence, the data set is stratified into 27 sub-samples and the average loss ratio is used as damage estimator (Elmer et al., 2010). FLEMOps+r is similar to regression tree models in two respects: the complete data space is partitioned in subspaces, and the average value of each subspace is used as prediction. However, there are essential differences in the partitioning process. The subdivision of FLEMOps+r has been developed based on expert knowledge and a number of different analyses of the data set (Thieken et al., 2005; Elmer et al., 2010). The partition is regular, e.g. each water depth class is combined with each building type class, and so forth. Regression tree algorithms partition according to an optimization criteria which results in irregular partitions.

The comparison of the performance of the different damage models is based on the following resampling procedure. 100 households are randomly drawn from the data set, and each model is applied to this random sample. The agreement between the predictions and the true values is quantified by three error measures: root mean square error (RMSE), mean bias, and correlation coefficient. This step is repeated 100 times, yielding 100 estimates for the prediction errors.

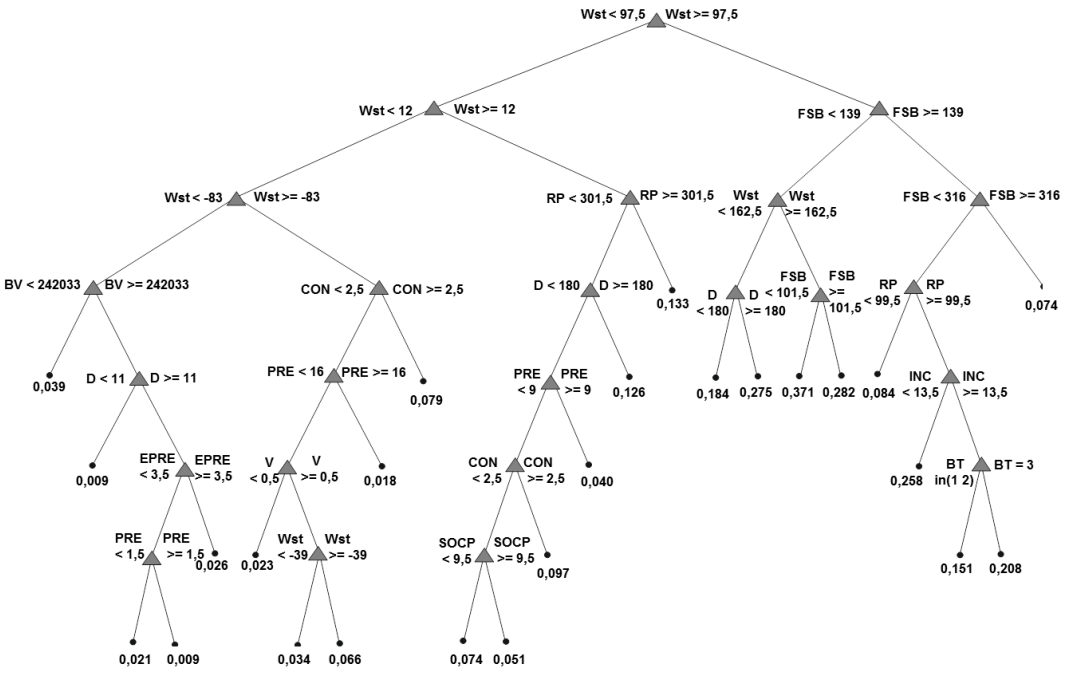

Fig. 3. Regression tree RT1 with 25 leaves for estimating building loss ratio (wst: water depth [cm]; fsb: floor space of building [m^2]; rp: return period [yr]; bv: building value [€]; con: contamination indicator [−]; pre: precautionary measures indicator [−]; inc: monthly net income [−]; epre: perception of efficiency of private precaution [−]; socp: socio-economic status [−]; bt: building type [−]). Terminal node values give the average loss ratio of all data values of the terminal node.

