EVALUATION OF STATISTICAL METHODS FOR PLUVIAL FLOOD RISK ASSESSMENT

Clemens Nocker, Gregor Laaha

Institute of Statistics, University of Natural Resources and Life Sciences, Vienna (BOKU)

Our Project is funded by the Interreg CENTRAL EUROPE Programme that encourages cooperation on shared challenges in central Europe and is supported under the European Regional Development Fund.

Evaluation of statistical methods for pluvial flood risk assessment

Version 9 July 2019

Authors DI Clemens Nocker, Assoc.Prof. DI Dr.techn. Gregor Laaha

Institute of Statistics, University of Natural Resources and Life Sciences, Vienna (BOKU)

Contracted by

Environment Agency Austria, Surface Water Unit

ENVIRONMENT AGENCY AUSTRIA umweltbundesamt

Co-funded by

Federal Ministry
Republic of Austria
Sustainability and Tourism

Acknowledgements

The study is based on the master thesis of DI Clemens Nocker, conducted at the Institute of Statistics in order of the Environment Agency Austria (UBA). We would like to thank Dr. Yvonne Spira for her constructive feedback, which has greatly improved the analysis. We further thank the Austrian Hail Insurance for providing flood-damage data, the Central Institute for Meteorology and Geodynamics (ZAMG) for providing precipitation data, the Austrian Federal Research Centre for Forests (BFW) and the Office of the Provincial Government Upper Austria for providing hydrographic and pedological data and geoinformation used in the study.

Contents

1. Context and goals of this study

1.1. Project context

Heavy rain events are a major environmental risk in Europe: they can hit any location with only very short warning time. Every year people die, thousands lose their homes, and environmental damages like water pollution occur. And the risks of heavy rain events are increasing all over Europe. In the project RAINMAN, partners from 6 countries have joined to develop and test innovative methods and tools for the integrated management of heavy rain risks by local, regional & national public authorities. These will be included in the RAINMAN-Toolbox, a set of five transferable tools and methods for municipalities and regional stakeholders.

The first tool of this toolbox is concerned with assessing and mapping heavy rain risks. It builds the foundation of the other tools, which will have their focus on the areas with a high risk. This study will test if and how well it is possible to assess those risks by using statistical methods.

1.2. Goals

The goal of the RAINMAN project is *"to reduce the losses in the natural and built environment caused by heavy rain"* (Rainman n.d.). Therefore an identification is needed which regions are under risk of pluvial floods occurring after heavy rain events. One approach to assess these risks is the usage of statistical modelling methods. The goal of this approach is to predict pluvial flood risk using reports of past flood damages on agricultural land and location characteristics. With this it is possible to focus the tools of the project to areas that are actually at risk. Furthermore, it can recommend collecting specific data which is not available yet even if the methods don't prove to be successful.

After specifying the methods, the most destructive events of the investigation period are identified and further investigated. In this event analysis it shall be determined which seasons are typical for big events that cause pluvial flood damages. In a further step the question shall be answered, if the cause of the floods was a continuous low intensity rainfall or a spontaneous heavy rainfall.

Based on this analysis the representative reported cells are investigated for their locations. The question is whether they have similar location values and therefore show locations with a higher vulnerability or if they differ. Each location variable is analysed separately to assess, if they show noticeable differences between cells with reports and cells without.

The first goal of the statistical models is to identify significant variables and their influence on the flood damages. In the last step the model with the best prediction values shall be determined and further discussed on practical usability.

1.3. Approach and structure

The first step of this study is building knowledge on the state of the art of prediction models for pluvial flood risk assessment. In the next step, out of the different data sets characteristic values need to be extracted for each cell. This is done with the programmes ArcGIS 10.3 (ESRI 2016) and R 3.4.3 (R Core Team 2017). After separating pluvial and fluvial flood damages an analysis based on single heavy rain events is conducted. Here the reports of pluvial flood damages are ordered by date and compared to the seasonal and meteorological conditions like precipitation amount and antecedent moisture conditions. The aim of this step is to identify typical seasons and to get a better understanding of the generating processes. Additionally, the locations of some of the affected cells are compared to each other before each location parameter is split into cells with reported pluvial flood damages and cells without. For that a presence/absence raster is calculated with a spatial resolution of 1x1km^2. Then the location analysis is summarised and possible interactions are identified. In the statistical modelling two different models are presented, one generalized linear model (GLM) and one random forest. Each model has two different data frames as inputs: the original data frame and the second one with aggregated classes. For the GLM weights are added and compared to each other. The random forest is carried out without weights, because no weighting is implemented in the current version of the package (v.4.6-14).

