Schröter, K., Lüdtke, S., Redweik, R., Meier, J., Bochow, M., Ross, L., Nagel, C., Kreibich, H. (2018): Flood loss estimation using 3D city models and remote sensing data. - *Environmental Modelling and Software*, *105*, pp. 118—131.

DOI: http://doi.org/10.1016/j.envsoft.2018.03.032

Flood loss estimation using 3D city models and remote sensing data

Kai Schröter[a], Stefan Lüdtke[a], Richard Redweik[b], Jessica Meier[a], Mathias Bochow[c], Lutz Ross[b], Claus Nagel[b], Heidi Kreibich[a]

a) German Research Centre for Geosciences GFZ, Section 5.4 Hydrology, 14473 Potsdam, Germany

b) VirtualcitySYSTEMS GmbH, 10789 Berlin, Germany

c) German Research Centre for Geosciences GFZ, Section 1.4 Remote Sensing, 14473 Potsdam, Germany

Highlights

Flood loss models for residential buildings are developed based on 3D city models and remote sensing data.

These multi-variable predictive models are validated using empirical data.

3D city models are readily available for urban areas and as standardized data they ease the spatial transfer of loss models.

Building vulnerability information is embedded into virtual 3D city models to support flood risk sensitive urban planning.

Abstract

Flood loss modeling provides the basis to optimize investments for flood risk management. However, detailed object-related data are not readily available to generate spatially explicit risk information. Virtual 3D city models and numerical spatial measures derived from remote sensing data provide standardized data and hold promise to fill this gap. The suitability of these data sources to characterize the vulnerability of residential buildings to flooding is investigated using the city of Dresden as a case study, where also empirical data on relative flood loss and inundation depths are available. Random forests are used for predictive analysis of these heterogeneous data sets. Results show that variables depicting building geometric properties are suitable to explain flood vulnerability. Model validation confirms that predictive accuracy and reliability are comparable to alternative models based on detailed empirical data. Furthermore, virtual 3D city models allow embedding vulnerability information into flood risk sensitive urban planning.

Keywords

flood risk; flood loss modeling; standardized data; random forests; vulnerability; virtual 3D city models

1 Introduction

With the transition to flood risk management, for which flood risk analyses are an essential basis, loss estimation is becoming increasingly important (Merz et al. 2010a; Bubeck et al. 2017). This is also recognised by the European Flood Directive (2007/60/EC) which requires the assessment and mapping of flood risk, and to draft flood risk management plans. Risk assessments aim to quantify the probability of expected losses resulting from interactions between hazard, exposure and vulnerability (e.g. de Moel et al. 2015).

Risk analyses are undertaken on different spatial scales (Meyer and Messner, 2005; de Moel et al. 2015). Our study focuses on the local or micro-scale, i.e. on the level of individual objects. These assessments are primarily undertaken to optimise investments for risk management concepts, including protection measures, urban planning, etc. Additionally, micro-scale, spatially explicit risk information enables communities, companies, and people to prepare for disasters (e.g. Takeuchi 2001; Merz and Thieken 2004). At the micro-scale the assessment is based on single elements at risk. For instance, in order to estimate the loss to a community in case of a certain flood scenario, losses are calculated for each affected object, e.g. buildings. These analyses require detailed object-related data. Accurate flood modelling at high spatial and temporal resolutions remains a significant challenge (Teng 2017). Likewise, this concerns the spatial resolution and geo-location of the exposed objects as well as object characteristics, which determine their vulnerability towards inundation. This is challenging, since such detailed data are hardly available (Apel et al., 2009).

A standard approach to determine the expected direct monetary loss to buildings are depth-damage functions based on the type or use of the building and the inundation depth (Grigg and Helweg, 1975; Smith, 1994; Penning-Rowsell et al.,2005). Accordingly, inundation depth is the variable which is most commonly included in flood loss models (Merz et al. 2010b; Gerl et al. 2016). However, making use of additional variables to explain vulnerability and to develop predictive models for loss estimation has been shown to offer substantial advancements to explain flood loss (Thieken et al., 2008; Schröter et al., 2014). The domain of flood loss modelling is experiencing a boom of tree based data analysis since Merz et al. (2013) have demonstrated the suitability and superior performance of regression trees and bagging decision trees for flood loss estimation. Also graphical models (Bayesian Networks) have been successfully applied to the domain of flood loss estimation, e.g. Vogel et al. (2012), Schröter et al. (2014). Decision trees have been used by Spekkers et al. (2014) to gain new insights into damage influencing factors for pluvial floods. Chinh et al. (2015), Hasanzadeh Nafari et al. (2016), and Wagenaar et al. (2017) have derived multi-variable flood loss models based on decision-tree approaches for the Mekong delta (Vietnam), Australia, and the Maas River (The Netherlands), respectively. The model validation experiment of Schröter et al. (2014) has shown that regression trees do also outperform traditional models in cross-regional and temporal transfer applications. Kreibich et al. (2017a) have proposed a novel flood loss model for meso-scale applications based on bagging decision trees which provides uncertainty information of loss estimates. In the broader context of flood hazard and flood risk assessment tree based model approaches have also been successfully applied (e.g. Wang et al. 2015, Chapi et al. 2017).

The application of multi-variable models poses high requirements on the availability of input data (Merz et al., 2013), which is particularly challenging for the spatial transfer of these models. Kreibich et al. (2017a) used empirical flood vulnerability data gathered via surveys to estimate the model input variables on municipality level. However, such data is costly and time consuming to collect and is therefore unavailable in many regions (Thieken et al. 2017; Kreibich et al., 2017b). By necessity, flood loss models are often taken from the literature and transferred in space and applied to different built environments and other settings (e.g. Balica et al. 2013). Model transfer and application is eased when data from standard databases or sources can be used because then the specification and definition of variables is similar across regions. For instance, Kreibich et al. (2010) used macro-economic data from the Federal Statistical Office Germany and the Federal Employment Agency together with geo-marketing data for the application of the multi-variable model FLEMOcs-Flood loss estimation model for the commercial sector on the meso-scale. As such, the Germany wide application of FLEMOcs was enabled. However, due to the spatially coarse data base, uncertainties in loss estimates are high. Gerl et al. (2014) examined the use of urban structure type information which is automatically derived from remote sensing data for flood loss estimation. Their analyses show that different urban structure types comprising the categories “closed block development”, “semi-open block development”, “mid-rise dwellings”, and “single-family/semi-detached houses” and the information about their specific location are valuable for flood loss modelling. However, they suggest that additional data about building characteristics which cannot be derived from remote sensing would be useful to make further advancements.

Detailed building location and characteristics can be stored in (virtual) 3D city models, which are based on CityGML. CityGML is an open standard application schema of the Geography Markup Language (GML), which represents the 3D geometry, 3D topology, semantics, and appearance of objects on different levels of detail (LOD). Hence, 3D city models provide a framework for detailed spatial information in terms of geo-located building footprints (based on cadastral data) which is useful to describe the exposure as well as detailed information about buildings characteristics and geometries which are useful to describe the vulnerability of residential buildings at risk. CityGML is the reference model for the building model in the Infrastructure for spatial information in Europe (INSPIRE; Directive 2007/2/EC). They have been widely adopted for diverse applications in environmental simulations (Biljecki et al. 2015) including also examples from flooding and flood damage assessment (e.g. Amirebrahimi et al. 2016). However, currently building information contained in 3D city models is available for low levels of detail (LOD), i.e. LOD1 or LOD2, whereas relevant data about building openings and/or interior structures of a building are available only for high levels of detail, i.e. LOD3 or LOD4. The concept of LODs is described in detail in section 2.2.2. Therefore, a combination with remote-sensing data might be advantageous. For instance, numerical spatial measures from remote-sensing data have been used to characterize the physical properties of landscapes (Uuemaa et al. 2009) and urban areas (e.g., Bochow 2010, Graesser et al. 2012) and have been used as proxies for information like “socio-economic characteristics” or “energy demand” (Jensen and Cowen 1999; Taubenböck et al. 2009).

The objective of this study is to develop multi-variable flood loss models which are based on standardized data sources to characterize the vulnerability of buildings towards flooding. We investigate the potential of virtual 3D city models and numerical spatial measures derived from remote sensing data to support the estimation of flood losses to residential buildings. Section 2 and 3 introduce the data sources and the methods applied. The results of these analyses are presented in section 4. Further, the potential of 3D city database systems to store data, embed flood loss modelling as a functional extension for risk assessment, and visualize results is explored. In this regard the prototype implementation and a case study application in the city of Dresden (Germany) are described in section 5. In section 6 the results are discussed and concluded in section 7.

2. Data

2.1 Study area

The city of Dresden (Saxony, Germany) is used as a case study for this research. Figure 1 shows the location of Dresden along the Elbe river banks and its tributaries. In the past, floods have caused severe impacts as for instance in June 2013, April 2006 and in August 2002. The flood in August 2002 caused more than EUR 1 Bn economic damage in the city of Dresden with losses to residential buildings of EUR 305 Mn (Kreibich and Thieken, 2009)

Dresden is characterised by a heterogeneous architectural structure including historical as well as modern multi-storey buildings in the densely built-up city center, and multi-storey residential buildings as well as one-to-three-storey developments in the neighboring city districts (Gerl. et al 2014) which is the outcome of a series of eventful historic developments with drastic impacts on the building stock development such as World War II, communist planned economy and the reunion of Germany. Given this diversity of building characteristics, differences in terms of building vulnerability towards flooding are expected.

Driven by the recent floods, flood risk assessment and management is a highly relevant topic on the urban planning agenda in the city of Dresden and comprehensive flood management concepts have been put into practice (Landeshauptstadt Dresden, 2011). As part of this planning hydro-numeric simulations have been conducted to determine inundation depth maps for historic floods and for design flood scenarios. For this study, inundation depth maps for three different water levels at the gauge Dresden are available for several focus areas in Dresden. In Landeshauptstadt Dresden (2011) flood impacts have been estimated using simple flood damage curves relating inundation depth to specific loss [€/m²] for different land use classes. This damage model had been derived for the river Rhine (ICPR 2001) and from this model the damage curve for residential buildings will be also compared to the outcomes of the flood loss models developed in this study.

Other data sources (cf. Fig. 1) are a dataset of computer aided telephone interviews (CATI) carried out after the floods August 2002 and April 2006 in Dresden (Thieken et al. 2007), building data from the 3D city model of Dresden in LOD1 and LOD2 (citydb) saved in the 3D City Database, spatial measures (SM) for the residential building stock derived from IKONOS hyper-spectral images mapped on the German Official Topographic Cartographic Information System (ATKIS) building blocks (Bochow 2010). These data sets provide input variables for the derivation of predictive flood loss models and are described in detail in the following sections. Beyond that, estimates of the regional stock of residential buildings (Kleist et al. 2006) are used to quantify the asset values of residential buildings exposed to flooding in the application case in the city of Dresden in section 3.4. A digital elevation model (DEM10; Federal Agency for Cartography and Geodesy in Germany (BKG)) with a spatial resolution of 10 m is used to estimate the elevation of residential buildings and to derive inundation depths maps for the flood scenarios used in the application case.