4 Results

4.1 Important damage-influencing parameters and their interactions

4.1.1 Regression trees

A first regression tree is grown with the stopping criteria that the minimum number of cases in terminal nodes is 30. This results in a tree with 25 terminal nodes (Fig. 3). The interpretation of a regression tree is straightforward. By looking at the tree, the important variables are disclosed. Out of the 28 candidate predictors given in Table 1, the regression tree considers twelve variables. Table 2 shows the twelve predictors and their relation to the predictand building loss ratio. In addition, Table 2 shows how many times a variable occurs as decision node. The number of occurrence and the position of a given decision node in the tree give an indication of the importance of the respective predictor. The more often a variable occurs, and the closer a decision node is to the root node, the more important is the variable and the node, respectively.

The root node and four other nodes of RT1 are water depth nodes; hence, not surprisingly, water depth is the most important predictor. This result is in accordance with previous flood damage analyses and most flood damage estimation models, i.e. stage-damage functions are based on this finding (Penning-Rowsell and Green, 2000; Merz et al., 2010). Other important variables (three decision nodes) are floor space of building, duration of inundation and precaution. In 18 out of 24 nodes, the relation between the predictor and relative damage is as expected (see Table 2: type of correlation). Table 2 shows that water depth is positively correlated with relative loss, i.e. branches of the regression tree with larger water depth have larger loss ratio. As expected, inundation duration, return period, contamination indicator and flow velocity indicator are also positively correlated with loss ratio, and to an extent with each other. Contamination and flow velocity indicator nodes are only present in the left part of the tree after the root node split, indicating that their influence on the loss ratio is particularly important in areas with smaller water depth (wst < 97.5 cm). This is confirming a previous study which showed that a significant influence of flow velocity on the structural damage of residential buildings is suspected above a critical impact level of 2 m of energy head or water depth (Kreibich et al., 2009). The precautionary measures indicator is negatively correlated with the loss ratio, i.e. the better the private precaution, the lower the loss ratio. This is confirming results of quantitative damage reduction of individual measures by Kreibich et al. (2005). They showed for instance that flood adapted use and interior fitting reduced the damage ratio for buildings by 46 % and 53 %, respectively. Since precaution nodes are only present in the left part of the tree after the root node split, precaution is important (three splits) but only for smaller water depths (wst < 97.5 cm). This is interpreted as hint that private precaution is most effective in areas

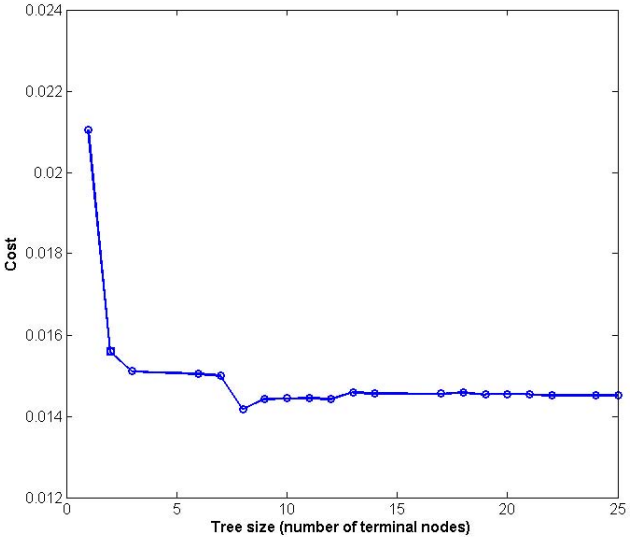

Fig. 4. Cost function of regression tree RT1.

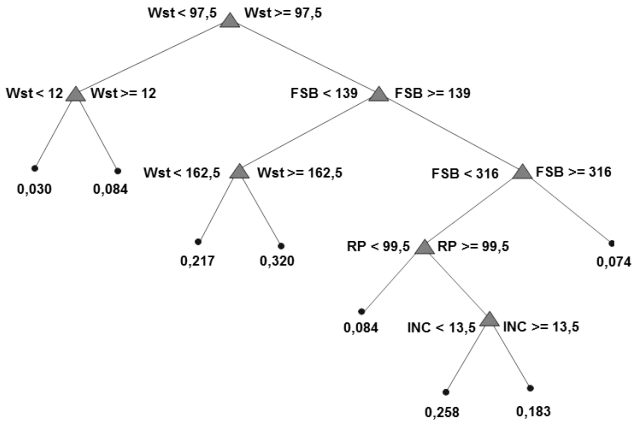

Fig. 5. Regression tree RT2 with eight leaves for estimating building loss ratio (wst: water depth [cm]; fsb: floor space of building [m^2]; rp: return period [yr]; inc: monthly net income [−]). Terminal node values give the average loss ratio of all data values of the terminal node.

with low flood water levels, which confirms expert judgment presented in ICPR (2002). As expected, perception of efficiency of private precaution is positively correlated with the loss ratio, i.e. households who perceived private precaution as efficient had lower loss ratios.