The report is structured as follows: First the data is described in detail with the study area first followed by each variable. Then the methods for the two analysis steps and modelling are presented. After that the results of the event and location analysis are shown separately. The final results are from the statistical models where the best results are presented and compared to each other through prediction diagnostics. At the end the results are summarized, discussed and the whole study is then summarized in the conclusion.

2. Review of current literature

Pluvial floods have been recognised as problem only a few years ago, and a number of studies are currently being published in this novel research field. However, the current literature on pluvial flooding focuses primarily on assessing which parts of a city are most likely to be flooded. Frequently, process-based analysis methods are used to calculate pluvial flood risk maps, while statistical methods are mainly used to calculate precipitation values. A few examples are shown below.

Similar to our study, Sörensen and Mobini (2017) used insurance claims for their study. They calculated the maximum rainfall volumes for 15 minutes to 12 days at a 50*50m grid at the city of Malmö, Sweden. Based upon this, they created a presence/absence grid of flooded cells and assigned them to systems, which indicated the drainage system type and the distance to major flow paths. They came to the conclusion that most of the biggest events were caused by heavy rainfall and occurred in summer. The amount of rain that fell was exceeded the sewerage system capacities, so about half of the amount was flowing overland. While the severe flooding events were caused by consistent rainfall during several days over the whole city, a few events were caused by highly localized rainfall. Events caused by the consistent rainfall were more evenly spread while the latter were around the main sewer where the water was led to lower areas.

Guerreiro et al. (2017) calculated a 10 year return period (RP10) of rain for most of Europe, using a regression model, and tried to predict how many from the over 500 cities are going to be flooded. To create an urban flood model they used the City Catchment Analysis Tool based on a DEM, which provides a simulation of urban hydrodynamics. The main problem they encountered was the scarce data availability especially for hourly rainfall data and DEMs with a high spatial resolution. They found that most urban floods were caused by interplay of heavy rainfall and the elevation of the cities. Cities in the north and west coastal areas of Europe had a smaller percentage flooded than Mediterranean or continental ones.

With a focus on pluvial flood in correlation to urban development Skougaard Kaspersen et al. (2017) selected four cities in Europe for their study "Comparison of the impacts of urban development and climate change on exposing European cities to pluvial flooding". Their choice fell on Vienna, Nice, Strasbourg and Odense, which show a quite different setting. They used a combined remote-sensing and flood-modelling approach to simulate the occurrence and extent of flooding. With the Horton's infiltration model and the Overland flow model MIKE 21 for the run-off and infiltration models, they got the result that for every 1% of absolute imperviousness 0 to 10% higher flooding is to be expected.

About ten years earlier in Britain the "Flooding from Other Sources" project was founded to tackle the flooding from other sources than sea or rivers. The two following studies were a result of this project.

A study by Hankin et al. (2008) was carried out with the aim of finding methods to improve pluvial flood mapping in urban areas. They reviewed available methods and difficulties in flood modelling. E.g. if an urban potential flood hazard map is needed and only limited financial resources are available, data with lower accuracy has to be used. They analysed four different approaches: Creating a buffer around historical flooding data, topographic analysis of LiDAR data, routing of blanket rainfall over a digital elevation model, and various levels of integration of sewer/drainage network models with other sources and pathways such as roads and small watercourses. The noticeable approach is the third one but the most important part for any of the techniques described in this study is accurate topographic data.

Falconer et al. (2009) summarized technical possibilities of providing warning systems for pluvial floods in urban areas based on the RF5 ("Feasibility study into expanding flood warning to cover other flood risks") project. According to conclusions from the RF5 project it is technically feasible to provide some warning service given planned improvements in the UK Met Office radar network. While the Pluvial Extreme Event Planning system (PEEPs) approach, which is supposed to map potentially vulnerable areas, doesn't have a stochastic element it still has many advantages over other methods like low costs or a broad indication of risk. For the PEEP only the topography was used to identify vulnerable areas but it was considered as sufficient as the sewer systems capacities were exceeded in such heavy rain events. Together with the other approaches highlighted in their study they should be able to assist organisations in dealing with different types of flooding.

Switzerland also identified pluvial floods (here integrated into surface water floods) as threat to society, which is pointed out in the following three studies:

Bernet, Prasuhn, and Weingartner (2017) aimed with their study to provide a method to differentiate between surface water floods and fluvial floods and assess the relevance of surface water floods in context of damages. The first goal was approached by assessing the distance of damage claims to rivers or lakes with respect to known fluvial flood zones. The main disadvantage of that method is that surface water floods, which eventually reached a watercourse, were partially classified as fluvial (a possibly more accurate manual classification proved to be too time-consuming and difficult). With this classification the study concluded that surface water floods caused as much damage claims as fluvial floods, but only one quarter of the total loss. The lower damages can be attributed to the lower water depth from surface water floods and the fact that the calculated damages stem only from buildings.