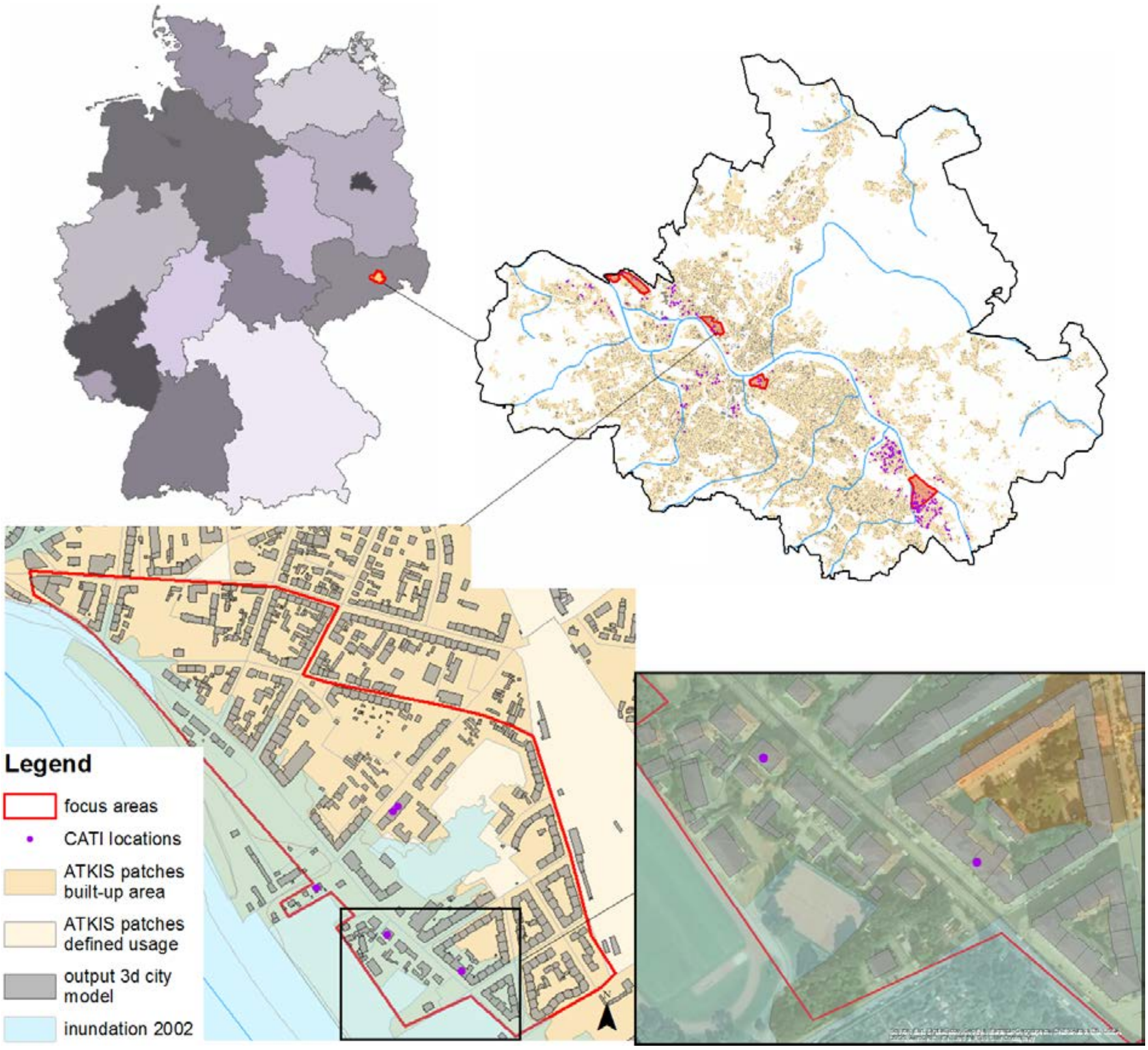

Figure 1 Map of Dresden in Germany with focus areas, interview locations, building outlines and ATKIS land use classes, the detail cutout shows the focus area Pieschen

2.2 Data sources

2.2.1 Computer aided telephone interview data

Empirical loss data have been collected from flood affected households via computer aided telephone interviews (CATI) after the major floods in August 2002 and April 2006. These questionnaires cover aspects about the flooding situation, early warning and emergency measures, precaution, building characteristics and the socio-economic status (Thieken et al. 2007). From these campaigns 292 geo-located interviews are available in the city of Dresden. Further, the interview data were supplemented by estimates of the return period of the flood peak discharge as well as aggregated indicators for flow velocity, contamination, flood warning, emergency measures, precautionary measures, flood experience, and socioeconomic variables. Building values were estimated according to the VdS guideline 772 1988-10 (Dietz 1999), and the loss ratio of the buildings has been calculated as the quotient of replacement costs and building value. However, the building value could be determined for only 80 buildings due a lack of information in the other interviews. As we are interested in investigating data sources to describe residential building resistance in flood loss models, from the broad range of information about flood influencing factors, those variables of the

CATI dataset are of central interest which provide specific information about the building characteristics. Beyond that, the interview datasets provide information about the flooding situation in terms of water depth and the flood impact in terms of relative loss to the individual buildings. These building related information used from the CATI data set are listed in Table 1. For a detailed description of the complete set of variables derived from CATI refer to (Merz et al. 2013).

Table 1: Variables available from computer aided telephone interviews, 3d city model and remote sensing

abbr.	Variable	Scale and range
flooding situation and impact from cati data set		
wd	Water depth	c: 212 cm below ground to 476 cm above ground
rloss	loss ratio of residential building	c: 0 = no damage to 1 = total damage
building characteristics from cati data set		
bt	Building type	n (1=multifamily house, 2= semi-detached house, 3=one-family house)
nfb	Number of flats in building	c: 1 to 12 flats
fsb	Floor space of building	c: 60 to 2,000 m²
bq	Building quality	o: 1=very good to 6=very bad
bv	Building value	c: 130,088 to 3,718,677 €
year	Construction year	c: 1800 to 2001
heat	Heating system	N (gas, coal, heating oil, night storage, wood pellet, long distance heating)
building characteristics from citydb data set		
ba	Building area	c: 37.6 to 362.5 m²
rt	Roof type	n: 3100, 3200, 3300, 3400, 3500, 5000
mh	Measured height	c: 4.42 to 19.01 m
sag	Storeys above ground	c: 1 to 5
year	Year of construction	c: 2000 to 2012
Variables of spatial measures data set		
BT_AREA	Area of a city block calculated as the sum of the pixel areas of the rasterized city block	c: 2302 - 43726
BT_LSI	Linear-Segment-Indicator calculated as the ratio between the first and second principal component of the pixel positions of a rasterized city block in the 2D space	c: 1.04 – 6.75

BT_DIST	Distance of a city block to the city center	c: 467 - 738
CL_POFA	Percentage of area of a Land use/Land Cover (LULC) class (or LULC class group) within a UST patch	c: 0.0726, … 0.339
CL_GRAV	Degree of compactness of a LULC class (or LULC class group) according to Newton's Universal Law of Gravitation	c: 0.00186, … 0.0286
CL_CEPE CL_CEPE_HEIGHT	Mean distance of class pixels from a central region (3% innermost pixels) of a building block normalized by the mean distance of the block boundary from the central region. A second variant of this basic type ignores pixels below a minimum height value in the DSM to exclude small buildings like garages.	c: 0.01, … 1 c 0.01, … 1
CL_NUMSEG	Number of segments of a LULC class (or LULC class group)	d: 2 … 37
CL_NUMSEG_PER_AREA	Number of segments of a LULC class (or LULC class group) divided by the area of the building block	c: 0.0322, … 0.319
CL_SEGORI	Mean value of the angles between the first Eigenvectors (indicating the orientation of the segments in the 2D space) of each pair of two segments of a LULC class (or LULC class group)	c: 0.318, … 40.2
CL_HEIGHT	Height of a LULC class (or LULC class group) *	c: 33.8, … 141
SEG2CL_AREA	Area (m^2) of the segments of a LULC class (or LULC class group) *	c: 68.1, … 830
SEG2CL_LSI	Linear-Segment-Indicator calculated as the ratio between the first and second principal component of a segment's pixel positions in the 2D space *	c: 1.27, … 4.79
SEG2CL_SEGDIST	The shortest distance between the centroid of a segment of a LULC class (or LULC class group) and the centroid of the closest segment of the same class (group) *	c: 8.45, …62.6
SEG2CL_VOL	3D volume of the segments of a LULC class (or LULC class group) *	c: 409, … 9764
SEG2CL_HEIGHT	Height of the segments of a LULC class (or LULC class group) **	c: 38.8, … 115

ust_type	The urban structure type of the building block	n: 2,3,4,5,7,8,B,F,I

c: continuous, d: discrete, o: ordinal, n: nominal, *: 4 variants: min, max, mean, sdev, **: 20 variants: min, max, mean, sdev, diff over all pixels per building followed by min, max, mean, sdev over all resulting values of the buildings within a building block

2.2.2 3D-city model data

For this study the City of Dresden provided a virtual 3D city model which is based on the international standard CityGML. CityGML is the Open Geospatial Consortium (OGC) standard for the representation, storage, and exchange of virtual 3D city models. CityGML is a XML based application schema of the Geography Markup Language 3 (GML3) (version 3.1.1). Hence, its usage is suitable for Spatial Data Infrastructures (SDI) in the context of OGC web services. Unlike purely graphical or geometrical models, CityGML represents semantic and topological aspects of features which are relevant for city models as well (Gröger and Plümer, 2012). Hence, CityGML based 3D city models can be used for visualization purposes and for thematic analysis. For that, CityGML provides a semantic "geospatial information model (ontology) for urban landscapes based on the ISO 191xx family" (Gröger et al., 2012). In general, CityGML defines classes and relations with respect to geometrical, topological, semantical, and appearance properties (Gröger et al., 2012). CityGML makes use of a level of detail concept discerning five levels (LOD0 to LOD4), where the same object can be represented in different LODs at the same time. For instance, in LOD0 a building is only represented by the horizontal 3D surface representing the footprint or the roof edge. In LOD1 buildings are represented as simple block models specified either by multi surfaces or a solid body. LOD2 introduces *BoundarySurfaces* as thematic features, e.g. *WallSurface*, *RoofSurfaces*, *GroundSurfaces,* and other surface information. LOD3 allows to model openings of *BoundarySurfaces* such as *doors* and *windows*. LOD4 adds interior structures of buildings as for instance *Rooms* or *BuildingFurniture*. The LODs describe geometrical and semantical aspects. That means the objects become more accurate and detailed with an increasing LOD level. An additional important design principle of CityGML is the coherent semantical-geometrical modeling, which means real-world entities are represented as features at semantic level and as geometry objects (that are assigned to the semantic features) at geometry level. It follows that "geometrical objects 'know' what they are" and "semantic entities 'know' where they are and what are their spatial extents" (Stadler and Kolbe 2007). For example, 3D city models can be queried for all wall surfaces with their material in an inundated area.