Thieken et al. (2005) showed that multi-family houses received a very high absolute building damage, but their loss ratio is smaller in comparison to single-family houses. This is confirmed here, since the branch containing multi-family and semi-detached houses has lower building loss ratios in comparison with the branch containing one-family houses. For four predictors (six decision nodes for floor space of building, building value, monthly net income, socio-economic status) their relation to loss ratio is not that obvious. All these cases show an inverse relation, i.e. higher values of the predictor correspond with lower loss ratios. Floor space of building and building value is strongly correlated (Fig. 1). The fact that the larger the building (and the higher its value) the lower its loss ratio is in accordance with the finding that single-family houses have a higher loss ratio in comparison with multi-family houses. It is also in accordance with findings of Thieken et al. (2005) who showed that the building loss ratio decreases if the total floor space of the building exceeds 120 m^2. The first split based on building floor space in RT1 is at 139 m^2, which is close to the number given by Thieken et al. (2005). Since floor space nodes are only present in the right part of the tree after the root node split, floor space is only important for larger water depths (wst > 97.5 cm), i.e. in cases where not only the cellar is affected.

Figure 4 shows the cost function of regression tree RT1. The cost is approximately constant for the sequence of subtrees from eight to 25 terminal nodes. The lowest cost is calculated for the tree RT2 which has been obtained by pruning RT1. RT2 consists of eight leaves and is much simpler than the full-size tree RT1. Figure 5 shows the simplified tree RT2. It considers four variables: water depth (wst), floor space of building (fsb), return period (rp) and monthly net income of the household (inc). These variables correlate to building loss ratio as in RT1.

4.1.2 Bagging decision trees

Similarly to regression trees, a first bagging decision tree is grown with the stopping criteria that the minimum number of cases in terminal nodes is 30. The number of trees in the ensemble is set such that the model error becomes stable. Figure 6 shows the out-of-bag feature importance. The ranking of the candidate predictors is water depth (wst), floor space of building (fsb), return period (rp), building value (bv), contamination (con), inundation duration (d), precautionary measures indicator (pre), and flow velocity (v). Other variables show very small feature importance. This list of more important variables for predicting the loss ratio is similar to the importance that has been obtained by growing regression trees, in particular if it is compared to RT1. All of the variables listed above also appear in RT1 with at least two splits. There is one exception, namely, flow velocity (v) appears only once in RT1. In a few cases there are (very small) negative values of feature importance. This means that permuting this variable has led to a slightly better prediction. This is a random effect and is not significant.

In a further step, a bagging decision tree BT2 is grown by limiting the candidate predictors to the more important variables, i.e. the eight variables listed above.

4.2 Performance of flood damage models

Figure 7 compares the performance of the four tree-based models with existing flood damage models, namely

Table 2. Damage-influencing variables of regression tree RT1 with 25 leaves.

Predictors		No. of decision nodes	Type of correlation*
wst	Water depth	5	+
fsb	Floor space of building	3	–
d	Inundation duration	3	+
pre	Precautionary measures indicator	3	–
rp	Return period	2	+
con	Contamination indicator	2	+
bv	Building value	1	–
inc	Monthly net income in classes	1	–
v	Flow velocity indicator	1	+
epre	Perception of efficiency of private precaution	1	+
bt	Building type	1	n.a.
socp	Socio-economic status (Plapp, 2003)	1	–

*All predictors show the same correlation at all their nodes.

