Multi-variable flood damage modelling with limited data using supervised learning approaches

Dennis Wagenaar[1], Jurjen de Jong[1], Laurens M. Bouwer[1]

[1] Deltares, Delft, The Netherlands
Correspondence to: Dennis Wagenaar (dennis.wagenaar@deltares.nl)

Abstract. Flood damage assessment is usually done with damage curves only dependent on the water depth. Several recent studies have shown that supervised learning techniques applied to a multi-variable dataset can produce significantly better flood damage estimates. However, creating and applying a multi-variable flood damage model requires an extensive dataset, which is rarely available and this is currently holding back the widespread application of these techniques. In this paper we enrich a dataset of residential building and content damages from the Meuse flood of 1993 in the Netherlands, to make it suitable for multi-variable flood damage assessment. Results from 2D flood simulations are used to add information on flow velocity, flood duration and the return period to the dataset, and cadastre data is used to add information on building characteristics. Next, several statistical approaches are used to create multi-variable flood damage models, including regression trees, bagging regression trees, random forest, and a Bayesian network. Validation on data points from a test set shows that the enriched dataset in combination with the supervised learning techniques delivers a 20% reduction in the mean absolute error, compared to a simple model only based on the water depth, despite several limitations of the enriched dataset. We find that with our dataset, the tree-based methods perform better than the Bayesian Network.

1 Introduction

Decision making in flood risk management is increasingly based on studies that quantify the flood risk rather than only the flood hazard. Flood damage estimation is therefore becoming increasingly important (Merz et al, 2010).Flood risk assessment supports policy makers to decide which flood risk management measures are most efficient in reducing flood risks and how much investment is cost-efficient. With the European Union Floods Directive (EC, 2007) now fully in place, national flood risk assessments are being developed with the final aim to support flood risk management plans. In the Netherlands, such flood damage assessment has been used to derive the optimal protection standard for flood protection (Kind, 2013; van der Most, 2014), using the current Dutch standard method for damage modelling (Kok et al., 2005). Also for insurance applications, more precise estimates of flood damages are required.

Flood risk assessments require flood damage models. These models typically predict the damage as fraction of the potential damage, based on the water depth, and average building repair and replacement costs for different types of buildings (Messner et al., 2007; Jonkman et al., 2008). Similar approaches are also applied to other natural hazards, for example for landslides (Papathome-Kohle et al., 2014) and the software package HAZUS can be used for floods, earthquakes and

hurricanes (Scawthorn et al., 2006). Alternative approaches to calculate flood risk do also exist, such as vulnerability indicators (Papathoma-Köhle, 2016).

When validated, simple flood damage models often don't perform well (e.g. Jongman et al., 2012). This is because water depth alone cannot explain the full complexity of the flood damaging processes and several studies have only found low correlation coefficients (typically below 0.5) between the water depth and the flood damage (e.g. Merz et al., 2013, Pistrika and Jonkman, 2009). Furthermore, often no local data is available on flood damage and therefore a relationship between the water depth and damage either needs to be estimated or transferred from other areas (Wagenaar et al., 2016). This can cause errors as simple models hold many implicit assumptions that may not be valid for the situation the model is transferred to. For instance, Elmer et al. (2010) showed that an event with a low flood probability could not use the same damage function as a flood event with a high probability. These implicit assumptions cause large unexplained differences between flood damage functions (Wagenaar et al.,2016; Gerl et al., 2016). Transferability however could be improved, when a model describes more variations of the damaging process, and when more variables are included in the damage models (e.g. flood probability is explicitly part of the model). Similar problems are also present in the modelling of other natural hazards. For example Fuchs et al. (2007) found that building materials are very important for debris flow damage modelling and that models can therefore not always be transferred in space and time.

Current approaches suffer from two main limitations: first, they rely on limited information and usually only take into account water depth as a predictor, and use a deterministic relation between water depth and some fraction of average maximum damages; secondly, they are deterministic in nature, while it has been shown that uncertainties in this approach are large, but generally not quantified e.g. in the Dutch standard method (Egorova et al., 2008). Some of the multi-variable methods are able to provide probability distributions, rather than deterministic estimates of damages.

Recently, multi-variable flood damage models have been created with a German dataset based on telephone interviews. Thieken et al.(2005) found that apart from the water depth also the contamination of the flood water and precautionary measures were important to estimate the flood damage. In Thieken et al. (2008) these extra variables were included in a simple multi-variable flood damage model as a surcharge. Using information from this same database, Merz et al. (2013) used regression and bagging trees and Vogel et al. (2014) used Bayesian Networks to predict the flood damage. Spekkers et al. (2014) applied regression trees to estimate pluvial flood damage. Van Oostegem et al.(2015) applied the Tobit estimation technique to a multi-dimensional dataset in Belgium to estimate pluvial flood damages. These multi-variable flood damage models have been shown to perform better than simple flood damage models in Schröter et al. (2014) (up to 25% reduction in mean absolute error, MAE), both tested on their own dataset and on datasets from other floods (Schröter et al., 2014). Also, some multi-variable approaches (Bayesian Networks, Bagging trees and Random Forests) generate probability distributions of estimated damages, and thus provide information on uncertainties of the estimates. Therefore, multi-variable flood damage models look like a promising approach to improve flood damage modelling.

The application of multi-variable flood damage models for flood risk management studies is still difficult because of the large data requirements. Running a multi-variable flood damage model for a new area requires for every object several

variables on the flood hazard and building characteristics that are not yet typically collected. Also creating new multi-variable flood damage models is currently rarely done because they also require records of flood damages at building level.

More commonly available (although still rare) are simple datasets that hold records with the flood damage that occurred for each building with sometimes a few other variables (such as location or water depth). Such datasets may have been created for compensation purposes or to build simple flood damage models but may miss most of the desired variables. An example of such a dataset is the flood damage dataset collected after the Meuse flood of 1993 in the Netherlands which is used here. Previously this dataset has been described in Wind et al. (1999) and in more detail in WL Delft (1994).In this paper we will explore the use of supervised learning techniques to build flood damage models based on a dataset that is very different from the datasets usedby previous studies (i.e. the German dataset applied by Merz et al. (2013) and Schröter et al.(2014)).) The dataset in this paper was collected by insurance experts directly after the flood for compensation purposes and covers all affected buildings. This is different from the German data which was collected a year after the flood for research purposes based on a sample of the affected buildings. The data is also different in that in the original study only a few variables were collected, in contrast for the German dataset all variables (except return period) were based on telephone interview answers.

In this study several methods are applied to enrich the Meuse 1993 flood damage dataset with extra flood hazard and building characteristic variables. We will answer the question of whether this enriched dataset from a different source then previous studies is also suitable to build a multi-variable flood damage model. The expectation is that the multi-variable models perform better than a model based on a single variable (water depth) and that even data with limited quality will improve the results.

2D simulations of the 1993 flood on the Meuse are used to enrich the dataset with additional flood characteristics. Cadastre data is used to enrich the Meuse dataset with extra building characteristics. Four different supervised learning techniques are then applied to this enriched dataset: a regression tree, bagging regression trees, random forest and a Bayesian network. A part of the dataset will be held back and will only be used for validation. This validation is then used to determine whether the enriched dataset combined with supervised learning techniques performs better than a traditional damage function based on the original dataset of water depths. In this paper we will focus on predicting absolute flood damages rather than relative flood damages. This is because the exact building values are not available.

2 Method

2.1 Datasets

2.1.1 Meuse 1993 damage dataset

The dataset available for this research is based on the Meuse flood of 22 December 1993 in the Province of Limburg in the Netherlands (WL Delft, 1994). Although no dike breaches occurred in this event, several towns and urban areas located

close to the river were affected. The flood caused a total of 254 million guilder (price level 1993) in direct damages, which is approximately 180 million euros today (price level 2016). The flood inundated 180 km^2, which is about 8% of the Province of Limburg. 32% of the damage pertains to residential buildings and content (furnishings). In this study only residential damage is considered. Other major damage categories were business (29%), government (24%) and agriculture (8%) (WL Delft, 1994). These categories are not considered because they are more heterogeneous and less data about them is available.

Damage information was collected in the context of a compensation arrangement for flood damages by the national government. All data was collected by sending damage experts from insurance companies to the affected buildings, several weeks after the flood event had occurred. Directly after the damage data was collected in 1994, the data was shared with WL Delft (now Deltares) to create a flood damage model. WL Delft received 5780 records for damage to residential buildings. The damage to privately owned residential buildings was collected by an organisation called "Stichting Watersnood 1993" , the damage to companies and the structure of rental residential buildings was collected by another organisation called "Stichting Watersnood Bedrijven 1993". So, in this set up of the damage collection, the building structure of rental residential buildings was collected by "Stichting Watersnood bedrijven", the organization that collected company damages. This is different from the organization that collected the rest of the residential damages. The structure damage to rental residential buildings was only shared with WL Delft (1994) in some partial aggregate form. WL Delft (1994) presumably distributed this partially aggregated rental residential building damage over the individual rental residential buildings. The exact method for this was however not reported and the original dataset is no longer available. Therefore, we had to work with a dataset which includes unknown manual actions. The structure damage data is therefore from inconsistent quality, the content damage however has no such problems. Furthermore, it is expected that the percentage of rental residential buildings in the affected area of Limburg is relatively low, limiting the impact of this data problem.

Another issue with the dataset is that for privacy reasons the exact locations of the buildings were not shared with WL Delft. Only the 6 digit postal code was available for this study, which makes it difficult to enrich the dataset, as between 1 and 20 buildings share the same 6 digit postal codes in the dataset.

In the original dataset the water depth (relative to the ground floor level) was estimated by the experts that surveyed the damage. The quality of the water depth estimate is questioned by WL Delft (1994; report 9, appendix A) because it was not the main aim of the survey and the experts visited several weeks after the water had receded. A plot of the water depth (see Fig. 1) and the damage doesn't show an obvious relation. The correlation between the water depth and the damage is weak (Pearson correlation coefficient = 0.18).

The final dataset also contains information on the number of inhabitants per building, whether the house has a basement and whether the house was attached to other houses. However, this data is not described in any of the available reports so the collection methods are not known, but the recorded values are clear enough to incorporate in this study. Two more variables are also included in the WL Delft dataset and also not described in any available report. These are emergency actions and ownership of the house. The meaning of the values found in the dataset for these variables is however not sufficiently clear, and could unfortunately not be taken into account in this study.

2.1.2 Upgraded Meuse 1993 dataset

To improve the dataset, additional information is required on both the flood hazard and exposure variables. The results of a 2D flood simulation and cadastre data were used to upgrade the dataset, in terms of hazard and exposure information, respectively. Because no observational data is available on flood characteristics other than the water depth, a simulation of the flood event was done. In the 2D flood simulation tool WAQUA (Rijkswaterstaat, 2013), a verified model of the state of the Meuse during the 1993 flood was available (Becker, 2012) and applied in this study to get extra variables. Using this model, a new simulation was run using a discharge boundary condition at Eijsden and a water level boundary condition at Keizersveer for the period 1 November 1993 to 31 Januari 1994. This simulation was used to create a maximum water depth map, a flood duration map, a flood return period, and a flow velocity map at a spatial resolution varying between 10 and 40 meters.

The maximum water depth and flow velocity are standard outputs of WAQUA. Flood duration is however not a standard output and is more difficult to get from a 2D flood simulation because the drainage also needs to be included in the schematisation (Wagenaar, 2012). During the 1993 Meuse flood, most drainage occurred because of the natural slope in terrain and therefore the 2D flood simulation implicitly includes most of the drainage because the discretised bed level is included. The flood duration can then be calculated by analysing the time-varying maps of the water depth and calculating for every cell the time between the moment a cell is inundated and the moment the cell is dry again. However, some cells in the digital elevation map in WAQUA are surrounded by cells that have a higher elevation. These cells do not drain in the 2D flood simulation and are still inundated at the end of the simulation. For these cells the flood duration has been calculated based on the change in water depth. If the water depth in a cell stays the same in the simulation for 24 subsequent hours the cell is considered dry at the moment this stable water depth is first reached.

Simulations were also ran with the same Meuse 1993 schematisation for design discharges with 1, 10, 50, 100, 250 and 1250 return periods. These discharges are based on HR2006 (Diermanse, 2004) and have discharges of respectively 1300, 2260, 2869, 3109, 3431 and 4000 m^3/s. The results of these simulations were combined to create a flood return period map for the Meuse 1993 situation. This map shows for each cell at what return period it first floods. Figure 2 shows that large water depths occurred and that most of the area floods frequently. The majority of the houses is however located in the safest areas with the lowest water depths and highest return periods.