CityGML organizes the features into different thematic modules which cover different thematic fields of virtual cities (e.g. Building, Bridge, Vegetation, Water Body, etc.) with each of them containing specific feature types. Still, the data model is flexible to combine any of these modules as needed by the application or application domain (Gröger and Plümer, 2012). In this research, the focus is on the data contained in the building module

For the Dresden case study a 3D city model in LOD1 and LOD2 was available and thus providing geometrical information about the location and the extent of the buildings (*building area*). Additional available thematic attributes on the building level are: *function*, e.g. residential, industry, etc., *roof_type*, *measured_height*, *storeys_above_ground,* and *year of construction*, see Table 1. The attributes function and *roof_type* are specified separately from the CityGML schema using code lists which specify the values given admissible ranges. To give an example, a *roof_type* encoded by the value 3100 is a gabled roof, a *roof_type* encoded by the value 3200 is a hipped roof. Roof types are useful to distinguish building types which is a relevant detail to characterize building vulnerability.

The usefulness of these variables for flood loss estimation will be explored within the model development step.

2.2.3 Numerical spatial measures derived from remote sensing data

Numerical spatial measures as implemented by Bochow et al. (2010) are based on standardized calculation procedures. They describe the amount, proportions, size, shape, spatial arrangement and distribution of land cover classes and classified image objects (e.g. houses, trees, lawns) within the building blocks. For some of these measures (e.g. the footprint size of a building) it is obvious or has been already shown that they directly affect the vulnerability to floods. For others hidden relations might exists which we aim to explore in this work.

Multispectral IKONOS Geo Ortho Kit satellite images (GeoEye Inc., 2006) acquired between 2004 and 2008 have been geometrically and atmospherically corrected, pan-sharpened to a spatial resolution of 1 m using principal component pan-sharpening, and mosaicked to a single image covering the complete city area of Dresden. A land use/land cover (LULC) classification has been produced using a maximum likelihood classifier. The initial classification result has been enhanced by incorporating object height information derived from LiDAR data using a decision tree built with the ERDAS Imagine Knowledge Engineer. Misclassified pixels (e.g., pixels of roofs and ground surface pixel with similar spectral properties) are reclassified according to their object height into a roof class or a ground surface class, respectively. In a further processing step the LULC classes are combined into thematic groups, as for instance the class group “vegetation” integrates the classes “meadow” and “trees”, which are stored as additional layers to the original LULC classes. The resulting layer stack contains the original 15 LULC classes, as well as 12 thematic class groups, namely vegetation, trees, soils, roofs, roofs_metal, roofs_tiles, roofs_flat, traffic, sports, roofs_industry, water, and shadow. In the next processing step sequentially numbered image segments consisting of adjacent pixels of the same class are created and stored for each LULC class and class group.

The resulting layers enable the computation of class specific and image object based numerical spatial measures as implemented by Bochow et al. (2010) which can be used as the basis for urban structure type mapping. The numerous spatial measures are organized into five categories assessing the size and shape, orientation, and distribution, percentage of area, neighborhood, or spatial position of classified pixels or image segments. They are calculated either based on the entire building blocks (taken from the ATKIS-DLM), based on individual classes or class groups within the building blocks, or based on the segments of a class within the city blocks. Furthermore, if present, the interior (backyard) of a building block or its border area can be considered separately for the calculation. For all basic features distinctive statistical parameters like minimum, maximum, mean, and standard deviation can be calculated. In this work we focus on the mean values for the basic spatial measures which are related to buildings, i.e. the class group “roofs”, see Table 1.

The Building blocks from the ATKIS-DLM data also serve as spatial mapping units for urban structure types. The urban structure types are distinguished in terms of their composition of different objects like buildings, trees, and surface material as for instance roof materials or vegetation types as well as about their distribution and arrangement within space. Within this study the following urban structure types are considered: semi open block development (2), terraced houses, (3), single/semi detaches houses (4), mixed types of buildings (5), sport facilities/playgrounds (7), garden plots (8), green spaces (B), water surfaces (F), traffic areas (I).

3. Methods

3.1 Model development

For the development of multi-variable flood loss models based on 3D city models and numerical spatial measures derived from remote sensing images two central tasks are addressed: a) selecting variables from the variety of candidate variables of standardized data sources according to their usefulness to explain flood loss, and b) deriving a good predictive model for flood loss to residential buildings.

Model development is carried out using a variety of sub-sets with different number of variables from the different data-sets described above, see Table 1. In this way, models of different complexity are derived and compared in terms of predictive performance also from the angle of model parimony. This comprises the development of flood loss models using i) the complete set of variables available from the surveys using computer aided telephone interviews (CATI_complete), ii) a reduced set of variables from computer aided telephone interviews which are related to building characteristics (CATI_reduced), iii) water depth at the building as the only explanatory variable referring to established stage damage functions (sdf), iv) geometrical information and thematic attributes provided by the 3D city model (citydb), v) building roof related numerical spatial measures derived from remote sensing images representing the mean values per building block (sm), and vi) the concatenation of 3D city model data and spatial measures (citydb_sm). All datasets are amended with the observations of water depth and relative building loss available from the empirical surveys (CATI) for the selection of variables and the derivation of predictive models.

The underlying question of variable selection is: which variables are most useful to explain flood damage? This step in data analytics aims to improve the predictive performance of a model in terms of faster and more cost-effective predictors in the sense that the trade-off between the number of explanatory variables used in the predictive model and the performance of the prediction is reflected according to the principle of parsimony. Accordingly, the task is to choose those variables from the data that will give as good or better predictive performance while requiring less data (Guyon and Elisseeff 2003). The general strategy involves a ranking of the candidate explanatory variables and an assessment of model predictive performance. While the ‘ad-hoc’ selection of variables requires deep knowledge of the problem domain, data-mining algorithms entail a variety of methods for the automated analysis of large data sets and also provide metrics for comparing the relevance of each candidate variable. Guyon and Elisseeff (2003) distinguish filter methods which rank the variables independent on the model approach like correlation coefficients, wrapper methods which assess subsets of variables according to the predictive performance within the chosen model approach like recursive feature elimination algorithms, and embedded methods which incorporate variable ranking and selection as part of the model derivation process as for instance classification and regression trees (Breiman et al. 1984). Embedded methods include the interaction with the model learning and are computationally more efficient than wrapper methods (Saeys et al. 2007).

Previous research (e.g. Merz et al. 2013, Schröter et al. 2014, Spekkers et al. 2014, Kreibich et al. 2017) has emphasized a number of advantages of tree based algorithms for problems faced in flood loss modeling. First, the approach is non-parametric, thus assumptions concerning the covariance between explanatory variables and response variable are not needed. Second, non-linear and non-monotonic dependencies can be represented by a tree. Third, these algorithms can deal with heterogeneous data, i.e. a mixture of continuous and categorical variables, as well as incomplete data. Fourth, no assumptions about the independce of data are needed. However, tree based models require large data sets to detect and represent complex inter-relationships. Random Forests (RF, Breiman 2001) are one member of Tree-Based algorithms which make use of bagging predictors (Breiman 1996) and have become popular across a broad range of disciplines due to its versatile applicability and efficient solution to multi-variable prediction problems facing complex interactions among variables with different scales (Huang et al., 2016).

For the problem addressed in this study the method of RF is a suitable approach because it first provides a concept to estimate the importance of candidate explanatory variables and thus enables a well-founded selection of variables and second is an efficient algorithm to learn models with superior predictive performance even for data with more variables (p) than samples (n) (Genuer et al. 2010; Huang et al., 2016). In this study we have applied the R package implementation randomForest by Liaw and Wiener (2002).

RF belongs to the family of ensemble methods: it uses many regression trees and strives to reduce the uncertainty associated with the selection of a single model, by aggregating an ensemble of alternative models. RFs are derived by generating many bootstrap replicas of the data set and by growing a regression tree on each replica. At each node of a regression tree a subset of explanatory variables (mtry) is randomly chosen and the best split is determined within this subset. This distinguishes RF from the bagging tree method where all variables are considered at each split. Sub-setting the variables considered for each split enables a more diverse set of variables to contribute to the ensemble prediction. This also brings the advantage that the individual trees of the forest are less correlated than in the bagging approach, and thus are less biased. A bootstrap replica is generated by randomly drawing with replacement observations of the sample n. On average, 37% of observations are excluded for building an individual tree. These observations are called out-of-bag observations (OOB) and are used to evaluate the predictive performance of the tree in terms of the out-of-bag error. While OOB error evaluation supersedes the need for cross-validation, bootstrapping and a minimum number of five data points in a node makes RF robust against changes in data and avoids overfitting. The response of RF is an aggregation of the responses of all individual regression trees in the ensemble while the distribution of responses provides an estimate of model structure and input data related uncertainty.

Further, RFs include an efficient way for ranking candidate explanatory variables based on variable importance, and thus supports the selection of the most suitable variables for the development of predictive models. In the permutation based approach the RF algorithm estimates the importance of a variable by evaluating the increase of the prediction mean squared error when data for that variable is permuted in the OOB sample: the higher the increase, the more important the variable. This approach is popular due to its unconditional properties and as the algorithm is sensitive to informative variables and to relations among variables it supports the identification of relevant variables (Hapfelmeier and Ulm 2013).

The underlying question for deriving a good predictive model is: how well does the model predict the target variable? To evaluate and compare the performance of the predictive models MSE is computed as the average squared deviation of the ensemble mean prediction and the out-of-bag observations as given in equation 1

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\overline{P_i} - O_i)^2 \quad (1)$$

With Pi the ensemble mean prediction of observation i, O the observation i and n the number of observations.

Additionally, the mean bias error (MBE, equation (2)) and the mean absolute error (MAE, equation (3)) which give information about the accuracy and the precision of the model predictions are used as model performance criteria.

$$MBE = \frac{1}{n}\sum_{i=1}^{n}(P_{50_i} - O_i) \qquad (2)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|P_{50_i} - O_i| \qquad (3)$$

MBE and MAE are calculated using the median (P_{50}) of the predictive distributions of the Random Forest models. The distribution of responses available from the Random Forest model offers additional information about the prediction uncertainty. We analyze this property using the 95-5-quantile range (QR_{90}, equation (4)) and the hit rate (*HR*, equation (5)) as indicators for sharpness and reliability of model performance, respectively (Gneiting and Raftery, 2007).

$$QR_{90} = \frac{1}{n}\sum_{i=1}^{n}(P_{95_i} - P_{5_i})/P_{50_i} \qquad (4)$$

$$HR = \frac{1}{n}\sum_{i=1}^{n} h_i \ ; \ h_i = \begin{cases} 1, if \ O_i \in \left[P_{95_i}, P_{5_i}\right] \\ 0, otherwise \end{cases} \qquad (5)$$

QR_{90} is a measure of sharpness of the prediction interval with smaller values representing smaller ranges of the model prediction interval. HR is an indicator for model reliability by quantifying the ratio of observations within the prediction interval defined by the 95 and 5 quantile range The 95–5 quantile range corresponds to a nominal coverage of 0.9, and thus a HR=0.9 indicates that the coverage of model predictions is equal to the nominal coverage representing a perfect reliability of model prediction on this level.