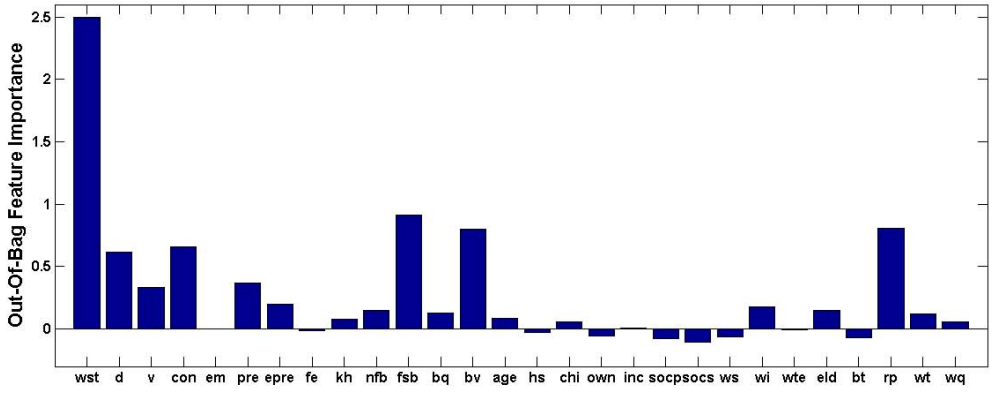

Fig. 6. Out-of-bag feature importance for bagging decision tree BT1.

stage-damage function approach and FLEMOps+r. 100 sets of 100 affected residential buildings are randomly drawn from the data set, each model is applied to each building record and the three error measures (root mean square error RMSE, mean bias MBE, correlation coefficient) are calculated and shown in boxplots.

There is clear improvement of the tree-based models compared to the established approaches. The model with the worst performance (exception: mean bias) is the stage-damage model; this is explained by the fact that it is a 1-dimensional model, considering only water depth as predictor.

The multi-variate model FLEMOps+r outperforms the stage-damage approach, however, it compares unfavorably with the tree-based models. It is interesting to compare the performance of the regression trees and FLEMOps+r. Both approaches are multi-variate and estimate flood damage by dividing the data set into sub-samples, and by using the average loss ratio of these sub-samples as estimator. RT1 and FLEMOps+r are of similar complexity; they use a similar number of sub-samples (FLEMOps+r: 27; RT1: 25), but a different number of predictors (FLEMOps+r: 6; RT1: 12). The prediction error of RT1 is significantly smaller. It is interesting to note that even the much simpler regression tree RT2 with only eight leaves and only 4 predictors shows clearly better results than FLEMOps+r. Thus, further development of flood damage models should take advantage of tree-based approaches, since a better representation of the damaging processes in comparison with stage-damage functions is possible due to the consideration of more predictors.

The performance of the four tree-based models is comparable. There is small improvement of BTs over RTs, and more elaborate trees (RT1, BT1) perform slightly better than the reduced trees (RT2, BT2). The main metric that improves in the BT over RT is the correlation coefficient. This makes sense since it reflects the potentially reduced variance across the bagging ensembles. Bagging averages across multiple candidate models that are not completely independent, since they are drawn from the same original data set under resampling. If they were independent, the variance of the average of m estimates would be $1\,m^{-1}$ of the variance of any one of the estimates. In bagging, a reduction in the variance of the estimate will still occur during the averaging process. The RMSE and Bias are unchanged. The differences across all