These maps (water depth, flow velocity, flood duration and return periods) were linked to the original damage records using cadastre data. The data of the cadastre has exact building locations, postal codes, living area within the residential buildings, the building footprint area and the construction year. The building year was used to filter the data to find the building stock of 1993. Then, based on the building locations the 2D flood simulation results were linked to the cadastre data.

This combination of cadastre data and 2D flood simulation data is then used to make the link with the original flood damage records. First per postal code a list is made of the damage records in the postal code area and ranked based on the water depth in the original damage records. Then another list is made of the objects per postal code according to the cadastre and

also ranked based on the simulated water depth. The cadastre objects combined with the 2D flood simulation data is then linked per postal code based on the water depth rank. This results in a join between the original damage records, cadastre data and 2D flood simulation results. Table 1 gives an overview of the available records in this combined dataset.

The method of joining cadastre objects with damage records within a postal code area based on water depth rank is error prone. The modelled water depth is on average 30 cm larger than then the recorded water depth. This is possibly because the difference in reference level of both data sources as the recorded water depth is relative to the floor level and the modelled water depth is relative to the digital elevation map. Not all houses have the same floor elevation and both the recorded and the modelled water depth are uncertain, because of recording and model imprecisions. It is therefore likely that some damage records have been linked to the wrong object. However, errors will likely be limited, because the join on postal codes is accurate. Object and flood variables are generally similar for buildings within the same postal code area (e.g. houses within a street are typically similar to each other) so these errors are expected not to significantly disturb the general trends in the data. The errors are therefore considered acceptable given that the purpose of the dataset is only to build a flood damage model. If significant errors are present this would result in a reduced performance of the supervised learning algorithms on the test set. A relatively simple alternative to this water depth rank method is also applied. In this alternative, the average value at all building locations in the postal code area was assigned to each of the objects in the postal code.

2.2 Supervised learning algorithms

Several supervised learning techniques have been applied to the enriched dataset to build multi-variable flood damage models. The different supervised learning techniques all have different ways to generalize the training data in such a way that it can give useful predictions of the total damage.

These multi-variable flood damage models are be compared to two different reference models to assess the value of the enriched dataset and to assess the value of multi-variable flood damage models in general. Below the different supervised learning algorithms applied are described in further detail.

2.2.1 Regression: Root function

The first reference model only uses the square root of the water depth (see formula 1) to predict the flood damage. This model represents the damage functions commonly applied today in flood risk management studies because many damage functions have approximately the shape of a root function (e.g. Scawthorn, C., et al., 2006; Thieken et al., 2008; Penning-Rowsell et al., 2005; Sluijs et al., 2000). Merz et al. (2012) applied the same method to get a reference damage function. The purpose of this reference model is to see the benefits of using more data.

The root function (1) is fitted to the dataset in such a way that the coefficients c_1 and c_2 are optimised to get the smallest possible error based on the total damage (td) and water depth (wdf) data. The values of the coefficients are optimized for the best fit with the ordinary least squares method. This is done with the Python package SciPy.

$$td = c_1 + c_2\sqrt{wdf} \quad (1)$$

2.2.2 Multi-variable linear regression

The second reference model uses multi-variable linear regression to fit a linear model to the data. This model represents more simple/traditional techniques to make a multi-variable model from data. The purpose of this reference model is to see the benefits of better techniques to build multi-variable models from data. Multi-variable linear regression is for example used in Islam (1997) to make multi-variable flood damage models.

To ensure that the model captures general trends and doesn't fit too strongly to the observed data (overfitting) the LASSO technique is used. This technique determines the coefficients in such a way that a penalty is applied for increasing the coefficients and using the variables more.

The multi-variable linear regression was carried out with the Scikit learn library in Python (Pedregosa et al. 2011). This library requires an alpha parameter to be set which determines the height of the penalty applied by the LASSO technique. Several alpha values were tried (0, 0.5, 1 and 10) and the model could predict the total damage in the test set best (on all indicators) with an alpha of zero. When alpha is zero the method is equal to the ordinary least square method and no overfitting prevention is in place. This shows that overfitting is no issue for this dataset with simple techniques such as linear regression.

2.2.3 Regression tree learning

Decision trees are a way to represent complex relationships between data and classes in a tree structure. A decision tree can be seen as a series of binary questions (nodes) leading to an answer in the form of a class (leaf). A question can be related to any variable at any value (e.g. is the water depth smaller than 0.5m).

A regression tree is similar to decision trees but instead of classes it results in real numbers. In theory, regression trees can be very large and have a separate leaf for each unique value in the dataset. However, it is more common to combine several similar unique values inside the same leaf and represent it with a summary statistic number (mean). In such a case the regression tree is an approximation of the relationship.

Regression tree learning algorithms can create optimal regression trees based on a dataset. In this paper the dataset consists of 4398 flood damage records (incomplete records are discarded) with 11 variables for each damage record (see table 1). The regression tree algorithm aims to split the dataset into subsets in such a way that the mean squared error (MSE) of the predicted total damage for all observations is minimized compared to the observed data. It does this by calculating the reduction for all candidate splitting variables according to their value and then picking the combination that maximises the MSE reduction (ΔI), this is shown in (2). n is the total number of observations in the node, y_n is the vector of observed target

values in the node and $\bar{y}$ is the mean of the target values in the node. y_{nL} and y_{nR} are vectors with the observed target values of the left and right group after the split and $\bar{y}_L$ and $\bar{y}_R$ are the mean observed target value for the left and the right group. The regression tree is grown by repeating this process at each node of the tree. This has been done with the Scikit learn library in Python (Pedregosa et al. 2011).

$$\Delta I = \frac{1}{n}\left(\sum (y_n - \bar{y})^2 - \sum (y_{nL} - \bar{y}_L)^2 - \sum (y_{nR} - \bar{y}_R)^2\right) \quad (2)$$

A regression tree algorithm keeps splitting the dataset into new branches until no more reductions in the MSE can be made. This can result in overfitting, which results in very large trees with only one data point per leaf. These very large trees are not a realistic representation of reality, and they typically perform badly when they have to predict the damage for a new data point that wasn't used for building the tree. There are several methods to prevent overfitting. The simplest methods require a minimum number of data points in a leaf or set a maximum number of nodes that the tree is allowed to contain. The disadvantage of these methods is that they sometimes don't build out a branch within the tree which at first doesn't look promising but which can make valuable homogeneity improvements deeper in the tree. A method called pruning is a more sophisticated method, in which the entire tree is first build with a subset of the data points, and then cut back based on its performance on data points that were not used for building the tree. This method was investigated in this research. This was done using *Matlab's 'Statistics and Machine Learning Toolbox'* (Matlab website), based on the work by Breiman et al. (1984), because the Python libraries do not support pruning. The performance of the pruning algorithm on this dataset was similar to a regression tree built with a combination of a minimum data point requirement per leaf and a maximum number of leaves (MAE with pruning in Matlab is 0.55 against 0.56 without pruning in Python). Therefore, the rest of the study was performed without pruning in the Scikit learn library in Python (Pedregosa *et al.2011*). Accordingly, the results shown do not include pruning.

2.2.3 Bagging regression trees

Another method to avoid overfitting and generally improve the accuracy of decision/regression trees is bootstrap aggregating, also called bagging. The idea behind the method is to resample the dataset multiple times and to build a new regression tree for each resampled dataset. This results in an ensemble of regression trees. The resulting flood damage is then the average of the ensemble of regression trees. Resampling is done by building several datasets by randomly picking records from the original dataset (each record is allowed to be used multiple times in the same dataset). Every resampled dataset therefore randomly leaves out a fraction of the observations and puts more weight on other observations because they are picked multiple times. Bagging regression trees also lead to probabilistic outcomes because the ensemble of trees can be seen as a probability distribution of the outcome.

2.2.4 Random Forest

A random forest is a more advanced variation of bagging regression trees. Apart from building multiple trees with resampled datasets it also randomly excludes a subset of variables at each decision split. This will result in an ensemble of regression trees each based on a different set of damage records and each leaving out a different number of variables at each decision split. For this paper the default settings of Scikit learn are applied, in our case this means 8 variables are left out at each decision split.

2.2.4 Bayesian Network

A Bayesian Network is a type of Probabilistic Graphical Model that represents a set of random variables and their conditional dependencies in a directed acyclic graph (DAG) structure. Each variable in the network may be observed or represented as a prior probability distribution and dependencies between variables are represented with edges representing joint probability distributions. The edges in a Bayesian Network are directed which means there is a direction in which the influence of one variable flows to the other. From this network, inference can be done in order to use knowledge of one variable to make predictions about other variables.

Bayesian Networks and Probabilistic Graphical Models in general are used in many different fields, such as bioinformatics (e.g. Mourad et al. (2011), image processing (e.g. Sudderth & Freeman, 2008) and speech recognition (e.g. Bilmes, 2002). Recently, they have also been applied to flood damage modelling (Vogel et al., 2014; Schröter et al. 2014; Van Verseveld, 2014). Schröter et al. (2014) found that their performance is often better than that of the different types of tree methods. Furthermore, a Bayesian Network can give its result as a probability distribution and does not require information about each variable in order make predictions. If fewer variables are available, the Bayesian Network handles this by adjusting the probability distribution of the outcome. This makes it ideal for transfer of models to other locations where less data is available than for the location where the model was originally based on. Furthermore, it returns (for each object) probability distributions rather than deterministic values, which is valuable for assessing uncertainties within the damage model estimates.

A Bayesian Network can be discrete, continuous or a combination. In this paper fully discrete Bayesian Networks are used, in which all variables are discretized into bins. Given a network the probability of particular set of discrete variable values can be calculated with the following formula:

$$P(X_i, \dots, X_n) = \prod_{i=1}^{n} P(X_i | parents(X_i)) \quad (3)$$

Where X_i are the variables and $parents(X_i)$ is the set of variables directed to X_i. The probability of a single variable value can be obtained by taking the sum of all the probabilities that contain the variable value of interest. The conditional probabilities are stored in conditional probability tables (CPTs). These tables show, for each combination of parent variable values, the probability of each possible output value.

A data-driven Bayesian Network can derive all its CPTs from the data and even derive its graph structure from the data. For this paper, two Bayesian Networks were made: A data-driven Bayesian Network with both the graph structure and the CPTs derived from the dataset and an expert network where the graph structure was estimated in an expert session but the CPTs were derived from the dataset. All calculations were done with a Python library called libpgm (Cabot, 2012). This library follows the methodology described in Koller and Friedman (2009).

The CPTs are learned with maximum likelihood estimation. This method estimates the (joint) probability distributions based on the number of observations. The discretisation assumptions have an impact on the maximum likelihood estimation. If the variables are discretised into a large number of bins more possible combinations of states are possible. These combinations of states grow exponentially with the number of bins of the parent variables. A too fine discretisation therefore quickly leads to more possible states than available data points. This results in a poor performance of the maximum likelihood estimation. Koller and Friedman (2009) call this one of the key limiting factors in learning Bayesian Networks from data. A too coarse discretisation on the other hand is also not desirable because it limits the precision of the Bayesian Network. For this study a balance was found by trying several discretisation resolutions until the best result was found based on the MAE criterion.

Discretisation was done by splitting the data into bins with an equal number of data points in each bin. This works better than making equal sized bins because of the large extremes in especially the damage data. Equal sized bins would either increase the number of bins, which is detrimental to the maximum likelihood estimation (having bins that contain no observations), or the bins would be so large that a majority of the data points would end up in the same bin, which would limit the Bayesian Network performance. The number of bins per variable was chosen based on the performance of a test set on the MAE criterion. This was done by varying the discretisation of the most important variables until the smallest error was found. For the Bayesian Network with the data-driven structure the number of bins chosen was slightly larger, because the network is less complex than the expert network.

The performance of the Bayesian Network on the testing data can be sensitive for discretisation. There are two possible alternatives for the discretisation method applied in this paper: An optimisation algorithm could be applied to determine the optimal discretisation, or a continuous Bayesian Network could be used (Friedman and Goldszmidt, 1996). Apart from solving the discretization problem the advantage of a continuous Bayesian Network is that it would probably perform better in predicting extreme values but a disadvantage is that the Bayesian Network is restricted to specific families of parametric probability distributions (Friedman and Goldszmidt, 1996). An optimization algorithm for the discretization can minimize the error produced by the discretizing but does not solve the fundamental problem of having too few data points.