3.2 Software and data availability

The model development of the 3dcfd module was implemented in R 3.4.1 (2017-06-30), using the attached packages scales 0.4.1, (Wickham, 2016), doParallel 1.0.10 (Analytics, R. and Weston, 2015a), foreach 1.4.3 (Analytics and Weston, 2015b), hydroGOF 0.3.8 (Zambrano-Bigiarini, M. 2014), tidyr 0.6.1 (Wickham and Henry, 2017), stringr 1.2.0 (Wickham, 2017), dplyr 0.5.0 (Wickham et al., 2017a), reshape2 1.4.2 (Wickham, 2007), readr 1.1.1 (Wickham et al. 20017b), , randomForest (Liaw and Wiener, 2002), ggplot2 4.6.12 (Wickham, 2007; Schloerke et al. 2017), _RODBC 1.3.15 (Ripley and Lapsley, 2017), RODBCext 0.3.0 (Zoltak et al. 2017) A prototype implementation for the case study Dresden is available online (Schröter et al. 2017 [doi:10.5880/GFZ.5.4.2017.001]). Further information can be obtained from the corresponding author of this paper.

4. Results

4.1 Variable selection

The RF implementation randomForest uses two parameters: *mtry*, the number of explanatory variables which are randomly chosen at each node, and *ntree* which defines the number of trees in the forest (Liaw and Wiener 2002). As the variable importance derived by RF is sensitive to these parameters (Genuer et al. 2010), we pursue an averaging approach to derive stable results for variable importance as a basis to select the variables from the different data sub-sets for the predictive models. The averaging is carried out for the variable importance obtained from a set of random forests with different numbers of trees (*ntree* ∈ [500, 1000, 2000, 3000]) whereby each realisation of *ntree* has in turn been derived using three values for *mtry*: the default p/3 and the recommended lower and upper limits p/6 and 2p/3 where *p* is the number of predictors in the dataset (Breiman 2001). The motivation for this variation and averaging of variable importance is to identify the variables which are consistently important across a reasonable range of settings for the RF algorithm. The results for the normalized mean importance of the variables related to building characteristics from the different data sub-sets are summarized in Figure 2.

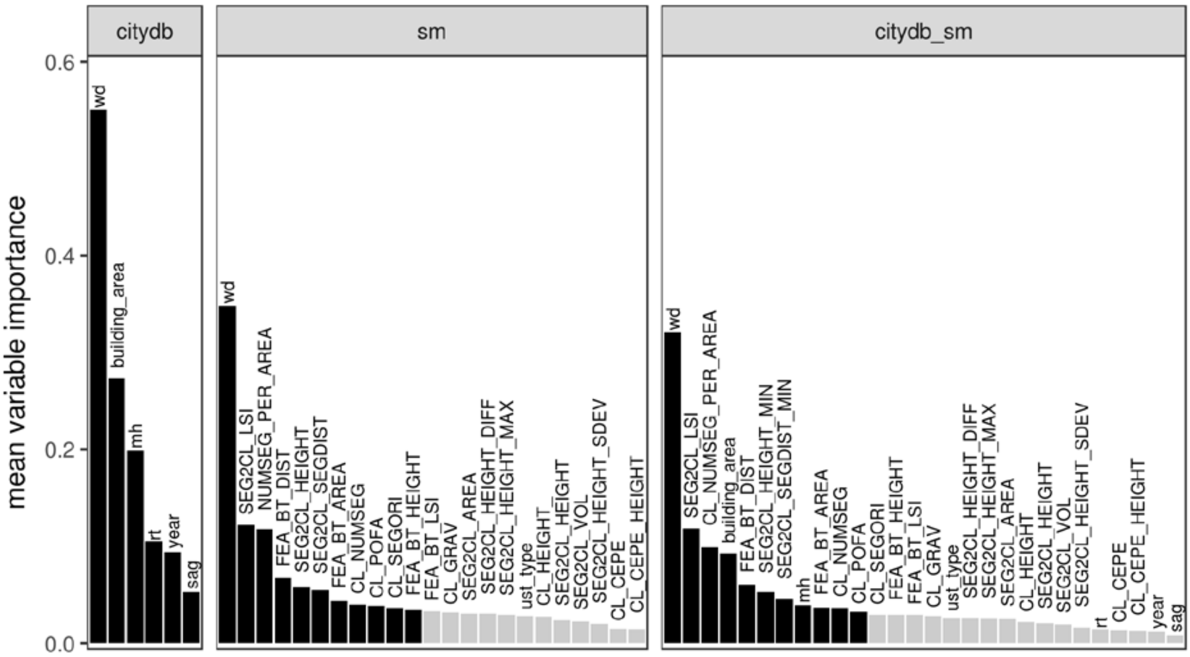

Figure 2: Ordered importance for building related variables for 3D City database and spatial measures data sub sets based on n= 80 observations including water depth from the cati data base. Top 10 important variables used for predictive modeling are colored black.

Water depth (*wd*) is the most important variable to explain flood loss in comparison to the other variables available from the different data sub-sets. The analysis of the citydb data sub-set results in a large importance of *building area* and *measured height (mh)*, followed by *roof type (rt)*. These variables are related to building characteristics as for instance the floor space which is a function of building footprint area and height or typical roof types for different building types. Merz et al. (2013) have shown that both building type and floor space are strongly correlated and that their relationship to relative building loss can be explained by the finding that single-family houses (with smaller floor spaces) have a higher loss ratio in comparison to multi-family houses (with larger floor spaces). From the spatial measures data sub-set the linear segment indicator (LSI) and the numbers of segments per area (NUMSEG_PER_AREA) achieve the highest importance for the estimation of relative flood loss. LSI is an indicator for the elongated shape of an area. Hence, the LSI for the roof areas of the building block provides information about the geometry of the building footprint which in turn provides proxy information for different building types, and/or building values.

NUMSEG_PER_AREA describes the density of buildings in the area with higher numbers being an indicator for the presence of single family houses and lower number an indicator for multi-family houses. Again, this information may serve as a proxy for building types and related properties. This complements the findings of Thieken et al. (2005) and Gerl et al. (2014) that the differentiation of building types is useful to explain flood loss. The analysis of the combined data sets citydb_sm confirms the importance of LSI, NUMSEG_PER_AREA, and building area information.

4.2 Model derivation and evaluation

The analysis of variable importance indicates that the data sources citydb and sm contain variables which are informative for the estimation of flood loss. On that basis, the performance of predictive models for flood loss to residential buildings using these variables is examined. The derivation predictive flood loss models uses the top 10 important variables from the outcomes documented in Figure 2.

First the two parameters of the RF algorithm need to be optimized with the aim to find stable models. To this end we repeatedly (i = 100 repetitions) build random forests with different *ntree* values (*ntree* ∈ [100, 500, 1000, …., 15.000]) and varying values for *mtry* (*mtry* = ∈ [p/6, … 2p/3]), record the mean OOB error of i repetitions and see the number of trees where the OOB error asymptotically reaches a stable minimum. Figure 3 shows the 90-quantile range and the median of OOB distributions grouped for *mtry* parameter values. For smaller forests (*ntree* < 500) the variability of OOB is large but decreases with growing numbers of trees in the RFs. RF models with 6000 trees seem to achieve stable predictions. For RF models above this size, the parameter *mtry* dominates the OOB error with optimum performance for *mtry* = 2 to 3 which is close to the recommended default value p/3 (for the data sub-set with p = 11 variables) (Liaw and Weimer 2002).

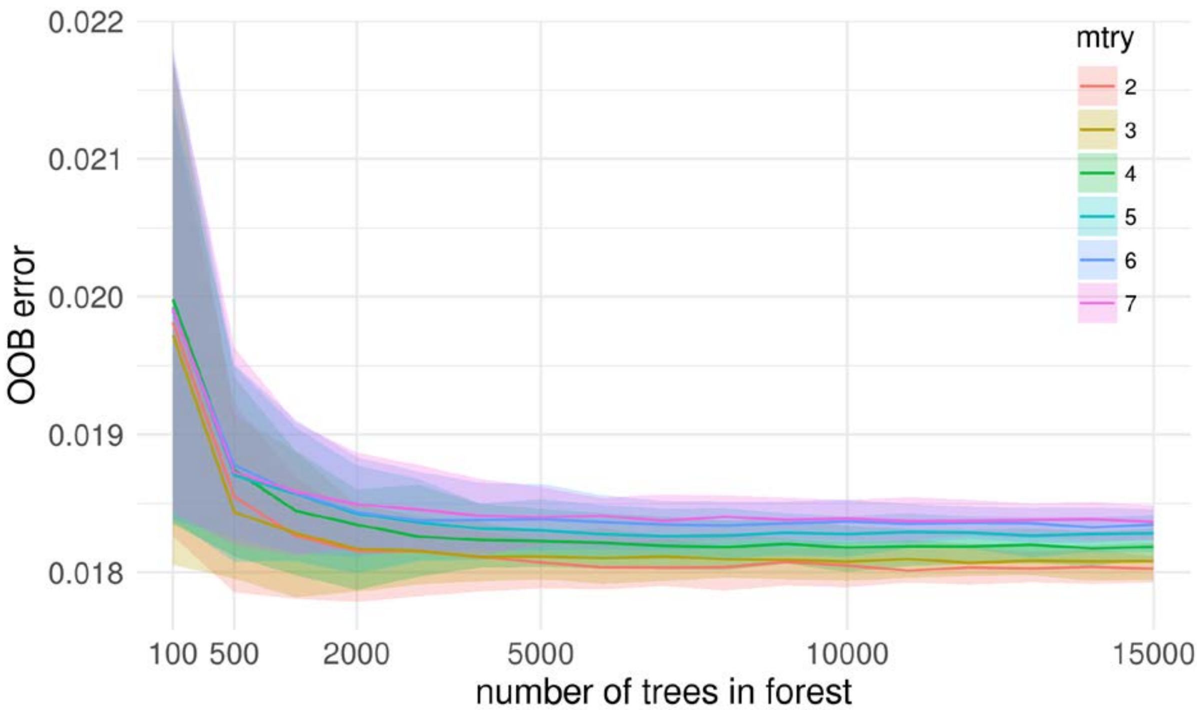

Figure 3: OOB error distributions of the citydb_sm data set for repeated simulations using different *mtry* values and increasing tree sizes

On this basis we work with RF models with *ntree* = 6000 and *mtry* = p/3 with p depending on the data sub-set used for learning predictive models for relative flood loss to residential buildings. Model performance is evaluated in terms of MSE, MAE, MBE, QR_{90}, and HR for all observed relative loss values in the data sample (n=80). The performance of RF models to reproduce the learning data is reported in Table 2. As mentioned earlier bootstrapping and the minimum number of data points in a node avoids overfitting of RF models. Model performance for independent data within a leave-one-out cross validation procedure is reported subsequently.