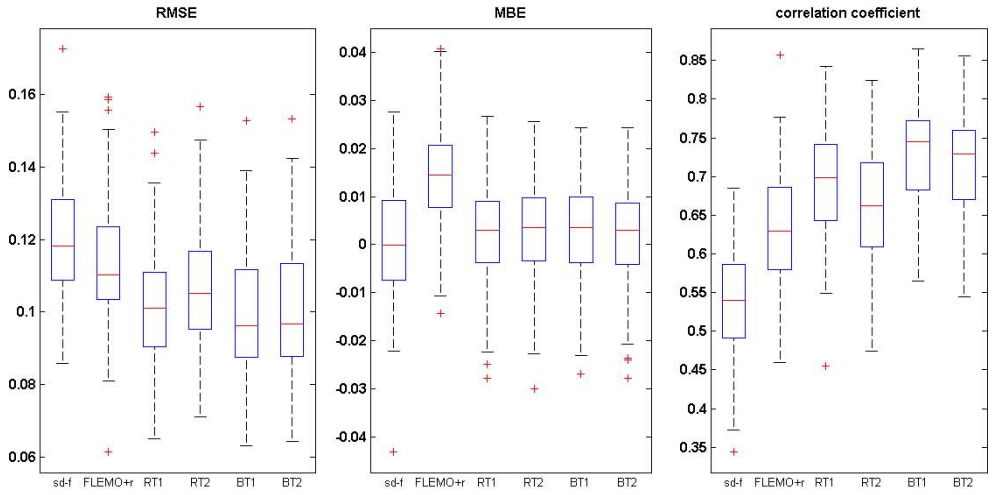

Fig. 7. Comparison of flood damage estimation models (sd-f: stage-damage function; FLEMOps+r; RT1, RT2: regression trees with 25 and 8 leaves, respectively; BT1, BT2: bagging decision trees with 28 and 8 predictors, respectively).

models in terms of the spread of the criteria is however not statistically significant at the 90 or 95 % level.

5 Conclusions

This paper reports about the use of tree-based approaches for the analysis of flood damage data and for the estimation of flood damage. Tree-based approaches are used to identify important damage-influencing variables and their relation to direct building damage. Summing up the results of the regression trees and bagging decision trees, the following damage-influencing variables have been identified as important: water depth, floor space of building (and the strongly correlated building value), return period, contamination, inundation duration and precautionary measures indicator. The high importance of water depth is in accordance with many previous studies and the traditional approach of using stage-damage functions for flood damage estimation. The importance of return period, contamination and precaution confirms previous findings (Thieken et al., 2005; Kreibich et al., 2005; Elmer et al., 2010) and these variables are used in the flood damage model FLEMOps+r (Elmer et al., 2010). The revealed importance of the floor space of building and the building value, two variables which are highly correlated, as well as of inundation duration, is interesting. There have been clues that they might be important (Thieken et al., 2005), but to our best knowledge, they have so far not been used for building damage modelling. In FLEMOps+r two variables describing the building are used, namely building type and building quality (Elmer et al., 2010) of which building type is highly correlated with the floor space and the building value (Fig. 1). Inundation duration is an important variable for estimating flood damage to agricultural crops (Förster et al., 2008; Tapia-Silva et al., 2011), but its importance for building damage may have been underestimated so far. Compared to former studies using this data set, such as Thieken et al. (2005), this study shows that tree-based models are very effective in identifying the important damage influencing variables and their interactions.

Tree-based models are a simple means to multi-variate damage modelling. Although damage processes are inherently multi-dimensional, damage models are often univariate, limited to water depth as predictor. It is shown that tree-based models perform better than existing models like stage-damage functions and the multi-variate FLEMOps model. Tree-based damage models are easy to understand and use. They permit to include both continuous, e.g. water depth, and categorical predictors, e.g. building type. Regression approaches have difficulties in handling categorical variables; tree-based models may be advantageous since they effectively decide whether or not to put these categories into fewer classes. Tree-based models allow for nonlinearities and predictor interactions and they do not use implicit assumptions about relationships between predictand and predictors (such as linear relations or normal distributions). An important advantage is their ability to exploit the local relevance of predictors. They avoid the need to find a parametric function which holds globally across all the data. For example, precaution appears only in the part of the regression tree RT1 with smaller water depths. This result confirms the hypothesis that private precaution is particularly effective when flood water levels are small; in areas with high flooding private precaution loses its ability to reduce damage.