The data-driven structure is also learned with the libpgm Python library. This library is using a constrained-based approach for structure learning, as is described in Koller and Friedman (2009). In a constrained based approach the structure is learned by calculating dependencies and conditional dependencies among the variables. When two variables are dependent regardless of what they are conditioned by, an edge (connection) is formed. The algorithm follows this procedure to create the entire network. The result is shown in figure 4 (left).

As an alternative to the data-driven structure a structure was also made in an expert meeting involving several Deltares flood damage/Bayesian Network experts (see acknowledgements). In the expert meeting the network was constructed based on a combination of expert judgement/logic and with the knowledge of figure 3 in this paper. The experts focused mainly on edges that they thought are relevant for estimating the flood damage. The result is shown in Figure 4 (right).

The relationship between the total, structural and content damage is known and not probabilistic: total damage = structure damage + content damage. Also, in our case the structure damage, content damage and total damage are always all dependent variables. Therefore, using a Bayesian Network to model this exact definitional relationship could only introduce extra errors and not add anything extra explanation.Therefore in the expert network it has been decided not to use the total damage variable. Instead the total damage is calculated as the sum of the expected value of the structure and the content damage. In the data-driven network the structure damage was not included by the algorithm. Therefore, the total damage variable itself is used for the data-driven network.

The advantage of an expert based network is that experts focus on the connections that matter most rather than on all possible connections. Furthermore, experts can include connections that are not found in this dataset but are expected to exist in theory or in an independent test set. The advantage of a learned network is that new and previously unknown relationships between variables can be discovered.It is expected that the Bayesian Networks in this manuscript are not very sensitive to overfitting during the CPT learning. Koller (2008) only mentions overfitting in the maximum likelihood estimation of Bayesian Networks in relation to discretization that is too fine and offers no techniques to counter overfitting in the maximum likelihood estimation. This expectation that overfitting isn't an issue was tested by testing the Bayesian Network on its own training data. If overfitting is an issue the model should do much better in predicting its own data then in predicting new data. This isn't the case (for the expert model) the MAE is even slightly worse when calculated on its own data (0.622), the correlation coefficient and R^2 are only slightly better (0.24 and 0.04) and only the mean bias error (MBE) is significantly better (-0.015). See results section for comparison.

2.3 Variable importance

In order to investigate the value of more data it is interesting to study the contribution of the different variables to the prediction accuracy. This can be done with bagging trees and the random forests methods. In the sci-kit learn library the "feature importance" (variable importance) can be returned. This importance is the (normalized) total reduction of the mean square error brought by the different variables. This feature importance can be calculated for all the regression trees in the ensemble and a general importance is computed by the sci-kit learn library by taking the average of the feature importances in the tree. This was applied in this study for the bagging trees. The variable importance has been separated for predicting the importance of the total damage, structural damage and the content damage. For the calculation of the variable importance the dataset is used in which the average per postal code is used for the new variables. The water depth rank is not used because it could transfer some of the importance of the original water depth value to the new variables.

3 Results

3.1 Model comparison

The different models are tested on a test set that was not used for training the models. Four indicators are used to rate the performance of the models: Mean Absolute Error (MAE), Mean Bias Error (MBE), the Pearson correlation coefficient, and the coefficient of determination (R^2). The MAE is the mean absolute error divided by the average damage, so a smaller MAE is a better model. The MBE is the average error, this differs from the MAE in that an overestimation is able to correct for an underestimation and the other way around. A low MBE shows that the sum of a large number of predictions will probably be very accurate. The Pearson correlation coefficient is a measure of the linear dependence between two variables. This measure is used to compare the predicted damages with the actual damages in the test set. A Pearson correlation of one means a perfect correlation, zero means no correlation and minus one a perfect inverse correlation. R^2 is the predictive capacity of a model compared to just using the average damage as a prediction. If the R2 is zero it means the independent variables add no predictive capacity compared to just using the average. When R2 is 1 it means the independent variables can explain all variation in the dependent variable. Table 3 shows the results for the different models.

Table 3 shows that given that the models can use all data, random forest and bagging regression trees perform best and equally well. These two methods reduce the MAE by 12% compared to a reference model using the same data (multi-variable linear regression). Bagging regression trees and Random Forest do perform significantly better than normal regression trees, as was also noted by Merz et al. (2013) for flood damages in Germany. Random Forest and Bagging regression trees also outperform the Bayesian Networks. The normal regression tree also works better than the Bayesian Networks. This contradicts earlier findings by Schröter et al. (2014), who found that in most cases Bayesian Networks outperformed the regression trees. Schröter et al. (2014) did however have a very different dataset from the one applied in this study.

Many explanations are possible for the relatively poor performance of the Bayesian Networks. The discretization of the data is a possible problem. Some trends could be too subtle to be captured by the rough discretization, but not enough data points are available for a more precise discretization. Perhaps there still is some space for improving the discretization, for example by applying an optimization algorithm to pick bin definitions in such a way that the available information is applied optimally (Vogel et al. 2012 applied such an algorithm). Another possible reason is that Bayesian Networks might be more sensitive to low quality data in combination with a small dataset. Some of the CPTs applied in the Bayesian Networks here are large and conditional probabilities are based on a relatively small number of observations. Some wrong observations may then have a relatively large impact on the damage prediction.

In the data-driven network the variable of interest (total damage) in our test is only influenced by the water depth. This is because the water depth relative to the ground floor is known while the content damage is not known, so the known water depth blocks all the influence of other variables and the unknown content damage has no influence because it is unknown (it is a target variable). The data-driven Bayesian Network is therefore in our test in practice only dependent on the water depth.

So the structure learning decides to ignore the other variables when the water depth relative to the ground floor is available. This is probably because the data-driven structure algorithms finds all variables equally important and therefore draws only the most important edges (connections) regarding the total damage. Other methods (e.g. as described by Riggelsen, 2008) for structure learning might be able to give better results.

The multi-variable linear regression reference model does a good job on the MBE but is clearly weaker on the other performance indicators, which shows that for predicting aggregate damages for e.g. policy studies, the more complex methods are less beneficial. This is different in cases where individual building damages are important, for instance for insurance rating purposes.

The reference root function has a very large bias compared to the other models. This is probably because the shape of the root function is inappropriate for this flood event.

3.2 Benefits of more data

The models were trained with different numbers of variables to see whether the additional data is valuable. As expected, the best performing model with a high number of variables always performs significantly better than the best performing model with fewer variables. More data therefore seems to add potential value to the damage prediction despite the possible quality issues in the additional data. The MAE of the best performing model with only the water depth (regression tree) can be reduced by a further 14% by the best model using all data (Random Forest). The MAE of the root function fitted to the data (representing common practice) can be reduced by about 20% using the Random Forest with all data.

The method to join the extra data with the original data based on water depth rank is not effective. Just taking the average value per postal code appears to work better. The water depth rank probably sometimes assigns extreme variable values to the wrong objects which disturb some correlations in the data.

3.3 Variable importance

The total importance of variables that were added in this study is about 30% (figure 5), the added variables therefore clearly add to the prediction accuracy. This assessment was done without the water depth ranking join because this could assign some of the importance of the original water depth to the modelled water depth. The original water depth is by far the most important variable. Construction year is an important variable for the structure damage but not for the content damage. This is as expected. Household size is quite important for the structural damage but insignificant for the content damage. This is less obvious but it could be that large families live on average in larger houses but do not have much more valuable contents on the ground floor. Return period is an important variable for both the structure and the content damage. This was also expected because the population in areas that flood more frequently are expected to have more flood experience, thus resulting in better preparedness and lower damages. This effect is visible in the data, with return period having an importance of about 10%.

4 Discussion and conclusion

Additional data improves flood damage modelling relative to a test set, even if this data comes from a collection of different sources and is of limited quality (error prone). The supervised learning algorithm is also important. Given the same data there are large differences between the algorithms. Random Forests and bagging regression trees perform significantly better than normal regression trees and multi-variable linear regression. The Bayesian Networks perform poorly compared to any of the tree based methods.

Our current approach doesn't show that the additional variables are beneficial for the Bayesian Networks. However, because the tree methods can benefit from the additional data it is likely that in some cases Bayesian Networks could also. The poor performance of the Bayesian Networks contradicts earlier studies (Schröter et al., 2014) and could be due to the discretization method, quality of the expert network, network learning algorithm or problems with data quantity or quality.

The test set that was applied in this paper for the validation of the model, was randomly selected from the data and consistently applied among all models. The indicators for model performance would have been more accurate if some form of cross-validation was used instead of a single test set. Expectations are that this would cause minor shifts in results but that it would not influence the conclusions of this paper.

This paper did not address another benefit of Bayesian Networks, Random Forest and Bagging trees, which is the incorporation of uncertainty. Bayesian Networks do this explicitly in the method and for Bagging Trees or Random Forest each tree can be seen as a possible damage estimate and together the trees represent a probability distribution.

The methods applied in this manuscript provide an uncertainty estimate for a single object. For policy decision making it is often useful to aggregate these uncertainty estimates to a total uncertainty for the entire flood event. This can be done with the assumption that all objects are perfectly correlated to each other (one tree will apply to the entire event but what tree is uncertain), or with the assumption that all objects are independent of each other (each object will have a different tree but what tree is uncertain). Both assumptions are however not completely correct (Wagenaar et al., 2016). The Bayesian Network framework might offer a middle way to model this correctly. If each object has a copy of the original Bayesian Network, and these Bayesian Networks are linked together based on the location of the objects, it can be explicitly taken into account that nearby objects are more likely to have similar damages. This could be an argument to prefer Bayesian Networks over tree based methods in the future.

The dataset applied in this paper had many limitations. The most important limitation is that the exact locations of the objects are unknown. Because of this, it was difficult to link building and flood characteristics to damage records. An attempt to do this by using water depth rank performed worse than just using the average variable values per postal code. Despite this limitation, the added data still produced significantly better damage estimates. Another problem with the dataset

is the unknown manual adjustment to an unknown share of data (rental residential buildings) for the structural damage records. These actions may have introduced a relationship between structural damage and some of the originally recorded variables that wasn't there in reality. This could in theory cause a slight overestimation in the prediction performance of the models on the test set. This effect on the results is however expected to be small, because most of the prediction improvements came from adding variables that were not available for doing the manual actions in 1994.

This study applied absolute damages rather than relative damages. This requires the supervised learning algorithms to implicitly also predict information about the values at risk besides the vulnerability. The algorithms can do this with variables such as living area, footprint area, building year and household size. This seems less error prone and better than estimating such values at risk with general rules of thumb based on assumptions about construction costs and content value. Such assumptions could cause extra errors, and therefore in this study absolute damages were used.

This paper trained flood damage models on just a single flood event. Ideally training data should consist of multiple events so that the spectrum of possible damages which the model is trained upon is larger. Especially for the transfer to other areas this would be important. Models that are trained on a single event could overfit on this event and this problem would not show up if the model is tested with data from that same event (even if this specific data wasn't used for training the model). A good example of this appears in the good performance of the regression tree based on only the water depth versus the fitted root function based on only the water depth. The root shape of a damage function which many expert models use (see section 2.2.1) and which makes physically sense, is performing much worse than a more flexible model that can adjust to other relationships between damage and water depth. This is explained by figure 6 which shows a downward sloping damage function after 90cm of water depth, a shape very different from damage functions normally found in the literature. The root function model therefore starts producing large errors after 90 cm while the regression tree can capture this trend well. This downward sloping makes physically no sense but could be explained by other variables such as return period. Return period could be a proxy for flood experience and better preparation because houses that experienced large flood depths in 1993 are probably on lower ground and also experience floods in general more often. This relationship is probably not true for other types of events, for example large flood depths due to dike breaches. So in that sense, the regression tree is overfitting on this single flood event.

Supervised learning can help to create and improve flood damage models. They have many theoretical advantages over deterministic damage functions based on only the water depth. The application of supervised learning in flood damage modelling remains challanging in practice, because of limited data availability. In this paper we utilized different data sources compared to previous studies to acquire this data and showed that also on this dataset the methods are beneficial, especially the tree based methods. Future work may merge available datasets from different events and from different countries in order to develop a model that can be applied using several hazard variables, and which also works in circumstances outside areas for which flood damage data is available.

Acknowledgments

We thank our colleagues Kathryn Roscoe for advice on the Bayesian Networks and our colleagues Karin de Bruijn and Marcel van der Doef for their input in constructing the expert Bayesian Network. Constructive comments and suggestions from two anonymous reviewers helped to improve this paper. This research has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement number 641811 (Improving predictions and management of hydrological extremes – IMPREX), see also http://www.imprex.eu.