Table 2: RF model performance using different data sub-sets with city DB and sm models also including water depth from the empirical data and icpr standard model performance, best performance values are marked bold

model approach	**mtry**	**MAE**	**MBE**	**MSE**	**QR**	**HR**
cati complete	9	0.015	-0.009	0.0016	0.25	0.95
cati reduced	3	0.016	-0.008	**0.0013**	0.23	0.95
cityDB model	2	0.021	-0.011	0.0018	0.24	**0.93**
sm model	3	0.019	-0.009	0.0028	0.25	0.95
cityDB_sm model	3	**0.014**	-0.010	0.0017	0.26	0.97
sdf model	1	0.030	-0.003	0.0018	0.19	0.69
icpr model	-	0.135	**-0.002**	0.0481	NA	NA

The empirical survey data can be assumed to be the most comprehensive and most detailed set of information available for flood loss estimation to individual residential buildings (Thieken et al. 2007; Merz et al. 2013). Therefore, the models based on these data sub-sets (cati complete and cati reduced) represent upper benchmarks for the models based on standardized data sources. In contrast, the sdf model which uses only water depth as predictor for relative building loss is a lower benchmark for the models based on standardized data sources, because it is expected that using additional variables improves predictive model performance (Schröter et al. 2014). Further, we also include the icpr model in the comparison which is the standard model for flood risk management and planning of mitigation measures in the city of Dresden (Landeshauptstadt Dresden 2011). The icpr stage damage function ($relative\ loss = (2wd^2 + 2wd)/100$)) has been originally derived for regional risk assessment in the River Rhine (ICPR 2001) and has been transferred without any further local adjustments to the city of Dresden in the Elbe catchment.

The RF models based on standardized data sources achieve comparable performance values as the models based on the cati complete and cati reduced data sets. In particular the citydb_sm model which uses both data from the 3D city database and the spatial measures outperforms the alternatives in terms of MAE and MSE. However, the citydb model achieves best scores in terms of QR and HR and thus provides smallest predictive uncertainty intervals in combination with most reliable predictions; recall that HR is evaluated using the 95–5 quantile range, and thus should ideally yield a nominal coverage of 0.9.

The sdf model shows inferior performance particularly in terms of MAE and MSE which indicates less accurate predictions. In contrast, for the sdf model the QR is underestimated in comparison to the other models as it is combined with a clearly smaller HR, which indicates that only 69 % of the observations are within the QR_{90} interval. This is a signal for the predictions of the sdf model being less reliable than from the other models, which cover more than 90 % of the observations. One explanation for this is the number of variables used in the sdf model and the other models: while the sdf model uses only water depth to predict relative loss to residential buildings, the other models use up to 10 additional variables which in turn reveal additional sources of uncertainty. Hence, the sharpness of the predictions is reduced resulting in larger quantile ranges, but as the results show, these predictions are more accurate and more reliable with HR close to the nominal coverage of 0.9. This insight supports similar findings by Merz et al. (2013), Schröter et al. (2014), and Wagenaar et al. (2017) that using additional variables improves the predictive performance of flood loss models and reliability. Both, the sdf model and icpr use only water depth as a predictor. In comparison, the sdf model yields clearly smaller MAE and MSE values which emphasizes the usefulness of local data to derive a flood loss model as also pointed out by Cammerer et al. (2013), and the increased flexibility of the RF approach to reflect underlying complexities in comparison to an analytical function (Wagenaar et al. 2017).

As a next step we investigate the performance of the models to predict independent data, i.e. data which is unused to derive the models. As the data sample with only 80 observations is rather small it is not sensible to conduct a split sample testing procedure. Instead, we follow a leave-one-out cross validation procedure to test the transferability of the models.

In this regard, the models are derived 80 times while excluding each time one of the records from the learning sample. For this independent data point we predict the relative building loss and evaluate the predictive performance in terms of MBE, MSE, and MAE. In this way, samples of performance values for the 80 model tests are obtained which are illustrated as box plots in Figure 4.

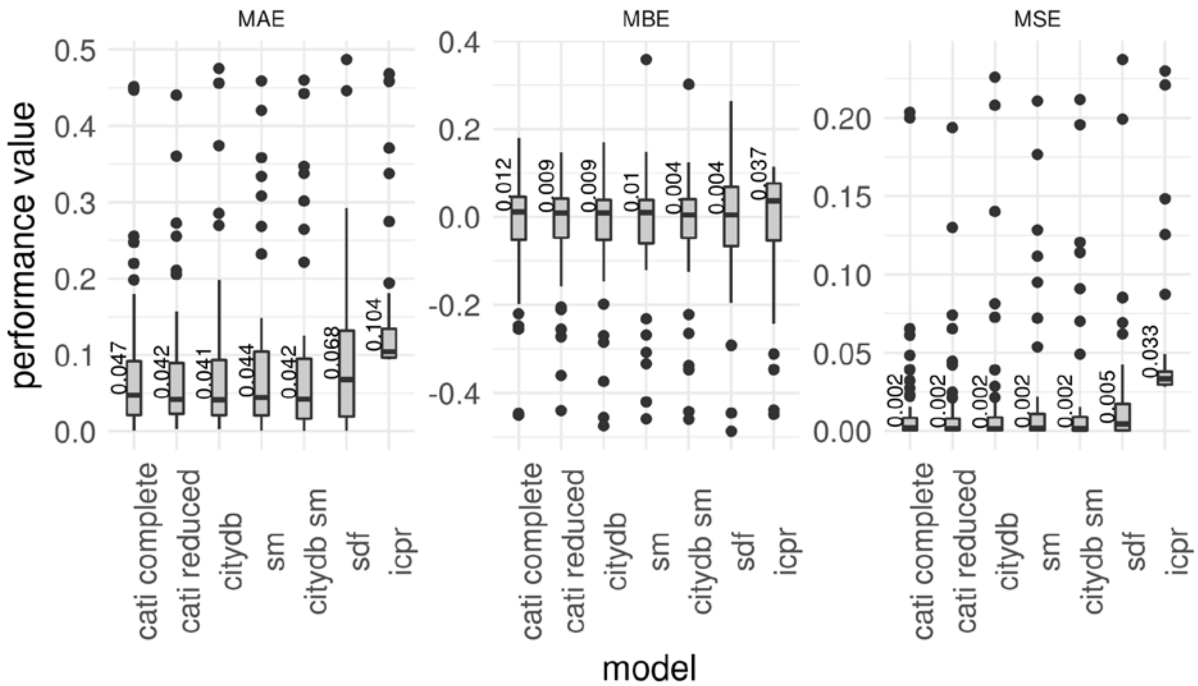

Figure 4: Model predictive performance in leave-one-out cross validation, median values for performance metrics are given as labels.

Again, the performance of the model based on standardized data sources is comparable the results of the cati based models. Both the citydb and citydb_sm models yield comparable results for MAE, MBE and MSE as well as in terms of median, inter quartile range, hinges, and outliers. The sdf model performs worse than the other models based on local data. Also the variability of sdf model performance is higher indicating less reliable model predictions in transfer applications. Similar to the performance values reported in, the icpr model performs clearly worse than the other models in particular regarding MAE and MSE. Obviously, models based on local data perform better than a model transferred from a different region which confirms the conclusions of Cammerer et al. (2013) and Schröter et al. (2014). In terms of MBE, the difference in performance of the icpr model is not that pronounced. This is due to the fact that for the given data sample, over and underestimation errors favorably cancel out.

Overall, the differences in model performance within the standardized data based approaches (citydb, sm, and citydb_sm) are rather small and it is impossible to conclude if one model is always better than all the others for all evaluation criteria. However, the results provide support for the hypothesis that standardized data sources provide useful information to describe the vulnerability characteristics of buildings for the estimation of flood losses to residential buildings with a comparable performance to models based on detailed empirical survey data. Hazard related information including inundation depth is still needed as an additional input variable.

For all models, the model performance boxplots show outliers, which indicate that a number of observations cannot be reproduced by the models well. These cases apparently deviate from the overall relationships derived from the data which offer too little support for capturing these particularities and adequately representing them in the models. Recall that flood loss estimation on the micro-scale, i.e. on the level of individual buildings, is particularly challenging (Dottori et al 2016), and as shown by Merz et al. (2004), given the large variability of individual objects is usually associated with large uncertainty. Neither the empirical survey data nor the standardized data used in this study do currently provide these details. In the future 3D city models could offer additional information, when data become available on higher levels of detail, i.e. LOD3 or LOD4. However, for current practical purposes the use of standardized data sources seems to enable flood loss estimation on a comparable level of performance as the empirical survey data. Further, 3D city models are readily available for urban areas and using the underlying standardized data for flood loss modeling will ease the spatial transfer of models. In addition, 3D city models also open for opportunities to embed risk oriented urban planning in these tools.

5. Prototype

5.1 Implementation

Beyond the usefulness of data about residential building characteristics for flood loss modeling, 3D city models offer additional possibilities to embed building vulnerability information into flood risk sensitive urban planning. This includes storing building related and other data in a standardized way, simulating urban and environmental processes and visualizing information (Gröger et al. 2012). To demonstrate the functioning and utilization of 3D city models for flood risk assessment and management, the flood loss model 'citydb' derived in section 3.2 has been implemented as the 3D city flood damage module (3DCFD) to the 3D City Database (3DCityDB). The 3DCFD module is applied to the city of Dresden (Germany) as an example. The 3DCityDB is a free 3D geo database to store and manage 3D city models, which is implemented as a schema compliant to CityGML 2.0 for the relational database management systems PostgreSQL/PostGIS and Oracle (Kunde et al. 2013). Hence, the 3DCityDB enables the efficient management and processing of large city models. The components and links of the 3DCFD prototype implementation are shown in Figure 5.

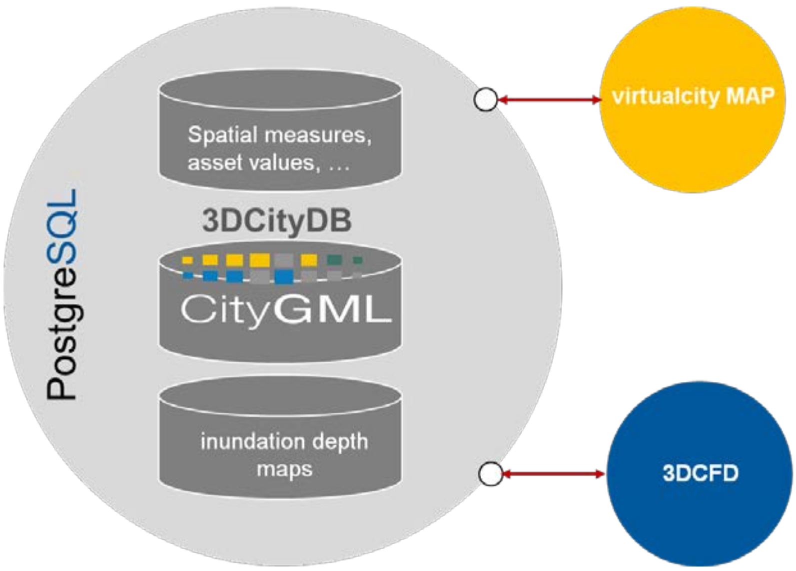

Figure 5 Prototype component diagram

The core of the PostgreSQL implementation form a number of database schema including the 3DCityDB which holds the geometrical, thematic, and generic attribute information of the 3D city model according to the CityGML standard. An additional schema contains further information about buildings like spatial measures or building asset values which are required for the calculation of economic loss. Further, spatial information about inundation extent and inundation depths are stored in a separate schema. The 3DCFD module is linked to the database using SQL queries. This module queries the attributes of the buildings affected by flooding and the inundation depths at the residential buildings for an inundation depth map of interest from the database. The result of this query provides the input to the flood loss model 'citydb', i.e. the RF flood loss model derived in section 3.2. The 3DCFD module calculates the relative loss for each affected building and returns them to the database where they are added to the features of the specific buildings using generic attributes. The concept of generic attributes allows extending CityGML applications during runtime by providing a name, a data type and a value. Thus, it is possible to add attributes that are not explicitly modeled in CityGML without changing the schema (Gröger et al., 2012). Loss modeling in combination with high resolution visualization allows identifying those buildings which should be protected with precautionary measures which is an important component of flood prevention and risk management concepts.