A disadvantage is that tree-based approaches only reflect the nature of the relationships that are contained within the available data and that large data sets are needed in order to identify complex relationships, especially in high-dimensional data spaces. This might hamper the application of this approach for flood damage analyses and modelling

in other regions where comprehensive, multi-dimensional databases do not exist. Although tree-based models allow multi-variate modelling, it has to be tested under which conditions such models are justified. Traditional flood damage models using only water depth as predictor have very limited data demand, and important damage-influencing variables, e.g. contamination, are hardly quantifiable. Regression trees as well as bagging decision trees can handle incomplete data: if data is missing, predictions are based by considering only the leaves that can be reached given the available data. However, depending on the data availability and the context of the damage assessment, the gain in performance of multi-variate models might be lost in real-world applications. For example, for the estimation of the cumulative loss in a large area with a large number of residential buildings simpler models might be the better choice, since we expect that differences between single buildings will play a smaller role for increasing numbers of households. The predictor selection problem is always a challenging one, and the "best" predictors in one setting may not be the best predictors in another setting. Nevertheless, as attention to flood damage prediction increases considering regional pooling or variation in predictors may be useful to provide for more robust predictor selection, especially where the predictors are generally correlated, and/or the contribution of a particular predictor may be meaningful only in a certain range of values of that predictor given the other predictors available, as was demonstrated with the data set considered in this paper.

The evaluation of model performance in this paper is based on random samples which are not independent from the data used for model development. Hence, our comparison of model performance does not give information about the transferability of the models. Future work will use independent flood damage data and will use a different model building design, in order to test specifically, to which extent models of different types can be transferred in space (the same flood event but different regions), in time (the same region but different flood events) or in space and time (different regions and different flood events). Further research focusing on the improvement of flood damage modelling and development of the damage model FLEMOps will analyse which variables and model structures are most suitable for estimating flood damage to residential buildings in respect to the transferability and applicability of different approaches.

Acknowledgements. This work has been funded by DFG (German Research Foundation, grant no. ME 1844/2-1) supporting the research stay of the first author at Columbia University.

The service charges for this open access publication have been covered by a Research Centre of the Helmholtz Association.

Edited by: L. Bouwer
Reviewed by: M. Kok and two anonymous referees

References

Ali, G. A., Roy, A. G., Turmel, M.-C., and Courchesne, F.: Multivariate analysis as a tool to infer hydrologic response types and controlling variables in a humid temperate cachment, Hydrol. Process., 24, 2912–2923, 2010.

Breiman, L.: Bagging predictors, Mach. Learn., 24, 123–140, 1996.

Breiman, L., Friedman, J., Olshen, R. A., and Stone, C. J.: CART: Classification and Regression Trees, Wadsworth, Belmont, CA, 1984.

Bubeck, P. and Kreibich, H.: Natural Hazards: direct costs and losses due to the disruption of production processes. WP1 Final Report of the EU Project Costs of Natural Hazards (ConHaz), http://conhaz.org/CONHAZ\protect\kern+.1667em\relax%20REPORT\protect\kern+.1667em\relax%20WP01_2.pdf, 2011.

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

Carlisle, D. M., Falcone, J., Wolock, D. M., Meador, M. R., and Norris, R. H.: Predicting the natural flow regime: models for assessing hydrological alteration in streams, River Res. Appl., 26, 118–136, 2010.

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, doi:10.5194/nhess-10-2145-2010, 2010.

Förster, S., Kuhlmann, B., Lindenschmidt, K.-E., and Bronstert, A.: Assessing flood risk for a rural detention area, Nat. Hazards Earth Syst. Sci., 8, 311–322, doi:10.5194/nhess-8-311-2008, 2008.

Gall, M., Borden, K. A., and Cutter, S. L.: When do losses count? Six fallacies of natural hazards loss data, B. Am. Meteorol. Soc., 90, 799–809, 2009.

Grunwald, S., Daroub, S. H., Lang, T. A., and Diaz, O. A.: Tree-based modeling of complex interactions of phosphorus loadings and environmental factors, Sci. Total Environ., 407, 3772–3783, 2009.

Handmer, J., Abrahams, J., Betts, R., and Dawson, M.: Towards a consistent approach to disaster loss assessment across Australia, Australian J. Emergency Manage., 20, 10–18, 2005.

ICPR (International Commission for the Protection of the Rhine): Non structural flood plain management, Measures and their effectiveness, International Commission for the Protection of the Rhine, Koblenz, 2002.

Iorgulescu, I. and Beven, K. J.: Nonparametric direct mapping of rainfall–runoff relationships: an alternative approach to data analysis and modeling?, Water Resour. Res., 40, W08403, doi:10.1029/2004WR003094, 2004.