References

Becker, A., 2012. Maas-modellen 5de generatie: Modelopzet, kalibratie en verificatie. Deltares rapport 1204280-000-ZWS-0011 (in Dutch).

Bilmes, J., 2002. Graphical models and automatic speech recognition. The Journal of the Acoustical Society of America 112, 2278 (2002); doi: http://dx.doi.org/10.1121/1.4779134

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

Cabot, 2012. Libpgm Python library., pythonhosted.org/libpgm, website accessed 10-08-2016: http://pythonhosted.org/libpgm/

Diermanse, F.L.M., 2004. HR2006 – herberekening werklijn Maas. Delft Hydraulics Q3623.00 (in Dutch).

EC, 2007. DIRECTIVE 2007/60/EC OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 23 October 2007 on the assessment and management of flood risks. Official Journal of the European Union. L 288/27

Egorova, R., van Noortwijk, J. and Holterman, S., 2008. Uncertainty in flood damage estimation. International Journal River Basin Management 6, nr. 2 (2008): 139-148.

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

Friedman, N. and Goldszmidt, M., 1996. Discretizing continuous attributes while learning bayesian networks. In Proc. ICML, page 157–165

Fuchs, S. , Heiss, K. and Hübl, J., 2007. Towards an empirical vulnerability function for use in debris flow risk assessment. Nat. Hazards Earth Syst. Sci., 7, 495–506, 2007

Gerl, T., Kreibich., H, Franco, G., Marechal, D. and Schröter, K., 2016. A Review of Flood Loss Models as Basis for Harmonization and Benchmarking. PLoS ONE 11(7): e0159791. https://doi.org/10.1371/journal.pone.0159791

HOWAS 21, website accessed 03-08-2016: http://dx.doi.org/10.1594/GFZ.SDDB.HOWAS21

Islam, K. M. N., 1997. The impacts of flooding and methods of assessment in urban areas of Bangladesh. PhD Thesis. Middlesex University

Jongman, B., Kreibich, H., Apel, H. B., Bates, P., Feyen, L., Gericke, A., Neal, J., Aerts, J.C.J.H., Ward, P. Comparative flood damage model assessment: towards a European approach. Natural Hazards and Earth Sciences, 12, 3733-3752. 2012

Jonkman, S.N., M. Bockarjova, M. Kok, P. Bernardini (2008). Integrated hydrodynamic and economic modelling of flood damage in the Netherlands. Ecological Economics, 66, 77-90.

Kadaster website. Accessed 25-10-2016: https://www.kadaster.nl/bag

Kind, J.M., 2013. Economically efficient flood protection standards for the Netherlands. Journal of Flood Risk Management, 7(2), 103-117.

Kok, M., Huizinga, H.J.,Vrouwenvelder, A.C.W.M. and van den Braak, W.E.W., 2005. Standaardmethode 2005, Schade en Slachtoffers als gevolg van overstroming. HKV, TNObouw, Rijkswaterstaat DWW.

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

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

Messner, F., E. Penning-Rowsell, C. Green, V. Meyer, S. Tunstall, A. van der Veen, 2007. Evaluating flood damages: guidance and recommendations on principles and methods. Floodsite report T09-06-01.

Mourad, R., Sinoquet, C., Leray, P., 2011. Probabilistic graphical models for genetic association studies. Brief Bioinform (2012) 13 (1): 20-33. DOI:https://doi.org/10.1093/bib/bbr015

Papathoma-Köhle, M., Zischg, A., Fuchs, S., Glade, T., Keiler, M., 2014. Loss estimation for landslides in mountain areas – An integrated toolbox for vulnerability assessment and damage documentation. Environmental Modelling & Software 63 (2015) 156-169.

Papathoma-Köhle, M., 2016. Vulnerability curves vs. vulnerability indicators: application of an indicator-based methodology for debris-flow hazards. Nat. Hazards Earth Syst. Sci., 16, 1771–1790, 2016

Pedregosa F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit learn: Machine Learning in Python. Journal of Machine Learning Research 12 (2011) 2825-2830

Penning-Rowsell, E.C., C. Johnson, en S. Tunstall. The benefits of Flood and Coastal Risk Management: A Manual of Assessment Techniques. Middlesex University Press, London, 2005.

Pistrika A.K., Jonkman S.N., 2009. Damage to residential buildings due to flooding of New Orleans after hurricane Katrina. Natural Hazards. Vol. 54 Issue 2, pp. 413-434

Riggelsen, C. 2008. Learning Bayesian Net-works: A MAP Criterion for Joint Selection ofModel Structure and Parameter. In ICDM, 2008 Eighth IEEE International Conference on Data Mining, pages 522-529.

Rijkswaterstaat, 2013. User's Guide WAQUA: General Information. Version 10.59, October 2013 (in Dutch).

Scawthorn, C., Flores, P., Blais, N., Seligson, H., Tate, E., Chang, S., Mifflin, E., Thomas, W., Murphy, J., Jones, C., and Lawrence, M.: HAZUS-MH flood loss estimation methodology II. Damage and loss assessment, Natural Hazards Review, 7, 72–81, 2006.

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

Sluijs, L., M. Snuverink, K. van den Berg, en A. Wiertz., 2000. Schadecurves industrie ten gevolge van overstromingen. Tebodin. Rijkswaterstaat DWW.

Spekkers, M. H., Kok, M., Clemens, F. H. L. R., and ten Veldhuis, J. A. E., 2014. Decision-tree analysis of factors influencing rainfall-related building structure and content damage, Nat. Hazards Earth Syst. Sci., 14, 2531-2547, doi:10.5194/nhess-14-2531-2014.

Sudderth, E., Freeman, W., 2008. Signal and Image Processing with Belief Propagation. IEEE Signal Processing Magazine Volume: 25, Issue: 2, March 2008. DOI: 10.1109/MSP.2007.914235

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

Thieken, A.H., Olschewski, A., Kreibich, H., Kobsch, S. and Merz, B., 2008. „Development and evaluation of FLEMOps - A new flood loss esimation model for the private sector." WIT Transactions on Ecology and the Environment 118 (2008): 315 -324.

Van der Most, H., Tanczos, I., De Bruijn, K.M., Wagenaar, D.J., 2014. New, Risk-Based standards for flood protection in the Netherlands. 6th International Conference on Flood Management (ICFM6), September 2014, Sao Paulo, Brazil.

Van Ootegem, L., Verhofstadt, E., Van Herck, K., Creten, T., 2015. Multivariate pluvial flood damage models. Environ. Impact Assess. Rev. 2015, 54, 91–100.

Van Verseveld, H., 2014. Impact Modelling of Hurricane Sandy on the Rockaways. Relating high-resolution storm characteristics to observed impact with use of Bayesian Belief Networks. MSc thesis Delft University of Technology - Deltares.

Vogel, K., Riggelsen, C., Kreibich, H., Merz, B., Scherbaum, F., 2012. Flood Damage and Influencing Factors: A Bayesian Network Perspective, in: Proceedings of the 6th European Workshop on Probabilistic Graphical Models (PGM 2012). Presented at the 6th European Workshop on Probabilisit Graphical Models, Granada, Spain, pp. 347–354.

Wagenaar, D.J., 2013. The significance of flood duration for flood damage assessment. Master Thesis, Delft University of Technology.

Wagenaar, D.J., De Bruijn, K.M., Bouwer, L.M. and De Moel, H., 2016. Uncertainty in flood damage estimates and its potential effect on investment decisions. Natural Hazards and Earth System Sciences, 16(1), 1-14.

Wind, H.G., T.M. Nierop, C.J. de Blois, J.L. de Kok, 1999. Analysis of flood damages from the 1993 and 1995 Meuse floods. Water Resources Research, 35, 3459-3466.

WL Delft, 1994. Onderzoek watersnood Maas, Deelrapport 1: Wateroverlast December 1993. WL Delft, Delft (in Dutch).

WL Delft, 1994. Onderzoek watersnood Maas, Deelrapport 9: Schade. WL Delft, Delft (in Dutch).

Table 1: Description of the variables in the flood damage dataset for the Meuse flood of 1993.

	Variable	Unit	Source	Pearson correlation on total damage
td	Total damage	Guilder (1993 value)	Original dataset[a]	1
sd	Structure damage	Guilder (1993 value)	Original dataset[a]	0.85
cd	Content damage	Guilder (1993 value)	Original dataset[a]	0.83
df	Water depth relative to floor	m	Original dataset[a]	0.18
dg	Water depth relative to DEM	m	Flood simulation[b]	0.18
bs	Basement	1=Yes, 2=No	Original dataset[a]	-0.04
dh	Detached house	1=Yes, 2=No	Original dataset[a]	0.08
hs	Household size	Number	Original dataset[a]	0.17
fv	Flow velocity	$m\ s^{-1}$	Flood simulation[b]	0.04
fd	Flood duration	h	Flood simulation[b]	0.05
rp	Return period	year	Flood simulation[b]	-0.09
ba	Building age	year	Cadastre[c]	0.01
la	Floor area for living	m^2	Cadastre[c]	0.04
fa	Footprint area building	m^2	Cadastre[c]	-0.02

[a] WL Delft, 1994

[b] 2D flood simulation data using WAQUA

[c] Basisregistraties Adressen en Gebouwen (BAG), version 2011 (Kadaster website).

Table 2: Results of different models for four indicators: MAE, MBE, R^2 and correlation coefficient. The models had access to all variables (except for the root function). The version of the dataset with the water depth rank join between the old and the new variables is used .

Calculation	MAE	MBE	R^2	Correlation coefficient
Root function	0.612	0.194	0	0.15
Multi-variable linear regression	0.578	0.055	0.07	0.27
Regression tree	0.561	0.065	0.03	0.31
Bagging regression tree	0.504	0.061	0.15	0.38
Random forest	0.508	0.054	0.16	0.39
Data-driven Bayesian Network	0.629	0.525	0	0.21
Expert Bayesian Network	0.607	-0.08	0.03	0.21

Table 3: The best performing model based on the MAE indicator with different number of variables.

Variables	Method	MAE	MBE	R^2	Correlation coefficient
Only water depth	Regression tree	0.564	0.071	0.08	0.306
Only original variables (waterdepth, household size, detached house, basement)	Bagging trees	0.551	0.052	0.07	0.345
All variables (water depth rank join)	Random Forest	0.508	0.054	0.16	0.394
All variables (average postal code join)	Random Forest	0.488	0.035	0.17	0.41

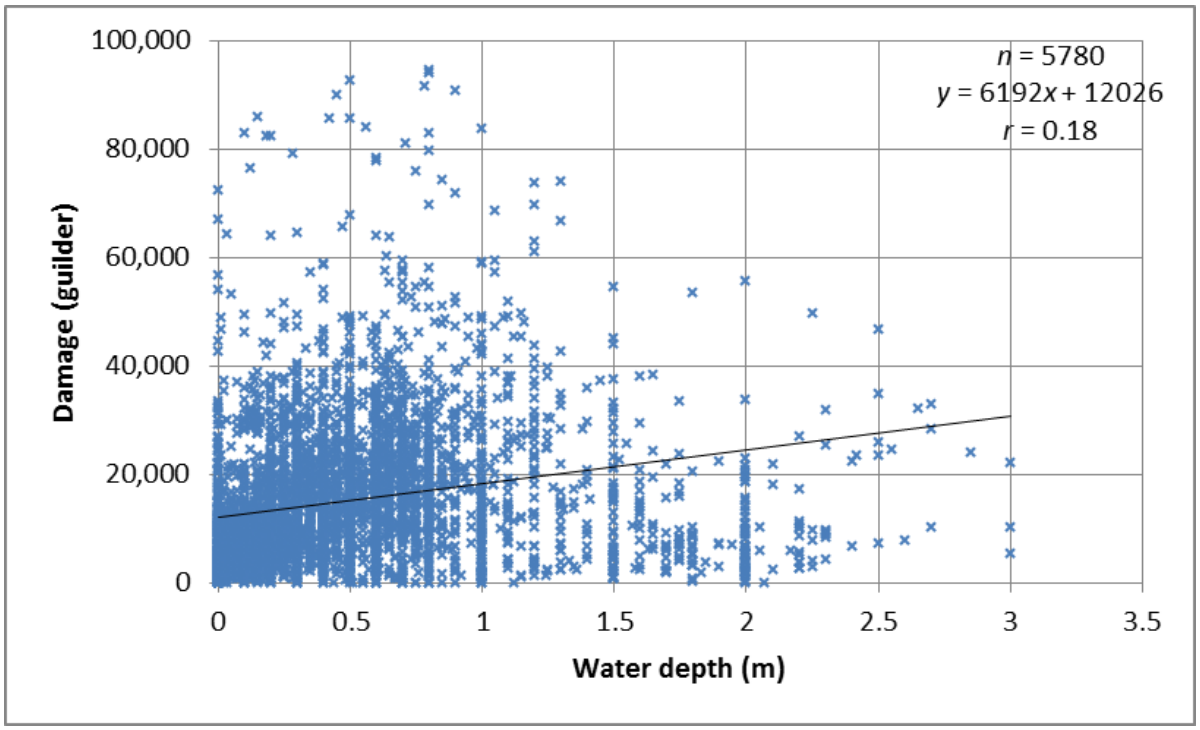

Figure 1: Scatter plot showing the relation between water depth and damage in the original data set..