The prototype implementation demonstrates the functioning using the 3DCFD module as an example. Driven by the idea to provide a flexible framework, alternative simulation modules can be linked to the database. Hence, similar to the 3DCFD module any other flood loss model which uses input data provided by the database can be linked to the framework using appropriate SQL queries. Further processing of the data to calculate economic losses to residential buildings as the product of relative loss and building asset values is straightforward and can be easily conducted using spatial SQL queries or GIS software tools.

In addition, the prototype offers functionalities to visualize the loss estimation results and the underlying data in an interactive web framework (virtualcityMAP). The user may browse through the 3D city model, colorize the residential buildings regarding their relative damage values caused by different flooding scenarios, and get detailed information for individual buildings, see Figure 6. This prototype application is based on the virtualcityMAP technology. It contains an integrated viewer for oblique images, 2D and 3D maps. The virtualcityMAP can be used from any computer or workstation with a modern browser which enables the execution of JavaScript and WebGL. WebGL is an API for rendering 3D computer graphics within a browser without the use of plugins; see Schröter et al. (2017) for an online version of the prototype.

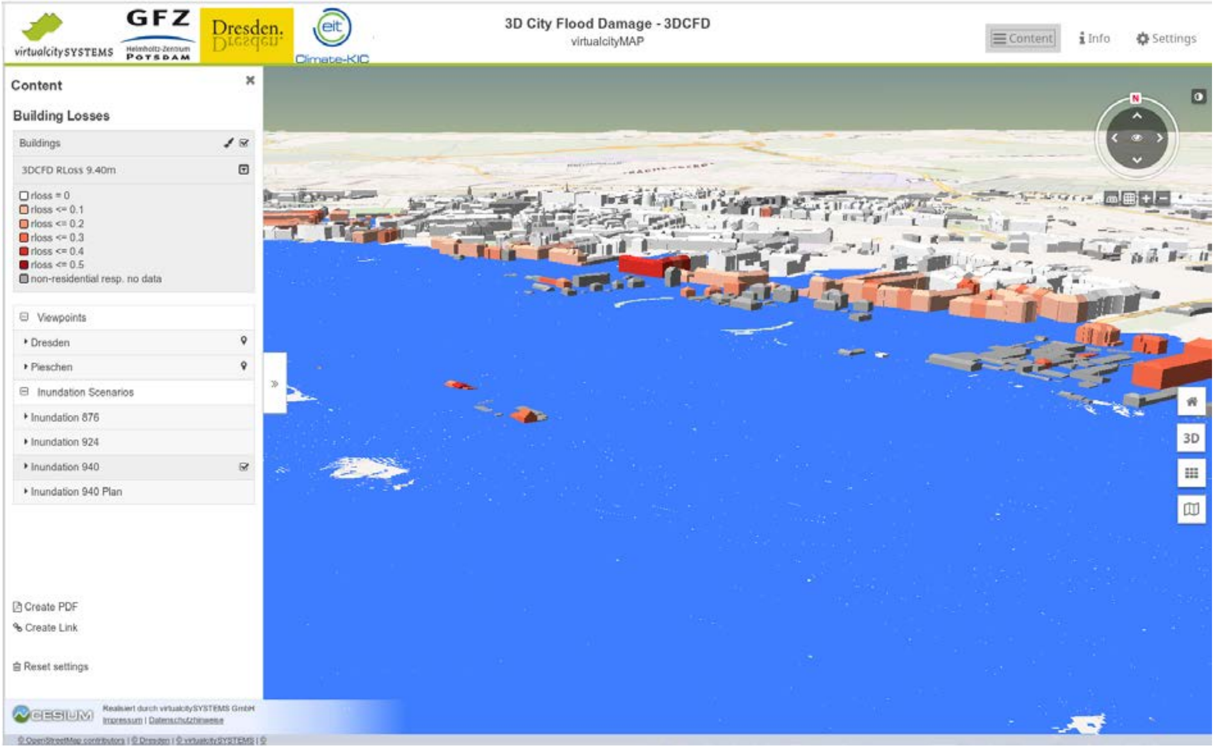

Figure 6: 3DCFD online prototype implementation showing the inundation area for the flood scenario 9.4 m in Dresden with color coded residential buildings according to relative building loss.

5.2 Application example Dresden

The 3DCFD module is used to estimate flood loss to residential buildings for three flood scenarios within the focus area of Pieschen in Dresden (see Figure 1). The flood scenarios are defined in terms of water levels above the Dresden gauge datum (112.3 m): 8.76 m which correspond to the June 2013 flood (50 years return period), 9.24 m, which is the statistically expected 1% (1 in 100) years flood, and 9.40 m which has been registered during the August 2002 flood (approx. 200 years return period). The 3DCFD loss estimation results are compared with the outcomes of the icpr model on the level of individual buildings (in total 272 residential buildings) as well as on the aggregated level of the entire focus area (33.6 ha). Figure 7 illustrates the inundation depth and the resulting relative losses to the buildings affected in the Pieschen area for the 9.4 m flood scenario which affects 95 residential buildings with maximum inundation depths of 1.5 m.

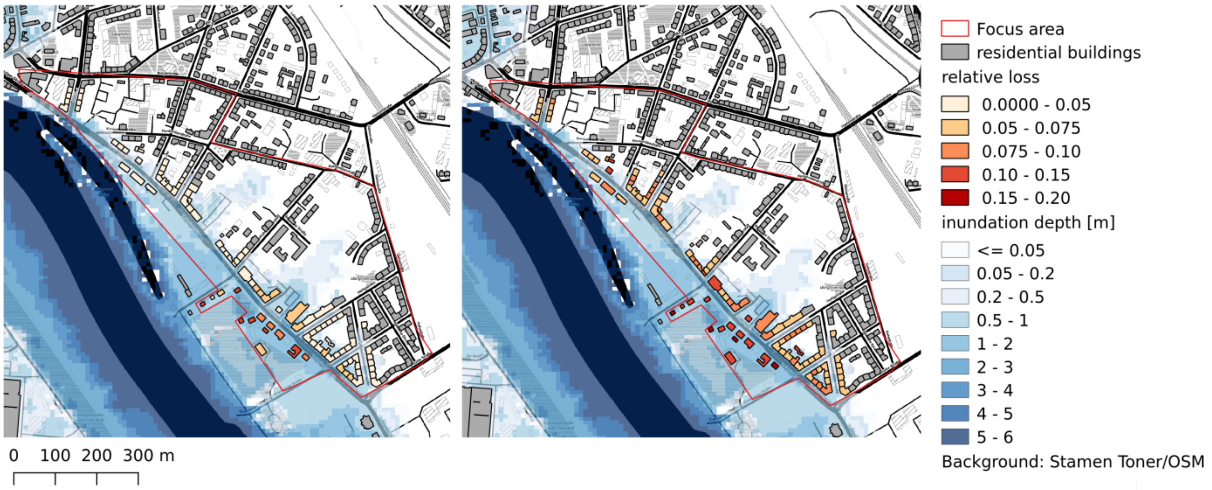

Figure 7: Relative loss estimates for residential buildings in the focus area Pieschen for 9.4 m flood scenario: left panel icpr model, right panel citydb model

As can be seen by the color codes of the affected buildings, the icpr model tends to predict lower relative damages than the citydb model. To illustrate this point, Figure 8 compares the relative loss estimates of both models as a function of inundation depth. The citydb model consistently estimates higher loss ratios than the icpr model with a more pronounced difference (ca. 0.05) for inundation depths values below 0.5 m. The plot also shows the relative loss values and inundation depths from the cati data set for the whole Dresden area but limited to the range of inundation depths present in the Pieschen focus area for the flood scenarios.

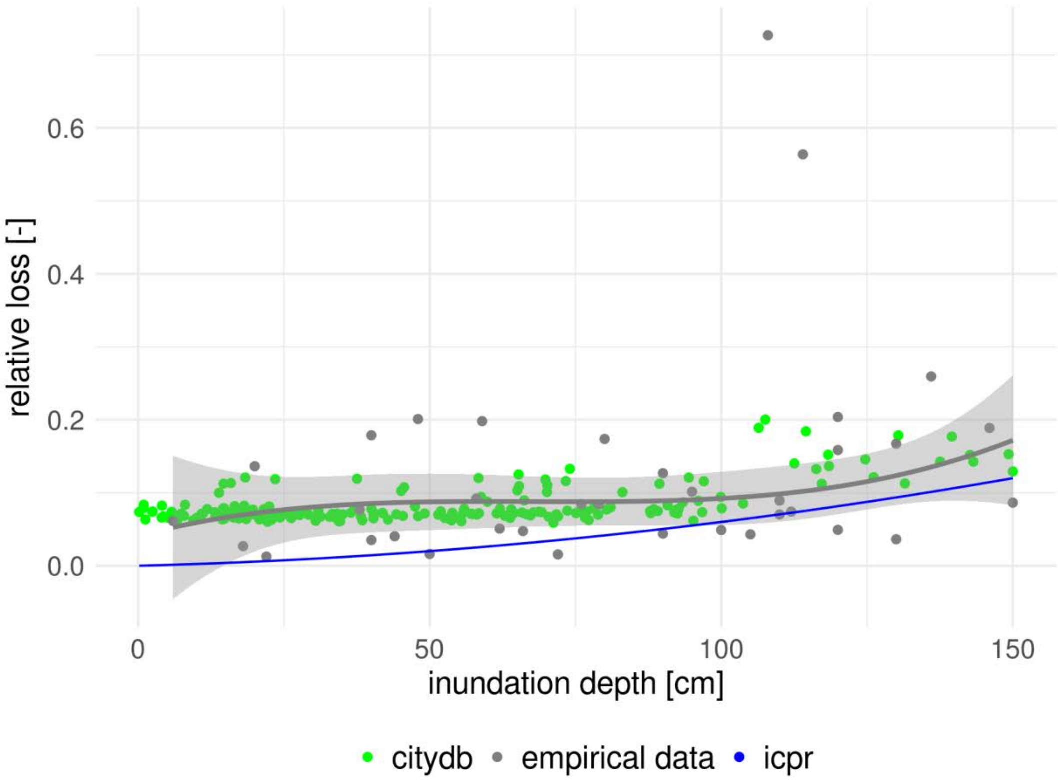

Figure 8: Relative loss to residential buildings as a function of inundation depth in the Pieschen focus area for different inundation scenarios using icpr and citydb flood loss models, empirical data are from the cati surveys for the whole Dresden area. The grey line is a spline approximation with standard errors to the cati data (neglecting the outliers).