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

Kreibich, H., Thieken, A. H., Petrow, Th., Müller, M., and Merz, B.: 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, 117–126, doi:10.5194/nhess-5-117-2005, 2005.

Kreibich, H., Müller, M., Thieken, A. H., and Merz, B.: Flood precaution of companies and their ability to cope with the flood

in August 2002 in Saxony, Germany, Water Resour. Res., 43, W03408, doi:10.1029/2005WR004691, 2007.

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

Kreibich, H., Seifert, I., Merz, B., and Thieken, A. H.: Development of FLEMOcs – A new model for the estimation of flood losses in companies, Hydrolog. Sci. J., 55, 1302–1314, 2010.

Kreibich, H., Seifert, I., Thieken, A. H., Lindquist, E., Wagner, K., and Merz, B.: Recent changes in flood preparedness of private households and businesses in Germany, Reg. Environ. Change, 11, 59–71, 2011.

Loos, M. and Elsenbeer, H.: Topographic controls on overland flow generation in a forest – An ensemble tree approach, J. Hydrol., 409, 94–103, 2011.

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

Mototch, N. P., Colee, M. T., Bales, R. C., and Dozier, J.: Estimating the spatial distribution of snow water equivalent in an alpine basin using binary regression tree models: the impact of digital elevation data and independent variable selection, Hydrol. Process., 19, 1459–79, 2005.

Nicholas, J., Holt, G. D., and Proverbs, D.: Towards standardizing the assessment of flood damaged properties in the UK, Struct. Survey, 19, 163–172, 2001.

Pappenberger, F., Iorgulescu, I., and Beven, K. J.: Sensitivity analysis based on regional splits and regression trees (SARS-RT), Environ. Modell. Softw., 21, 976–90, 2006.

Parker, D., Tapsell, S., and McCarthy, S.: Enhancing the human benefits of flood warnings, Nat. Hazards, 43, 397–414, 2007.

Penning-Rowsell, E. C. and Green, C.: New Insights into the appraisal of flood-alleviation benefits: (1) Flood damage and flood loss information, J. Chart. Inst. Water E., 14, 347–353, 2000.

Plapp, S. T.: Risk perception of natural catastrophes – an empirical investigation in six endangers areas in south and west Germany), Karlsruher Reihe II – Risikoforschung und Versicherungsmanagement Band 2, edited by: Werner, U., Verlag Versicherungswirtschaft, Karlsruhe, 2003 (in German).

Schnell, R., Hill, P. B., and Esser, E.: Methods of the empirical social sciences, 6th Edn., Oldenbourg, München, Wien, 535 pp., 1999 (in German).

Solomatine, D. P. and Dulal, K. N.: Model trees as an alternative to neural networks in rainfall–runoff modeling, Hydrolog. Sci. J., 48, 399–412, 2003.

Tapia-Silva, F.-O., Itzerott, S., Förster, S., Kuhlmann, B., and Kreibich, H.: Estimation of flood losses to agricultural crops using remote sensing, Phys. Chem. Earth Pt. A/B/C, 36, 253–265, 2011.

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, W12430, doi:10.1029/2005WR004177, 2005.

Thieken, A. H., Kreibich, H., Müller, M., and Merz, B.: Coping with floods: preparedness, response and recovery of flood affected residents in Germany in 2002, Hydrol. Sci. J., 52, 1016–1037, 2007.

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, edited by: Proverbs, D., Brebbia, C. A., and Penning-Rowsell, E., WIT Press, 315–324, 2008.

Torgo, L.: Inductive Learning of Tree-based Regression Models, Dissertation, University of Porto, Porto, 1999.

Wei, W., Watkins Jr. and D. W.: Data mining methods for hydroclimatic forecasting, Adv. Water Resour., 34, 1390–1400, 2011.

Wind, H. G., Nierop, T. M., de Blois, C. J., and de Kok, J. L.: Analysis of flood damages from the 1993 and 1995 Meuse floods, Water Resour. Res., 35, 3459–3465, 1999.

Zhai, G., Fukuzono, T., and Ikeda, S.: Modeling flood damage: case of Tokai Flood 2000, J. Am. Water Resour. As., 41, 77–92, 2005.