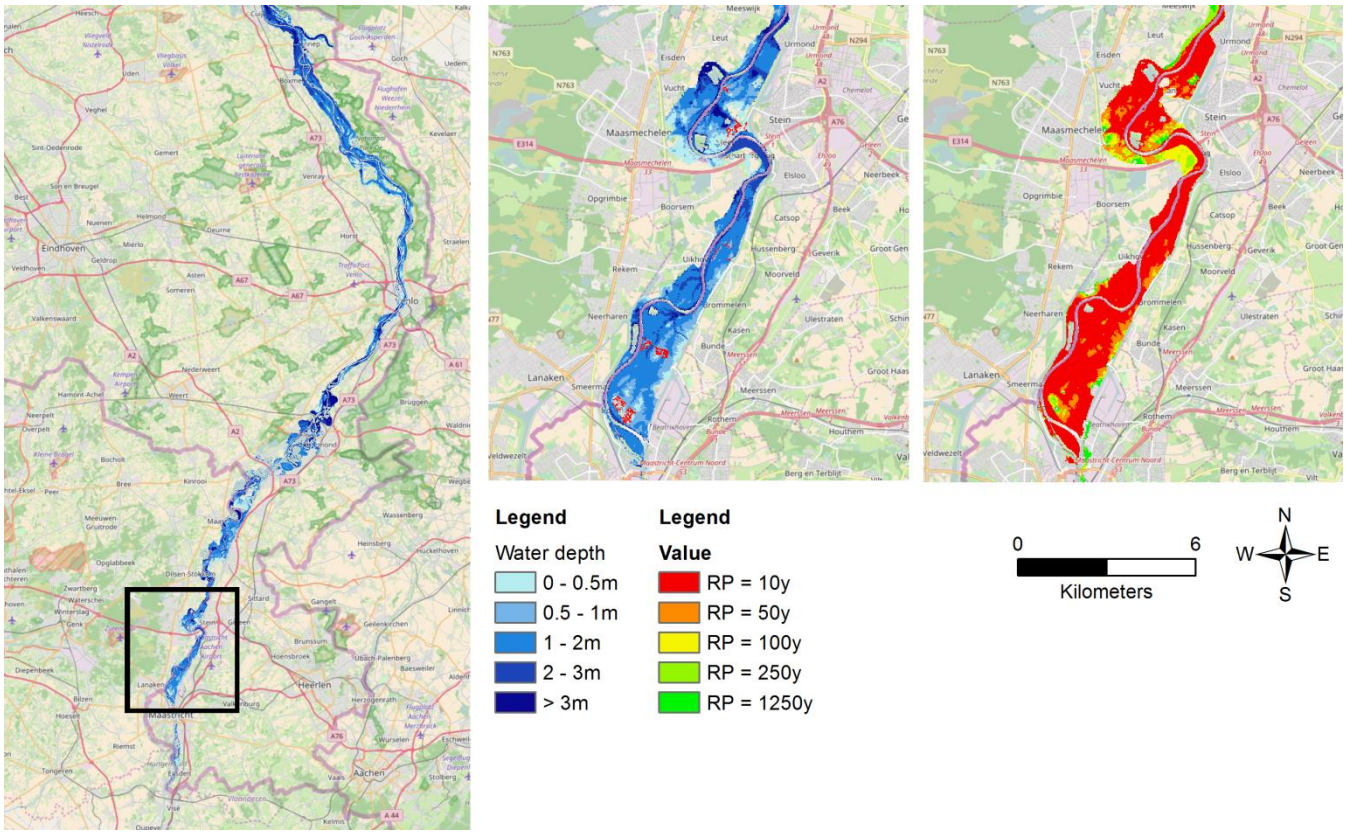

Figure 2: Left the simulated water depth for the entire study area in Limburg. In the center the simulated water depth and affected population (in red) for an example area. On the right the return period at which areas start flooding for the example area. The example area is defined in the box in the left picture.

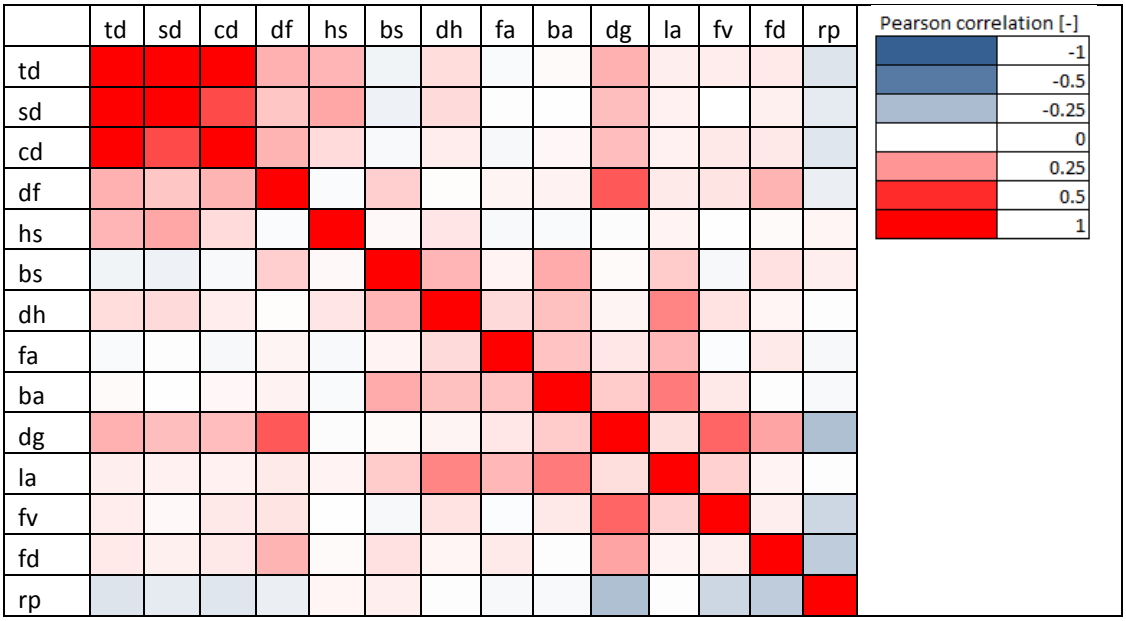

Figure 3: Correlation coefficients between the different variables. See Table 1 for a description of the abbreviations).

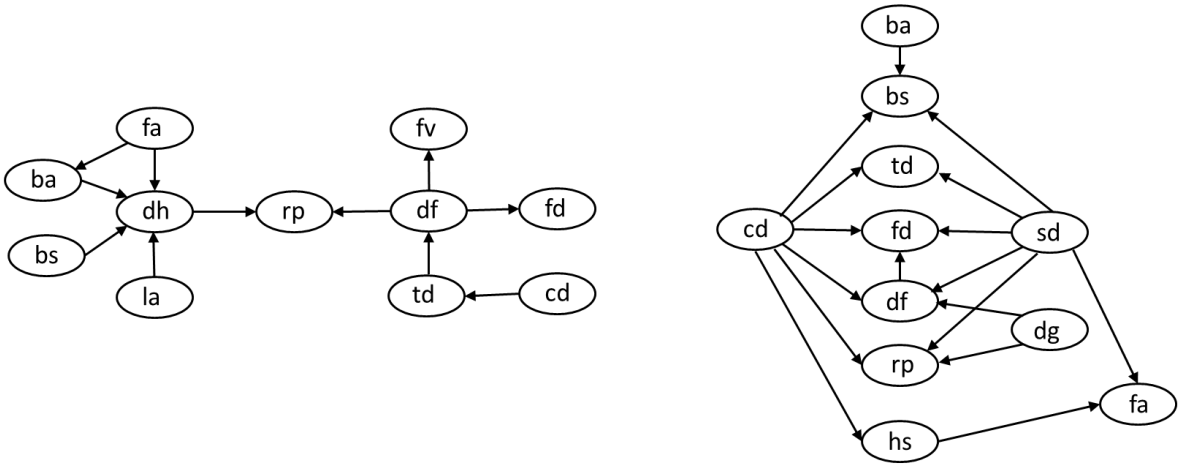

Figure 4: Bayesian Network learned from data (left) and Bayesian Network constructed by experts (right). Note that not all variables are used in the network.

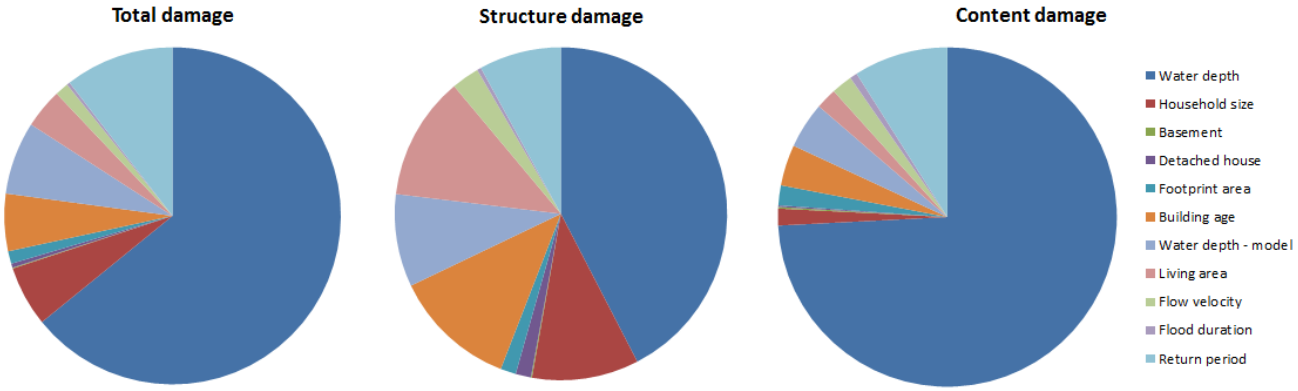

Figure 5: Variable importance: The contribution of different variables in reducing the error in the bagging regression trees.

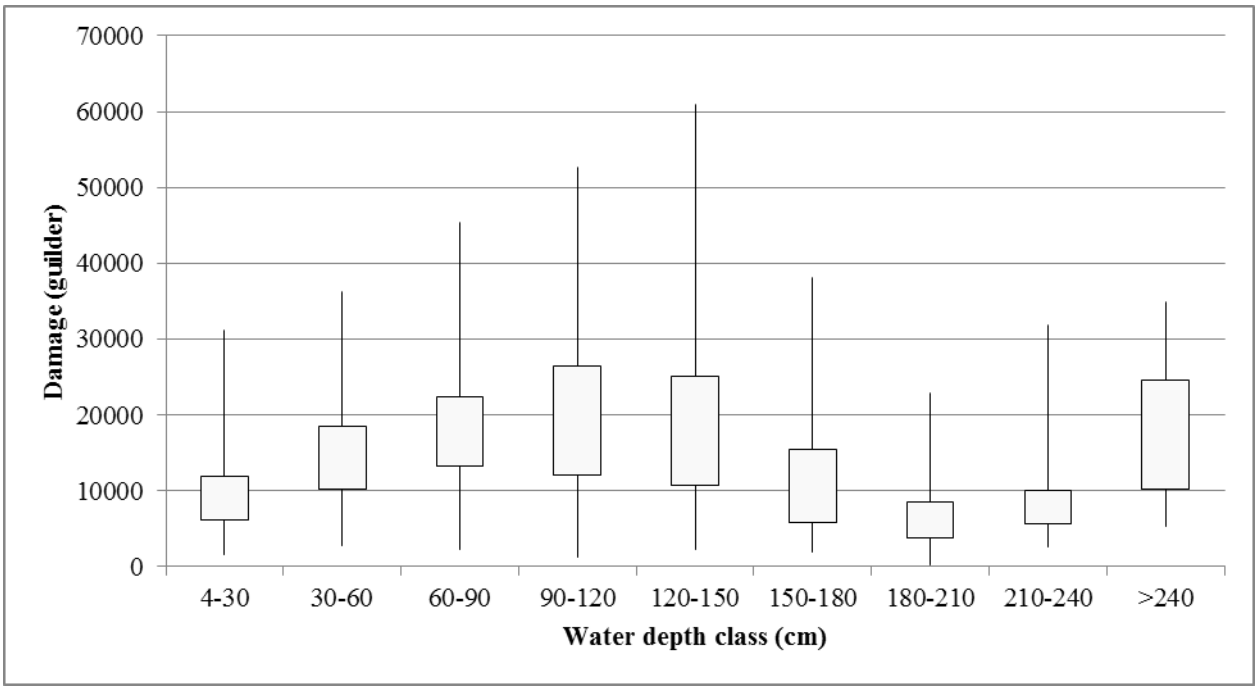

Figure 6: Box-plots of the Meuse flood of 1993 per water depth class. The box shows the 33-66% interval and the lines show the 5-95% interval. The number of observations per water depth class are: 1171, 1221, 865, 247, 96, 109, 87, 30 and 18.