The citydb model nicely reproduces the mean of the empirical relative loss values along the range of inundation depths which is approximated by a spline interpolation (grey line in Fig.8) without taking the outliers into account. As the citydb model uses additional variables like building area and measured height, it reproduces parts of the variability of relative loss for comparable inundation depths which is also visible for the cati data.

The implications of these differences in relative loss estimates on the estimation of risk in terms of the expected annual damage (EAD) are investigated for the Pieschen focus area. For the calculation of economic loss the regional values for residential building stock derived by Kleist et al. (2006) are disaggregated to the level of individual buildings. The product of relative loss and building value yields an estimate for the economic loss in terms of the reinstatement costs for the affected buildings in the different flood scenarios, Table 3. EAD is calculated as the probability weighted sum of losses for different flood scenarios. Probability weights are defined as the differences of flood probabilities (inverse of return period) between different flood scenarios

Table 3: Flood loss to residential buildings in the focus area Pieschen for different flood scenarios using icpr and citydb flood loss model

Flood scenario	approx. Flood probability	Loss icpr model [€]	Loss citydb model [€]
9.40	0,005	3,611,000	7,928,000
9.24	0,01	2,245,000	6,848,000
8.76	0,02	629,000	3,940,000
	EAD	**33,200**	**118,000**

The differences in economic losses for both models are considerable. Notably, larger differences occur for higher probable flood scenarios which are associated with smaller inundation depths where the differences in the loss models are most distinct. This directly translates into considerable deviations in the calculation of the EAD as the probability weighted mean of loss incurred for discrete flood scenarios (Kaplan and Garrik 1981). Using the citydb model instead of the icpr model yields a difference of a factor 3.5 in EAD which results in a substantially different assessment of investments in flood mitigation measures. These findings are in line with the results published by (Wagenaar et al 2016) and underline the high relevance of flood loss model uncertainty in risk assessments. In this respect, the 3DCFD prototype makes a contribution to overcome the reluctance to use more sophisticated and complex modeling approaches instead of overly simplified models either due to convenience and/or lower efforts.

6. Discussion

Flood loss modeling on the micro scale, i.e. on the level of individual buildings, is important to optimize investments for the implementation of flood risk management concepts in urban areas. The application and even more the spatial transfer of multi-variable flood loss models are challenging, because input data need to be available with high spatial detail. Standard data sources may ease the set-up of flood loss models and spatial transfer.

3D city models can store detailed information of building locations and building characteristics. However, presently 3D city models mostly include information only on LOD1 or LOD2 which corresponds to geometric building information. Therefore, numerical spatial measures derived from remote sensing data are another promising standardized data source to feed flood loss models. This paper investigates the potential of 3D city models and numerical spatial measures to support the estimation of flood losses to residential buildings. The study is carried out in the city of Dresden where data on relative flood losses and inundation depths are available from empirical surveys which can be used to derive and to validate new flood loss models. For this purpose Random Forests are used to first analyze the usefulness of variables available from 3D city models and remote sensing based spatial measures and to identify the most important ones to explain flood loss.

The sample size of 80 available data points for the study may influence the outcomes of variable importance and predictive model performance analyses. On the one hand it is possible that the Random Forest does not find any reasonable splits in the data space, and thus does not provide any meaningful predictions. On the other hand it may overfit the data and achieve overly optimistic performance results. The study includes several steps to reduce these potential influences. First bootstrapping i.e. using only a subset of data to learn an individual tree of the forest, a minimum number of five data points per node, and a limited number of candidate variables for each split effectively reduce overfitting. Second, within the leave-one-out cross-validation procedure model predictive performance is tested against independent data. Third, the models derived are assessed using more complex data sets for model learning and a simple standard flood loss model as benchmarks. The performance of the models using standardized input data ranges between both benchmarks.

The acid test for the derived flood loss models will be the application and performance testing in another city. In this context, the ability to assimilate local data will be an interesting question to answer.

For practical applications the trade-off between number of input variables, model performance and effort to collect input data needs further investigation. In our study we assumed that all input data are readily available with the 3D city model. Also, the various spatial measures are derived from remote sensing using automated classification algorithms. Therefore, we assume that all variables will be generated within the same data processing operation. However, the collection of individual variables may require substantial efforts. Therefore, the knowledge about their relevance/importance for loss modeling is valuable to set priorities.

The exemplary calculation of the expected annual damage within the application case in the city of Dresden shows the potential effects of model biases on the outcomes of economic assessments and decisions about efficient measure. This emphasizes the importance of model validation and selection. In this regard, the information about predictive uncertainty provided by the predictive distribution of Random Forests offers additional possibilities for quantifying the reliability of model predictions. How this additional information can be used to support decision making and how it can be visualized to illustrate and communicate uncertainty are interesting questions for following research. Overall the 3D city model based technology shows good opportunities to use more sophisticated and complex modeling approaches by exploiting available data sources.

Beyond that, 3D city models also open for opportunities to embed building vulnerability information into flood risk sensitive urban planning including the storage of a variety of ancillary information, simulation of urban and environmental processes and visualization. Concerning simulations, the use of building geometries as boundary conditions for hydro-numerical simulation of flood dynamic processes is obvious. Concerning visualization, this tool provides large potential to increase transparency in risk assessment, to support the identification of damage hotspots as well as to improve risk communication and raising awareness.

7. Conclusions

The assessment of variable importance reveals that those variables of standardized data sources which provide information about building geometric properties such as building area, height, roof type, shape, and density of buildings are most suitable to explain flood loss. All these variables are directly or indirectly linked to different building types and associated characteristics.

Predictive flood loss models are derived using a set of important variables. Model validation confirms that using multiple variables for flood loss modeling improves the predictive performance and reliability. The results also show that standardized data sources provide useful information for the estimation of flood losses to residential buildings. For current practical purposes the use of standardized data sources enables predictive flood loss modeling with a comparable performance as modeling based on detailed empirical survey data.

Further, 3D city models are readily available for urban areas and using the underlying standardized data for flood loss modeling will ease the spatial transfer of models. In the future, even more can be expected from 3D city models, when data become available on higher levels of detail, i.e. LOD3 or LOD4, which provides details about building openings, sill level heights and building internal facilities.

Acknowledgements

Parts of this work have been supported by funding for the 3D city flood damage module (3DCFD) pathfinder project through the Climate Knowledge and Innovation Community (Climate-KIC) by the European Institute of Innovation and Technology (EIT). We are thankful to the city of Dresden for the provision of data including the flood scenarios, the virtual 3D city model, and the digital terrain model.

References

Analytics, R. and Weston, S.: doParallel: Foreach Parallel Adaptor for the “parallel” Package. [online] Available from: https://CRAN.R-project.org/package=doParallel, 2015a.

Analytics, R. and Weston, S.: foreach: Provides Foreach Looping Construct for R. [online] Available from: https://CRAN.R-project.org/package=foreach, 2015b.

Amirebrahimi, S., Rajabifard, A., Mendis, P., and Ngo, T.:A framework for a microscale flood damage assessment and visualization for a building using BIM–GIS integration. International Journal of Digital Earth, 9(4): 363-386. DOI:10.1080/17538947.2015.1034201, 2016

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

Balica, S. F., Popescu, I., Beevers, L. and Wright, N. G.: Parametric and physically based modelling techniques for flood risk and vulnerability assessment: A comparison, Environmental Modelling & Software, 41, 84–92, doi:10.1016/j.envsoft.2012.11.002, 2013.

Biljecki, F., Stoter, J., Ledoux, H., Zlatanova, S., Çöltekin, A.: Applications of 3D City Models: State of the Art Review. ISPRS International Journal of Geo-Information, 4(4): 2842, 2015.Bochow, M. : Automatisierungspotenzial von Stadtbiotopkartierungen durch Methoden der Fernerkundung. Dissertation, Universität Osnabrück, 189 p, 2010

Bochow, M., Taubenböck, H., Segl, K., and Kaufmann, H.: An automated and adaptable approach for characterizing and partitioning cities into urban structure types. In Proceedings of the 2010 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Honolulu, HI, USA,25–30 July 2010; pp. 1796–1799, 2010

Breiman, L., Friedman, J., Stone, C. J. and Olshen, R. A.: Classification and Regression Trees, New Ed., Taylor & Francis Ltd., 1984.

Breiman, L.: Bagging predictors, Machine Learning, 24(2), 123–140, doi:10.1007/BF00058655, 1996.

Breiman, L.: Random Forests, Machine Learning, 45(1), 5–32, doi:10.1023/A:1010933404324, 2001.

Bubeck, P., Kreibich, H., Penning-Rowsell, E. C., Botzen, W. J. W., de Moel, H. and Klijn, F.: Explaining differences in flood management approaches in Europe and in the USA - a comparative analysis: Differences in flood management approaches, Journal of Flood Risk Management, 10(4), 436–445, doi:10.1111/jfr3.12151, 2017.

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

Chapi, K., Singh, V. P., Shirzadi, A., Shahabi, H., Bui, D. T., Pham, B. T. and Khosravi, K.: A novel hybrid artificial intelligence approach for flood susceptibility assessment, Environmental Modelling & Software, 95, 229–245, doi:10.1016/j.envsoft.2017.06.012, 2017.

Chinh, D. T., Gain, A. K., Dung, N. V., Haase, D. and Kreibich, H.: Multi-Variate Analyses of Flood Loss in Can Tho City, Mekong Delta, Water, 8(1), 6, doi:10.3390/w8010006, 2015.

de Moel, H., Jongman, B., Kreibich, H., Merz, B., Penning-Rowsell, E. and Ward, P. J.: Flood risk assessments at different spatial scales. - Mitigation and Adaptation Strategies for Global Change, 20, 6, p. 865-890, 2015.

Dietz H. Wohngebäudeversicherung Kommentar, 2nd ed. Karlsruhe, Germany: VVW Verlag Versicherungswirtschaft GmbH, 1999.

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 2007, L288/27, 2007.

Dottori, F., R. Figueiredo, M. L. V. Martina, D. Molinari, and Scorzini, A. R.. "INSYDE: A Synthetic, Probabilistic Flood Damage Model Based on Explicit Cost Analysis." *Nat. Hazards Earth Syst. Sci.* 16, no. 12 2577–91. doi:10.5194/nhess-16-2577-2016, 2016.

Genuer, R., Poggi, J.-M. and Tuleau-Malot, C.: Variable selection using random forests, Pattern Recognition Letters, 31(14), 2225–2236, doi:10.1016/j.patrec.2010.03.014, 2010.

GeoEye Inc: IKONOS Imagery Products Guide, Version 1.5, http://glcf.umd.edu/library/guide/IKONOS_Product_Guide_jan06.pdf, last access: Jan 9th 2018, 2006.

Gerl, T., Bochow, M. and Kreibich, H.: Flood Damage Modeling on the Basis of Urban Structure Mapping Using High-Resolution Remote Sensing Data, Water, 6(8), 2367–2393, doi:10.3390/w6082367, 2014.

Gerl, T., Kreibich, H., Franco, G., Marechal, D., and Schröter, K..:A Review of Flood Loss Models as Basis for Harmonization and Benchmarking. *PLOS ONE* 11, no. 7: e0159791. doi:10.1371/journal.pone.0159791, 2016.