Response to review Referee 1:

Thank you very much for your helpful and detailed comments and suggestions. The number of helpful suggestions, and also detailed comments on the text, is great and much appreciated. They contribute a lot to improving our manuscript. Below, we respond to the individual comments.

1. *The authors may want to reconsider the title of the manuscript. "Data mining" is very prominent in the title, but I think this does not reflect the main focus of this paper. As a matter of fact, the aim of this manuscript is not primarily to do classical data mining on a huge data set (i.e. clustering, anomaly detection, classification), but rather to employ various unsupervised learning algorithms with the aim of finding the best model to explain flood damage with a couple of independent variables (which of course is a part of data mining). Thus, the aim is to compare methodologies for a specific application example, rather than discovering patterns in a huge data set. To emphasize the focus of this work (multivariate flood damage modeling, limited data), I would suggest to rephrase the title to "Multi-variable flood damage modeling with limited data using supervised learning approaches."*

   We agree that the word "Data-Mining" is broad and also covers many things not done in this paper. Our motivation for using it was to follow the terminology used in Merz et al. (2013), an early publication about the application of supervised learning algorithms in flood damage estimation. However, considering the comments about this by both referees and that we actually agree that "Data-Mining" is a very broad term; we propose to change the title as proposed in this review. The new title will be: "Multi-variable flood damage modeling with limited data using supervised learning approaches".

2. *Results and conclusions should be pointed out more clearly in the abstract. Please provide concise information regarding the improvements instead of pointing out a "significant improvement" and mentioning that some models "perform better".*

   We will mention the improvements in the goodness of fit (GOF) values in both the abstract and the conclusions. Also we will mention the exact differences in performance among the different models.

3. *With respect to the presentation quality, I advise to work on the language, on the structure of the manuscript and on the presentation of results. The "common thread" and the main takehome messages are not fully clear and concise throughout the manuscript. The discussion is relatively short, even though there are plenty of interesting aspects in this research that would be worth discussing, and that need to be discussed against the background of uncertain input data. Some formulations are too difficult to understand from a linguistic point of view. For instance p2, l23: "More commonly available (although still rare) are simple datasets that hold records with the flood damage that occurred for each building with sometimes a few other variables (such as location or water depth)."*

   We will go again through the manuscript and look critical at the language. Also we will add more text to the discussion (see more about changes to the discussion section, in our response to referee 2).

4. *The authors might want to think about the reference function. The root function is a simple, univariate function, which serves as a reference for sophisticated mul-tivariate methods. Total damage and water depth are correlated with r = 0.18; I would assume that this value doesn't change significantly when calculating the correlation between total damage and the square root of the water depth. So the reference model is actually a rather bad model, possible improvements regarding the GOF when using more advanced methods seem natural. It might be interesting to include a more sophisticated regression model as a reference (e.g. using LASSO, as this includes both variable selection and regularization).*

The purpose of the reference model is to compare the supervised learning algorithms with what is currently typically applied in flood risk management studies. Very few studies (outside academic research) already apply multi-variable functions and therefore we chose a uni-variate function as a reference. Secondly, many expert damage functions look like a square root function. We therefore believe this is a relevant reference model as it represents what is commonly applied. However, we agree that it is a poor alternative to the techniques later applied in this research. We therefore will add the LASSO technique as a second reference.

5. *Results with respect to themost important variables should be reported in greater detail. It is not clear to me which of the variables are actually beneficial for modeling total damage. The correlation coefficients in Table 1 do not provide any information on that, neither do the other tables or the results section. Variable selection is not discussed at all in this manuscript. For instance, total damage in the Bayesian networks is apparently (c.f. p25, l25-32; Figure 3) influenced by water depth (data-driven network) or water depth and structure damage (expert network), implying that there is no added value of using additional data. Table4 somehow indicates that the increase in GOF is primarily dependent on the algorithms applied, rather than on additional data.*

Variable selection was not one of the initial purposes of this manuscript. The referee is therefore right that no information about it can be found in the manuscript. However, we agree that variable selection provides some important extra information regarding the benefits of extra data. It is therefore very relevant to this manuscript and therefore will add it to the revised manuscript. We will use out-of-bag techniques to say something about variable importance.

6. *Given that the benefits of additional data are emphasized in this manuscript (p11, line 5; p11, line 26-27), it is worth discussing that care has to be taken when introducing additional data to a model. Even though pruning and bagging is mentioned, this topic is not really emphasized in this manuscript. Showing awareness of regularization and penalization of additional variables is of prime importance.*

We fully agree with the referee that potential overfitting is an essential topic (the mentioned methods, regularization, penalization, bagging and pruning, are all methods to avoid overfitting). In the study a lot of care went into avoiding overfitting of the tree based methods. Furthermore, our testing set was not used for training the models, so problems with overfitting would come back in the GOF indicators. If, for example, the different methods

to avoid overfitting would not be used on the regression trees, the GOF would be much worse on the testing data and nearly perfect on the training data. Below, we describe the considerations that we will clarify in the revised paper on the different statistical approaches:

Tree based models

For the tree based methods we avoided overfitting with a minimum data required per tree leaf combined with a maximum number of splits per tree. For the simple regression tree method we compared this technique with pruning and found similar results. The random forest and bagging tree methods are in itself already less sensitive to overfitting, however the same techniques were applied to avoid overfitting. Regularization is not commonly used to avoid overfitting in tree based methods.

Bayesian Network

Overfitting was not addressed in the manuscript. The assumption was that Bayesian Networks are not very prone to overfitting and also the applied library has no settings to avoid overfitting. This assumption seems correct. If instead of the separate test set the Bayesian Network is tested on the learning data, the GOF indicators show no improvement. If overfitting would be a problem the GOF indicators would be much better when the training data is used for testing. This line of argumentation will be added to the Bayesian Network section.

7. *Uncertainty estimation is mentioned as one of the main merits of the methods used in this manuscript (p2, l16). However, uncertainties are neglected in the discussion section of this paper. Even though a number of sources for uncertainty are pointed out (e.g. data collected by different organizations; exact locations of buildings are not known; water depth is only based on estimates and has been questioned by experts; collection methods for variables "inhabitants", "basement" and "attached buildings" are unknown; uncertain join of data based on water depth rank), implications are not discussed.*

Uncertainty estimation was not a purpose of this manuscript, however, uncertainty is an important issue in flood damage estimation and some of the methods applied could help quantify the uncertainty. That is why the uncertainty estimation qualities of the different techniques were presented in the discussion section. The uncertainty in the input data is discussed throughout the manuscript. In the discussion we will add an overview of these uncertainties and discuss the implications.

8. *(a) Table 1: last column should read "Pearson correlation on total damage"(there are 3 different damage variables in the data set).*
(b) Table 2: caption: "...algorithms". Column names for col 2 and 3 need to be more specific, "water depth" is also part of the "original variables". However, the authors may wish to reconsider if this table is really needed, all information presented is the text.

(c) Table 3: It is not clear to which dataset (water depth only / original data set / all variables) these values refer to. I guess it is the data set containing all variables, except for the root function? In addition, the authors may wish to consider adding a GOF-measure for the explained variation (i.e. R2) to the table.

(d) Table 4: It is not really clear how the "best performing" models have been selected – seemingly on the correlation coefficient? Table 3 indicates that RMSE and correlation coefficient of bagging regression tree and random forest show almost identical GOF.

(e) May I propose to combine Tables 3 and 4 by reporting all values (i.e. RMSE, r, and maybe R2 for all 6(5) methods, structured by input dataset). This would also incorporate the information from Table 2.

A) Agree, will be done. B) Agree, we will omit this table. C) We will clarify this table and add the R2 indicator. D) The best performing method has been selected based on a combination of the GOF indicators, in case of similar results only one was picked. This will be clarified in the revised paper and in the cases of similar results this will be mentioned. E) We will add the R2 GOF to both tables. Merging the tables is possible but we feel that this is not very practical, as it would result in one complex table rather than two simple ones. Now the tables neatly address a different question (best performing algorithm; and improvements when more data is added).

9. *(a) Figure 1: please rephrase caption, e.g. "Scatter plot showing the relation between water depth and damage in the original data set".*

(b) Figure 2: It might be interesting to check if plotting the affected houses atop the water depth is easier to understand. It seems that some houses on the left map are not located in the inundated area at all, albeit they are labelled as "affected objects". A comparison is quite difficult, because the map sections are not identical (right map is shifted slightly towards north-west).

(c) Figure 3: The caption should be rephrased – td, sd and cd are no "predictors" (as mentioned at p5, l32).

(d) Figure 4: please rephrase the capture, e.g.: "Bayesian Network learned from data (left) and Bayesian Network constructed by experts (right). Note that not all variables are used in the network."

(e) Figure 5: The authors might reconsider plotting only themean value for each class – boxplots for each bin would be more informative. Please include the number of observations to for each category.

A) We will rephrase it to the suggested phrase. B) It is correct that the left map is a map of all objects in the 1993 situation rather than the affected objects. We can plot them on top of each other only showing the actually affected objects. C) We will change the word "predictors" to "variables". D) We will rephrase this in the suggested way. E) We will use boxplots and add the number of observations.

10. *Please adhere to the journal standards concerning references (see guidelines for authors). References should be formatted accordingly and consistently, and references should be sorted alphabetically.*

We will improve the reference list according to the journal standards.

11. *References to relevant literature are sparse, while the number of references to gray literature is relatively high. Especially sections 1 and 2 would benefit from some additional references.*

We will add some key references to peer-reviewed papers in the introduction section, and to the section describing the different algorithms. For the description of the dataset there is unfortunately mostly gray literature available, except for the paper by Wind et al. (1999) in GRL which we already quote.

12. *Please adhere to the journal standards concerning references (see guidelines for authors). References should be formatted accordingly and consistently, and references should be sorted alphabetically.*

See 10

13. *References to relevant literature are sparse, while the number of references to gray literature is relatively high. Especially sections 1 and 2 would benefit from some additional references.*

See 11

Specific comments:

1. *p1, line 8: "Flood damage assessment is usually done with damage curves only dependent on the water depth." – I would agree that most assessments include water depth as the main determinant of direct damage, but against the background of recent research, I would disagree that it is still state of the art to build flood damage assessments solely on water depth (c.f. Dutta et al., 2003; Kreibich et al., 2005; Thieken et al., 2005; Apel et al., 2009; Elmer et al. 2010; Merz et al. 2013; van Ootegem et al. 2015; Gerl et al. 2016). I would advise to slightly rephrase this sentence, indicating that more sophisticated, multivariate approaches (including hydrological modeling) are on the rise.*

A distinction should be made here between the scientific literature and actual flood risk management studies. We will add a sentence that multi-variable approaches have been carried out in recent academic research, especially in Germany.

2. *p1, line 20–21: "Because flood risk management becomes increasingly risk-based, flood damage estimation is increasingly important in flood risk assess-ment." Please rephrase, this is unclear.*

We will rephrase it into: ”Decision making in flood risk management is increasingly based on studies that quantify the flood risk rather than only the flood hazard. Flood damage estimation is therefore increasingly important.”

3. *p1, line 23: “...flood risk assessments are...”*

We will change this.

4. *p1, line 27: “These models typically predict the fraction of damage...” – the authors may wish to clarify what the denominator of the fraction is by adding e.g. “...as percentage of total possible damage”.*

We will add “as percentage of potential damage”

5. *p2, line 11: “... based on a German dataset based on ...” – please rephrase.*

We will change it into “with a German dataset based on ..”