Gneiting, T. and Raftery, A. E.: Strictly Proper Scoring Rules, Prediction, and Estimation, Journal of the American Statistical Association, 102(477), 359–378, doi:10.1198/016214506000001437, 2007.

Graesser, J., Cheriyadat, A., Vatsavai, R.R., Chandola, V., Long, J., and Bright, E.: Image Based Characterization of Formal and Informal Neighborhoods in an Urban Landscape. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 5, 1164-1176, 2012.

Grigg, N. S. and Helweg, O. J.: State-of-the-art of estimating flood damage in urban areas, Water Resour. Bull., 11(2), 37–390, 1975.

Gröger, G., Kolbe, T. H., Nagel, C. and Häfele, K.-H.: OpenGIS City Geography Markup Language (CityGML) Encoding Standard, Version 2.0.0. Technical report Doc. No. 12-019. Wayland (MA), USA: OGC. 2012.

Gröger, G. and Plümer, L.: CityGML - Interoperable semantic 3D city models. ISPRS Journal of Photogrammetry & Remote Sensing 71:12–33. 2012.

Guyon, I. and Elisseeff, A.: An Introduction to Variable and Feature Selection, Journal of Machine Learning Research, 3(Mar), 1157–1182, 2003.

Hapfelmeier, A. and Ulm, K.: A new variable selection approach using Random Forests, Computational Statistics & Data Analysis, 60, 50–69, doi:10.1016/j.csda.2012.09.020, 2013.

Hasanzadeh Nafari, R., Ngo, T. and Mendis, P.: An Assessment of the Effectiveness of Tree-Based Models for Multi-Variate Flood Damage Assessment in Australia, Water, 8(7), 282, doi:10.3390/w8070282, 2016.

Huang, B. F. F. and Boutros, P. C.: The parameter sensitivity of random forests, BMC Bioinformatics, 17(1), doi:10.1186/s12859-016-1228-x, 2016.

ICPR: Rhine-Atlas, International Commission for the Protection of the Rhine, Koblenz. [online] Available from: https://www.rheinatlas.de, 2001.

Jensen, J.R. and Cowen, D.C.: Remote Sensing of Urban/Suburban Infrastructure and Socio-Economic Attributes. Photogrammetric Engineering and Remote Sensing 65(5), 611-622, 1999.

Kaplan, S. and Garrick, B. J.: On The Quantitative Definition of Risk, Risk Analysis, 1(1), 11–27, doi:10.1111/j.1539-6924.1981.tb01350.x, 1981.

Kleist, L., Thieken, A. H., Köhler, P., Müller, M., Seifert, I., Borst, D. and Werner, U.: Estimation of the regional stock of residential buildings as a basis for a comparative risk assessment in Germany, Nat. Hazards Earth Syst. Sci., 6(4), 541–552, doi:10.5194/nhess-6-541-2006, 2006.

Kreibich, H., Botto, A., Merz, B. and Schröter, K.: Probabilistic, Multivariable Flood Loss Modeling on the Mesoscale with BT-FLEMO, Risk Analysis, 37(4), 774–787, doi:10.1111/risa.12650, 2017.

Kreibich, H., Thieken, A.H., Haubrock, S. and Schröter, K.: HOWAS21 the German flood damage database. in: Flood damage survey and assessment: new insights from research and practice (D. Molinari, S. Menoni, F. Ballio, eds.) AGU, Geophysical Monograph 228, Wiley, Chapter 5, 65-78, 2017b

Kreibich, H., Seifert, I., Merz, B. and Thieken, A. H.: Development of FLEMOcs – a new model for the estimation of flood losses in the commercial sector, Hydrological Sciences Journal, 55(8), 1302–1314, doi:10.1080/02626667.2010.529815, 2010.

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

Kunde, F., Nagel, C., Herrerruela, J., Ross, L. and Kolbe, T. H.: 3D-Stadtmodelle in PostGIS mit der 3D City Database, in Tagungsband der FOSSGIS-Konferenz., 2013.

Landeshauptstadt Dresden. “Plan Hochwasservorsorge Dresden.” Dresden: Landeshauptstadt Dresden, Geschäftsbereich Wirtschaft, Umweltamt, June 23, 2011.

Liaw, A. and Wiener, M.: Classification and Regression by randomForest, R News, 2(3), 18–22, 2002.Merz, B. andThieken, A. H.: Flood risk analysis: Concepts and challenges. - Österreichische Wasser- und Abfallwirtschaft, 56, 3-4, p. 27-34, 2004.

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

Merz, B., Hall, J., Disse, M. and Schumann, A.: Fluvial flood risk management in a changing world. - Natural Hazards and Earth System Sciences (NHESS), 10, 3, p. 509-527, 2010a

Merz, B., Kreibich, H., Schwarze, R. andThieken, A.: Review article 'Assessment of economic flood damage'. - Natural Hazards and Earth System Sciences (NHESS), 10, 8, p. 1697-1724, 2010b.

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

Meyer V. and Messner F.: National Flood Damage Evaluation Methods: A Review of Applied Methods in England, the Netherlands, the Czech Republic and Germany. UFZ-Discussion Papers, www.floodsite.net, 2005.

Penning-Rowsell, E., Johnson, C., Tunstall, S., Tapsell, S., Morris, J., Chatterton, J. and Green, C.:The Benefits of Flood and Coastal Risk Management: A Handbook of Assessment Techniques. London: Middlesex University Press, 2005.

Ripley, B. and Lapsley, M.: RODBC: ODBC Database Access. [online] Available from: https://CRAN.R-project.org/package=RODBC, 2017.

Saeys, Y., Inza, I. and Larrañaga, P.: A review of feature selection techniques in bioinformatics, Bioinformatics, 23(19), 2507–2517, doi:10.1093/bioinformatics/btm344, 2007.

Schloerke, B., Crowley, J., Cook, D., Briatte, F., Marbach, M., Thoen, E., Elberg, A. and Larmarange, J.: GGally: Extension to “ggplot2.” [online] Available from: https://CRAN.R-project.org/package=GGally, 2017.

Schröter, K., Kreibich, H., Vogel, K., Riggelsen, C., Scherbaum, F. and Merz, B.: How useful are complex flood damage models?, Water Resour. Res., 50(4), 3378–3395, doi:10.1002/2013WR014396, 2014.

Schröter, K., Redweik, R., Lüdtke, S., Meier, J., Bochow, M., Kreibich, H., Ross, L. and Nagel, C.: 3D-ctiy Flood Damage Module prototype implementation, doi:10.5880/GFZ.5.4.2017.001, 2017.

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

Spekkers, M. H., Kok, M., Clemens, F. H. L. R. and ten Veldhuis, J. A. E.: Decision-tree analysis of factors influencing rainfall-related building structure and content damage, Nat. Hazards Earth Syst. Sci., 14(9), 2531–2547, doi:10.5194/nhess-14-2531-2014, 2014.

Stadler, A. and Kolbe, T. H.: Spatio-semantic coherence in the integration of 3D city models. In Proceedings of the 5th International ISPRS Symposium on Spatial Data Quality ISSDQ 2007 in Enschede, The Netherlands, 13-15 June 2007, edited by Alfred Stein. ISPRS Archives. ISPRS. 2007.

Takeuchi, K. : Increasing vulnerability to extreme floods and societal needs of hydrological forecasting. Hydrol. Sci. J. 46(6), 869–881, 2001.

Taubenböck, H., Wurm, M., Setiadi, N., Gebert, N., Roth, A., Strunz, G., Birkmann, J. and Dech, S.: Integrating remote sensing and social science - The correlation of urban morphology with socioeconomic parameters, 2009 Joint Urban Remote Sensing Event, pp. 1-7, 2009.

Teng, J.: Flood inundation modelling: A review of methods, recent advances and uncertainty analysis. Environmental Modelling and Software 90, 201–216, 2017.

Thieken, A. H., Müller, M., Kreibich, H. and Merz, B.: “Flood Damage and Influencing Factors: New Insights from the August 2002 Flood in Germany.” *Water Resources Research* 41 (12): 1–16. doi:10.1029/2005WR004177, 2005.

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

Thieken, A. H., Olschewski, A., Kreibich, H., Kobsch, S. and Merz, B.: Development and evaluation of FLEMOps - a new Flood Loss Estimation MOdel for the private sector, London., 2008.

Thieken, A., H. Kreibich, M. Müller, and J. Lamond: Data collection for a better understanding of what causes flood damage: experiences with telephone surveys: in Flood damage survey and assessment: new insights from research and practice (D. Molinari, S. Menoni, F. Ballio, eds.) AGU, Geophysical Monograph 228, Wiley, Chapter 7, 95-106, 2017.

Uuemaa, E., Antrop, M., Roosaare, J., Marja, R., and Mander, Ü. (2009). Landscape Metrics and Indices: An Overview of Their Use in Landscape Research. Living Reviews in Landscape Research, 3,2009.

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

Wagenaar, D. J., Bruijn, K. M. de, Bouwer, L. M. and Moel, H. de: Uncertainty in flood damage estimates and its potential effect on investment decisions, Natural Hazards and Earth System Sciences, 16(1), 1–14, doi:10.5194/nhess-16-1-2016, 2016.

Wagenaar, D., de Jong, J. and Bouwer, L. M.: Multi-variable flood damage modelling with limited data using supervised learning approaches, Nat. Hazards Earth Syst. Sci., 17(9), 1683–1696, doi:10.5194/nhess-17-1683-2017, 2017.

Wang, Z., Lai, C., Chen, X., Yang, B., Zhao, S. and Bai, X.: Flood hazard risk assessment model based on random forest, Journal of Hydrology, 527, 1130–1141, doi:10.1016/j.jhydrol.2015.06.008, 2015.

Wickham, H.: Reshaping Data with the reshape Package, Journal of Statistical Software, 21(12), 1–20, 2007.

Wickham, H.: ggplot2: Elegant Graphics for Data Analysis, Springer-Verlag New York. [online] Available from: http://ggplot2.org, 2009.

Wickham, H.: scales: Scale Functions for Visualization. [online] Available from: https://CRAN.R-project.org/package=scales, 2016.

Wickham, H.: stringr: Simple, Consistent Wrappers for Common String Operations. [online] Available from: https://CRAN.R-project.org/package=stringr, 2017.

Wickham, H. and Henry, tidyr: Easily Tidy Data with 'spread()' and 'gather()' Functions. [online] Available from: https://CRAN.R-project.org/package=tidyr, 2017.

Wickham, H., Francois, R., Henry, L. and Müller, K.: dplyr: A Grammar of Data Manipulation. [online] Available from: https://CRAN.R-project.org/package=dplyr, 2017a.

Wickham, H., Hester, J. and Francois, R.: readr: Read Rectangular Text Data. [online] Available from: https://CRAN.R-project.org/package=readr, 2017b.

Zambrano-Bigiarini, M.: hydroGOF: Goodness-of-fit functions for comparison of simulated and observed hydrological time series. [online] Available from: https://CRAN.R-project.org/package=hydroGOF, 2014.

Zoltak, M., Ripley, B. and Lapsley, M.: RODBCext: Parameterized Queries Extension for RODBC. [online] Available from: https://CRAN.R-project.org/package=RODBCext, 2017.