6. *p2, line 11ff: The authors might want to add some additional references to their literature overview about multi-variate flood damage models. In addition, it might be of interest for the reader to know about the types of covariates used in these studies.*

We will reference to the suggested literature of point 1 and discuss the methods they applied.

7. *p2, line 13: “Spekkers et al. (2014) did something similar ...” – please specify.*

We will change this in: “Spekkers et al. (2014) applied regression trees to estimate pluvial flood damage”.

8. *p2, line 14: “These multi-variable flood damage models have been shown to perform better...” – the authors may want to provide some quantitative indication regarding how much the performance of these multi-variate models exceeded the performance of simple flood damage models.*

We will add some GOF values from Schröter et al. (2014).

9. *p2, line 27: “...that is used here, and previously described...” – please rephrase*

We will change this into: “…which is used here. Previously this dataset has been described in Wind et al. (1999) and in more detail in WL Delft (1994).

10. *p2, line 29: "...very different from the datasets used so far (fewer variables, different sources of variables and different country)." – please explain in more detail. What is meant by "different sources" and why is data from the Netherlands expected to be "very different" from data from Germany? Also, this seems to refer only to the data set used by Merz et al. (2013) and Vogel et al. (2014).*

With different source we mean that the data was collected by insurance experts directly after the floods for compensation purposes and covers all affected buildings. This is different from the German data which was collected a year after the flood for research purposes based on a sample of the affected buildings. The data is also different in that in the original study only a few variables were collected, most of the other variables are added later. In contrast for the German dataset all variables (except return period) were based on telephone interview answers. A few studies also applied datasets different from the GFZ data. These studies did however use different analysis methods (e.g. Dutta et al. 2003) or focused on pluvial flood damage (Van Oostegem et al., 2015 and Spekkers et al., 2014). We will add this explanation to the revised manuscript.

11. *p3, line 11: "The dataset used in this study is based on..."*

We will change this as suggested.

12. *p3, line 12: "...in the Netherlands (WL Delft, 1994)."*

We will change this as suggested.

13. *p3, line 13: 180 km^2*

We will change this as suggested.

14. *p3, line 14: "32% of the damage pertains to residential buildings and content, for this study only the damage to this category is used". – please rephrase.*

We will change this into: "…32% of the damage pertains to residential buildings and content. In this study only residential damage is considered. "

15. *p3, line 14: Please explain briefly why you decided not to consider damage to business and government buildings.*

We will add the sentence: "Other damage categories are not considered because they are more heterogeneous and less data about them is available."

16. *p3, line 17: I think the term "citizen household" is not very common. Maybe replace with "private households"?*

See 17.

17. *p3, line 17ff: Please use a consistent, clear terminology. Distinguishing between "citizen households" (p3, l17), "companies" (p3, l18), "rental residential buildings" (p3, l21) "residential buildings" (p3, l25) and "rental houses" (p3, l26) and "privately owned residential buildings" (p3, l22) is confusing.*

We will change: "rental houses" to "rental residential buildings" and "citizen households" to "privately owned residential buildings".

18. *p3, line 20–23: "The building structure ... content for the same structure." –please rephrase these two sentences to make this more clear.*

We will rephrase this into: "In this set up of the damage collection, the building structure of rental residential buildings was collected by "Stichting Watersnood bedrijven", the organization that collected company damages. This is different from the organization that collected the rest of the residential damages. From the company damages less information was shared to WL Delft (1994), the source of the dataset for this study."

19. *p3, line 23: What is meant by "building content"? Furnishings?*

Building content is a commonly applied term in flood damage literature (both German and US studies use it consistently). However, the reviewer is right that there are UK studies that apply the word furnishings instead. At the first mentioning of the word building content we will add the word furnishings between brackets.

20. *p3, line 25: "The dataset did not include the building structure damage to all rental houses" – It is not clear to me until now, if the data have simply been collected by two different companies (as p3, l17ff imply) or if these two companies have also collected different types of data? Based on the text I assumed that structural damage to rental houses has been collected by "Stitching Watersnood Bedrijven 1993"?*

We will clarify this already a bit earlier (see point 18). Our source for the data, the WL Delft report (1994) combined two different sources for the building data, and in one source (rental houses) the structure damage was available only in some unknown aggregate form. Probably because the rental residential building damage was collected per owner and one owner could own multiple buildings. This reason is however speculation and was therefore not mentioned in the manuscript. The bottom line is that the sum of all building values is known (we

verified this with Wind et al., (1999)), but that the distribution of this value over individual objects is uncertain for a part of the structure damage. The share of rental buildings is however expected to be low in this rural area and therefore we expect this to not substantially affect our results. We will mention these issues more explicitly in the revised paper.

21. *p3, line 27: "Several manual actions were undertaken..." – please explain/provide some insight into what type of actions this could have been.*

We speculate that the organization collecting the data had information on the total structural damage to rental houses, and divided this over the rental objects, based on the number of inhabitants. We will add this in the revised paper.

22. *p3, line 30–31: So, apparently the "manual actions" are not known at all. Please refer to possible impacts of these manual actions on the results in the discussion.*

We will add a paragraph about this in the discussion.

23. *p4, line 6: as a matter of fact, this correlation between water depth and damage is almost negligible. Other studies have found more obvious relationships between water depth and damage (e.g. Merz et al., 2003; Pistrika et al., 2009; Prettenthaler et al., 2010). The assumption that water depth as the main determinant of direct damage does not seem to hold in this case. Please discuss possible reasons for this weak correlation in the discussion (is this only due to the questionable quality of the water depth data mentioned at p4, l4?).*

Given our large dataset size (about 4000+ records) the correlation with water depth is not high, but not negligible either. The correlation coefficient is about half of what other studies found, however in looking at the variable importance we see that water depth is still by far the most important variable. The point of this study is that even though the correlation coefficient is weak and the dataset has some issues (as in many other cases around the world), we can still get significantly better damage estimates with this "limited data". We will emphasize this a bit more in the abstract and the conclusion of the paper. However, our estimates have not improved as much as one would have hoped, and this point will also be added to the discussion.

24. *p4, line 9: "However, this data is not described..."*

We will change this as suggested.

25. *p4, line 23: "... and 40 meters."*

We will change this as suggested.

26. *p5, line 1: The authors may wish to explain shortly how return levels are computed.*

P5 line 1-6 already contains this explanation. We will clarify that this return period here differs from the return period variable in the GFZ dataset, in that we use the return period for any flood at the specific object location and not the return period of the flood that actually occurred. Our hope is that this return period is a good proxy for flood experience of the population, while in the GFZ dataset it says something about the magnitude of the flood. This context will be added before the explanation of how the return periods are determined to clarify the explanation to the reader.

27. *p5 line 5: I do not understand why Figure 2 would show that most of the area floods frequently. Isn't this just a map about water depth?*

Correct, we removed the information on flood frequency from the draft paper, but the text remained. However given the previous comments on the return period variable (comment 26), we will add the map again to the revised paper so that the reader has a better understanding of the meaning of this variable.

28. *p5, line 17: "The method of joining cadastre objects with damage records within a postal code area based on water depth rank is error prone." – This is a quite straightforward approach, which is understandable given the lack of further information. However, this join is probably linked with relatively high uncertainty, depending on the spatial resolution of the DEM used and the (uncertain) expert estimation of water depth in the first place. It was mentioned that between 1 and 20 buildings share the same 6 digit postal code (p4, l2), so mismatches are likely to occur in postal codes with a larger number of buildings. The authors are probably right that houses within a postal code area are similar to some extent, but I am not sure if this is true for variables like "household size" or "floor area for living". Are water depths within a postal code area similar, too, or are the ranks clearly distinguishable? The problem in case of a large number of mismatches is, that this just seemingly increases precision of the analysis. It might be worth testing if results change when simply using a mean/median value for all buildings within one postal code.*

This suggestion is appreciated, and we will perform this test.

29. *p5, line 29: "Several data mining (sometimes called machine learning) ..." – please rephrase. Even though these are closely linked and often used as synonyms, data mining and machine learning are not exactly identical. Rather, machine learning is a sub-field of data mining, i.e. data mining is not only restricted to machine learning methods.*

As mentioned in point 1 of the detailed comments, we will remove the word data-mining from the manuscript and apply the term "supervised learning" instead, as helpfully suggested by the reviewer.

30. *p5, line 31: "...based on all independent variables (thus excluding total, content and structure damage)." – please, rephrase. This might be confusing to some readers, as the BN (Figure 4; p9, l34) includes content damage and structure damage.*

We will omit the part "based on all independent variables (thus excluding total, content and structure damage). ". This addition is indeed so obvious that it can be confusing.

31. *p6, line 6: "...because many damage functions in the literature have this shape".– please provide references, additional to Merz et al. (2012).*

In Wagenaar et al. (2016) there is a figure with damage functions from different studies. Most of the damage functions have approximately the root function shape. For example, HAZUS (Scawthorn et al. 2006), MCM (Penning-Roswell, 2005), Tebodin (Sluijs et al., 2000) and Flemo (Thieken et al., 2008). We will add this to the revised paper.

32. *p6, line 8: may I suggest to use different variable names (variable names withsubscripts, e.g. dt for total damage) in the formula? df is a common abbreviation for degrees of freedom.*

We thank the reviewer for the suggestion to use more consistent abbreviations. df stands for "depth relative to floor", we will change this into "wdf", adding the "w" of water. The use of sub-scripts then would not be necessary.

33. *p6, line 8: "...to get the smallest possible error based on the total damage and water depth data. The optimization of the coefficients is done with the Python package SciPy" – please rephrase and clarify (e.g. "...are optimized using ordinary least squares estimation from the Python package SciPy").*

We will change this into: "The values of the coefficients are optimized for the best fit with the ordinary least squares method. This is done with the Python package SciPy.

34. *p6, line 15: "However, it is more common to ..."*

We will change this as suggested.

35. *p6, line 19: "...with 11 variables for each damage record."*

We will change this as suggested.

36. *p6, line 21 "...reduces maximally..." replace with "...is minimized..."*

We will change this as suggested.

37. *p6, line 22 and p6, line 25: "...by calculating the MSE reduction for all..." and "...is the reduction in MSE of total damage ..." ("MSE error" is redundant).*

We will remove the second "MSE error"

38. *p6, line 23: abbreviation MSE is already explained in p6, l20.*

We will just use the abbreviation here.

39. *p6, l. 24ff: please try to integrate the formula and the explanation of variables more naturally into the flow text. The sentence "The regression tree... (Pedregosa et al. 2011)." might be added at the end of the page.*

We will rewrite this part as suggested.

40. *p7, line 10: "the Matlab Statistical Toolbox (Matlab website)" – replace with "Matlab's 'Statistics and Machine Learning Toolbox"'*

We will change this as suggested.

41. *p7, line 11: "Python libraries do not support pruning" – I think there are custom implementations of pruning in Python. The authors might want to look at sgenoud's fork of the scikit-learn package at github.*

The main version of Scikit learn (well-known Machine Learning library in Python) doesn't support pruning. An internet search didn't yield any alternative in major libraries that do support this. With some effort we might be able to find and use a GitHub implementation of pruning in Python by someone. However, we did successfully run the pruning algorithm in Matlab and found no better results than in Python without pruning (and using alternative methods to avoid overfitting). Furthermore, pruning is mostly relevant for traditional regression trees and not for Random Forests or Bagging trees. Traditional regression trees are currently far from the best performing algorithm

and including pruning is therefore not expected to influence the conclusions of this paper in any way. We therefore would not apply a Python implementation of pruning for this paper.

42. *p7, line 11: "performance of pruning was similar" – can you provide some information about the method and results of the comparison?*

We compared them based on the RMSE indicator. We will run the Matlab script again to get the exact values, and report these in the revised paper.

43. *p8, line 9: the authors might want to add references to these fields of application.*

We will look up references for applications in the different fields.

44. *p8, line 27: please cite the URL as a normal reference, i.e. "All calculations were done using the Python library libpgm (Cabot 2012)."*

We will change this as suggested.

45. *p9, line 6: "...balance was found by trying several discretization resolutions in order to gain the best results." – please rephrase and add more concise information ("...trying several discretization results until the best solution was found based on xxxxx criterion")*

We will rewrite this. The criterion was the RMSE. We changed the number of bins the different variables are divided in, and calculated after each change the RMSE. We then applied the number of bins with the smallest RMSE. This action was done by hand and not by an algorithm (hence manual).

46. *p9, line 13: "This was done manually by varying the discretization of the important variables until the smallest error was found" – this is rather vague. What is "manually"? What do you mean by "important variables" and what is the "smallest error"?*

See 45.

47. *p9, line 7–15: please make this paragraph more concise*

We will rewrite that paragraph in a more clear and structured manner. The content of the paragraph is however highly relevant and the rewrite will focus on style only.

48. *p9, line 28–32: please rephrase, focus on methodology and advantages/disadvantages of a manually established network. Contributing experts other than the authors should be added in an "Acknowledgements" section rather than in the text.*

We will add a more detailed discussion on the advantages/disadvantages of expert networks versus learned networks. The main advantages of an expert network are that the overfitting problem is less relevant and that experts take into account the variables/connections that are practically important. Advantage of a learned network are that new and previously unknown relationships between variables can be discovered. Also, we will add the experts contributing to the expert network (and not being authors of the paper) to the acknowledgements section, as suggested.

49. *p9, line 33 – p10, line 1: "The total damage is … and the content damage." Please explain in more detail, this is not fully conclusive to me.*

We will rephrase this into: "The relationship between the total damage, structural damage and content damage is known and not probabilistic: total damage = structure damage + content damage. Also, in our case the structure damage, content damage and total damage are always all dependent variables. Therefore, using a Bayesian Network to model this exact definitional relationship could only introduce extra errors and not add anything extra explanation."

50. *p10, line 5: please provide information about important independent variables within the results section.*

We will add this analysis. See point 5 of the main points.

51. *p10, line 15 : "(with different better training data)" – please rephrase*

We will remove the section between brackets and replace this with a new sentence. "Schröter et al. (2014) used another dataset, with more variables per damage record and applied more reliable collection methods".

52. *p11, line 1: the authors may wish to put section 3.2 into the discussion section.*

This is a good suggestion; we will split paragraph 3.2, the first paragraph and the table will remain in section 3. The second and third paragraph will be moved to the discussion.

53. *p11, line 7–8: "The relatively good performance … is striking." – the authors may wish to replace "striking" with "worth noting". Actually, given the rather bad fit of the root function (as explained by the authors in the following*

paragraph) and the concern about overfitting with regression trees, I assume that both the authors and the reader would have expected this behavior, at least to some extent.

We will use the word "worth noting" instead of "striking". The bad fit of the root function is most unexpected here, given that most damage functions look like root functions and given that root functions are a logical relationship between water depth and damage. Our initial thought was that the root function performed bad because it only had the water depth as input. This expectation turned out wrong after seeing the good performance of the regression tree with the same information. At the description of the root function, additional argumentation for expecting a root function will be provided.

54. *p11, line 10: I think boxplots would be a more informative representation for Figure 5. Also, the conclusion of a "downward slope" based only on the means for each class should be interpreted with care. It has to be noted that variance/number of outliers gets smaller for data points with water depth > 1.3m.p11, line 12: This is an interesting peculiarity of this data set. While it seems to be plausible that preparedness effects might mitigate total damage (note the very weak correlation of -0.09 in this case), it is counter-intuitive that return period is negatively correlated with water depth. Basically, events associated with high return periods are rare events with high water depth, i.e. the higher the water depth, the greater the return period. Under the assumption that values for return periods are relatively homogeneous for the Meuse flood (which was one actual event with a certain return period), this would mean that areas with a high water depth get flooded more frequently at relatively higher water depths. Yet, I would assume that they get flooded more frequently, but at lower level. So, in the case of the Meuse flood, areas with high water depth showed lower return periods. Does this indicate possible inaccuracies of the flood return period maps?*

The reviewer misunderstood the meaning of our variable "return period". In this study we used the flood return period at a location for any flood not the return period of the flood that actually occurred. Therefore, there is a lot of variation in the return period variable within our dataset. In point 26 and 27 we propose ways to avoid this confusion. The reviewer is correct that the number of observations in figure 5 is smaller at the higher water depths (however the number of observations remain large). As suggested we will add box plots to solve this problem in figure 5. As for the possible inaccuracies in the flood return period map, these are expected to be insignificant. The absolute values might be inaccurate but in relative terms the return periods are expected to be good (low areas near the river have a frequent return period, high areas far from the river have an infrequent return period). For our study only the relative return periods are important. We will add a return period map so the reader can see that the return periods make sense.

55. *p11, line 21: "...are different from each other in more ways than just the water depth" – please rephrase.*

We will rephrase this into: “it shows that there are relevant differences between floods that cannot be expressed with the water depth variable alone”.

56. *p11, line 22: While overfitting based on a single variable is a valid concern, con-cluding to use multiple variables to avoid overfitting might be erroneous if the use of extra variables is not penalized.*

In fact it’s not the overfitting on the event that is an argument for multi-variable damage functions; it’s the physically unrealistic downward sloping damage function itself. This downward sloping could only occur if some other factor plays an important role. We will rephrase this part of the paper to make this clear. For the penalization see point 6 of the main points.

57. *p11, line 25: rephrase as “Discussion and conclusion”*

We will rewrite this part as suggested.

58. *p11–p12: please work on the discussion section, a large portion of page 12 (l10-l26) is is mainly about potential advantages of BN that are not visible in the results of this study.*

We will shorten the section about potential advantages of Bayesian networks not shown in this manuscript. Many points will be added to the discussion section, see points: 7,21,23 and our response to the comments from the other referee.

59. *p11–p12: please work on the discussion section, a large portion of page 12 (l10-l26) is is mainly about potential advantages of BN that are not visible in the results of this study.*

We will rephrase into: “..but what tree is the correct tree is uncertain”.

60. *p12, last paragraph: please rephrase your final conclusions, this is somewhat clumsy from a linguistic point of view; e.g. split the first sentence into two sentences at the third “and”: “In this paper we utilized different data sources compared to previous studies to acquire this data and showed that also on this dataset the methods are beneficial, especially the tree-based methods” – simplify, rephrase; “One possible way forward is to...” replace with “Future work may include ...”; etc.*

We revise the first sentence in this paragraph using the suggestions of the reviewer.

Review referee 2:

Thank you very much for your thoughtful and interesting comments. They contribute a lot to further improving our manuscript better.

1. *First of all I suggest to change the title a bit since according to my opinion the term "data mining" is a bit misleading in comparison to the work undertaken in this paper. What about just "multi-variable flood damage modelling with limited data"?*

We agree with the referees comment about the title. This is addressed in detail in our comment to point 1 of referee 1. The title suggested by referee 2 is very similar to the title suggested by referee 1, we picked the first suggestion because it's a bit more specific. We will now change the title to "Multi-variable flood damage modeling with limited data using supervised learning approaches".

2. *Second, I have the feeling that some of the existing (and relevant) literature on this topic is not included in the Introduction so far. It would be interesting to see more than the presented references to (mostly) Dutch researchers and the Potsdam group, e.g., by broadening the focus a bit towards works on flooding with sediment transport – here similar problems are described that somehow the deposition height is the only available parameter, but in turn this parameter is not fully representing the processes leading to loss. Examples include the works of Papathoma-Köhle or Fuchs, to just drop some names.*

We thank the referee for the suggestion to widen the scope for the introduction and possibly the discussion. We will mention the vulnerability indicators in Papathoma-Köhle (2016), and Papathoma-Köhle et al. (2014) will be used to very briefly describe the state of the art in landslide vulnerability.

3. *Third, in the discussion on vulnerability of buildings exposed to flood hazards there are some works not comparing direct losses, but the degree of loss, which is a relative measure taking into account the different building values. As such, and I am not com-pletely familiar with Dutch building regulations, different loss heights are also a result of different values of the elements at risk. How did the authors consider this challenge during their analysis (which is also perfectly mirrored by Figure 1)?*

This is an important issue that will be addressed in the revised manuscript. We will mention in the introduction that we aim for predicting absolute damages rather than relative damages. In the discussion we will discuss the advantages/disadvantages of using absolute/relative damages. The values at risk are included indirectly in variables such as living area, footprint area, building year and basement. Making this relative would be useful if exact building values were available. However, since these building values are not available, general rules of thumb would be needed for building values. This would introduce extra errors, and therefore we decided to use absolute flood damages.

4. *Fourth, I kindly would like to suggest that the Results and Discussion (!) sections are more carefully written since so far, the first includes lots of discussion, and the current Conclusion and Discussion section is rather short. This should also include some paragraphs on the uncertainties behind the analysis, as mentioned in the Methods section.*

We agree with the referee. The conclusion and discussion will be revised considerably, as this suggestion was also made by referee 1. The discussion will focus more on the impact of uncertainties in our dataset and the way they might impact the conclusion. The conclusion will focus more on the goodness of fit indicators and the relationship to the limited data. See also our replies to points 7,21, and 23 from referee 1.

5. *Fifth I would like to recommend that the authors show a more detailed situation as the one presented in Figure 2 – the current scale is hardly readable. A possible solution is to show the overall extent as an inlet map and then in the main map just a zoom of the most interesting river section or so. For the legend: the water depths of 0.5, 1.0 and 2.0 m are not clearly distinguishable, and technically should be presented differently (e.g., by using the ">"). For some of the other Figs. presented I also would like to recommend to clearly state the abbreviations (e.g., td, sd, cd,: : :) in the Figure caption.*

We will add extra zoomed in maps of an interesting river section. We will also change the legend colors and improve the notation in the legend. In case of abbreviations in the figures, we will add the meaning to the figure caption. In cases of figures with many abbreviations, we will reference to the table that has the meanings listed.

6. *Finally, I would recommend to extent the discussion on Fig. 5 – as already indicated there may be variables other than the water height responsible for the loss height available…*

The discussion of figure 5 will be extended and moved to the discussion section (see also our replies to the comments of the first referee).

New References (not already included in the first manuscript):

Bilmes, J., 2002. Graphical models and automatic speech recognition. The Journal of the Acoustical Society of America 112, 2278 (2002); doi: http://dx.doi.org/10.1121/1.4779134

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

Fuchs, S. , Heiss, K. and Hübl, J., 2007. Towards an empirical vulnerability function for use in debris flow risk assessment. Nat. Hazards Earth Syst. Sci., 7, 495–506, 2007

Gerl, T., Kreibich., H, Franco, G., Marechal, D. and Schröter, K., 2016. A Review of Flood Loss Models as Basis for Harmonization and Benchmarking. PLoS ONE 11(7): e0159791. https://doi.org/10.1371/journal.pone.0159791

Islam, K. M. N., 1997. The impacts of flooding and methods of assessment in urban areas of Bangladesh. PhD Thesis. Middlesex University

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

Mourad, R., Sinoquet, C., Leray, P., 2011. Probabilistic graphical models for genetic association studies. Brief Bioinform (2012) 13 (1): 20-33. DOI:https://doi.org/10.1093/bib/bbr015

Papathoma-Köhle, M., Zischg, A., Fuchs, S., Glade, T., Keiler, M., 2014. Loss estimation for landslides in mountain areas – An integrated toolbox for vulnerability assessment and damage documentation. Environmental Modelling & Software 63 (2015) 156-169.

Papathoma-Köhle, M., 2016. Vulnerability curves vs. vulnerability indicators: application of an indicator-based methodology for debris-flow hazards. Nat. Hazards Earth Syst. Sci., 16, 1771–1790, 2016

Penning-Rowsell, E.C., C. Johnson, en S. Tunstall. The benefits of Flood and Coastal Risk Management: A Manual of Assessment Techniques. Middlesex University Press, London, 2005.

Scawthorn, C., Flores, P., Blais, N., Seligson, H., Tate, E., Chang, S., Mifflin, E., Thomas, W., Murphy, J., Jones, C., and Lawrence, M.: HAZUS-MH flood loss estimation methodology II. Damage and loss assessment, Natural Hazards Review, 7, 72–81, 2006.

Sluijs, L., M. Snuverink, K. van den Berg, en A. Wiertz., 2000. Schadecurves industrie ten gevolge van overstromingen. Tebodin. Rijkswaterstaat DWW.

Sudderth, E., Freeman, W., 2008. Signal and Image Processing with Belief Propagation. IEEE Signal Processing Magazine Volume: 25, Issue: 2, March 2008. DOI: 10.1109/MSP.2007.914235

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

Thieken, A.H., Olschewski, A., Kreibich, H., Kobsch, S. and Merz, B., 2008. „Development and evaluation of FLEMOps - A new flood loss esimation model for the private sector." WIT Transactions on Ecology and the Environment 118 (2008): 315 -324.

Van Ootegem, L., Verhofstadt, E., Van Herck, K., Creten, T., 2015. Multivariate pluvial flood damage models. Environ. Impact Assess. Rev. 2015, 54, 91–100.