Bernet et al. (2018) used several hydrodynamic models to predict surface water floods in rural areas. They used a binary response variable (wet or dry) and a contingency table to assess the model performance. However, due to biased predictions about effective precipitation and insufficient representation of topographic structures, their grid-based models were not able to predict flooded areas due to surface water flooding. Instead of improving the addressed shortcomings, they recommended the communication and quantification of the uncertainties of the model. Finally, they recommend using a standardized method for the documentation and reporting of surface water floods.

With four test cases, Zischg et al. (2018) compared two 2D inundation models to each other: one is based on insurance claims and the other on observed inundation areas. For model validation they also used validation metrics based on a contingency table. Both models produced similar results, but the model based on insurance claims was better concerning areas with high densities of values at risk. Nevertheless, insurance data has to be carefully pre-processed, for example surface water floods and groundwater floods should be filtered out, which was not done for that study. Another advantage of insurance data is that they also covers small events and are more consistent over time. Whether insurance data is able to reconstruct flood areas of past events remains an open question. The main limitation for that approach, according to this study, remains in the limited data availability due to privacy protection.

As a fellow study from Austria, Zahnt, Eder, and Habersack (2018) wrote a paper about challenges through pluvial floods titled “Herausforderungen durch pluviale Überflutungen - Grundlagen, Schäden und Lösungsansätze“. Their research area is in four communities in Styria, Austria, where they researched why and how damages occurred and how the citizens protect themselves. For this purpose, the flooding season of 2016 was analysed, events documented and affected persons interviewed. In the end, they concluded that there is not enough knowledge and education about the risk of pluvial floods and there is a high need for action.

The most similar study in comparison to ours is called “Assessing urban areas vulnerability to pluvial flooding using GIS applications and Bayesian Belief Network model” from Abebe, Kabir, and Tesfamariam (2018). In their research, they also use a number of possible influencing factors (land cover, population density, slope, soil drainage class, drainage density, DEM, rainfall and drainage capacity) with the goal of obtaining a risk map for Toronto, Canada. As information about flooding events they use basement flooding. They focus their method on the Bayesian Belief Network model which is similar to a statistical method. The model sets conditional probabilities for each variable and their child nodes. The main point in using this approach was that it can quantify uncertainty and consider interdependencies between the variables. They conclude that the most influential factor is population density followed by land cover related parameters and slope.

3. Data

With almost 12,000km² Upper Austria is the fourth biggest federal state of Austria ('Land Oberösterreich - Administrative Gliederung' n.d.). The climate is mostly Central European transitional characterised by oceanic and continental influences. The mean air temperature over the years 1981 to 2010 is 7.6 °C. The average annual precipitation amounts to 1,150 mm, but there are strong regional differences over the study area. In the Bohemian mass it is the lowest with less than 800 mm while in the Alps it is at 1,600 to 1,800 mm ('Land Oberösterreich - Klima in Oberösterreich' n.d.).

The data investigated is a presence/absence raster of pluvial flood damages and 15 different variables with a possible influence on flood damages. From some of these variables, classes have been aggregated for technical reasons. The cell size in the applied overall raster is 1x1km*.

3.1. Flood damage data

The data on the flood damages are based on reports from the Austrian Hail Insurance. There, the agricultural flood damages are defined as *"(…)flood damage near rivers as well as damage caused by heavy rainfall/slope water. In addition, damages fromcontinuous rainfall, followed by silting up and dying of the very young plants after hot subsequent phases. This means that damage is not only reported on hills, but also in flat areas"* (Spira Yvonne 2018). As the data covers damages of agricultural land, flood damages for other land uses, like e.g. settlements is not considered. It is worth noting, that some variables might explain ideal agricultural land rather than pluvial flood vulnerability. The data of the locations of floods were delivered for the years 2007 to 2013 as shapefiles with 100x100 m cells, which contained information about the day of the flood (for 2007 and 2008) or the first and last day a flood occurred in the corresponding year (for 2009 to 2013).

The main information source for the decision whether the reports were based on pluvial or fluvial floods is 'DORIS Atlas 4.0' (2018) where the flood extents of 2002 and 2013 are shown. Additionally, newspapers like "meinbezirk.at" (Uibner Florian 2018) or "Vienna.at" (APA/Red 2013) have repeatedly reported larger damages caused by fluvial floods. It could also be the case that heavy rainfall caused flooding but the water flows into a river. To take this into account, floods at steeper slopes going down into a river were marked as "pluvial". Furthermore connected flood cells that were located along a river were also marked as "fluvial". In the end, 12.392 cells were marked as "pluvial" and cells marked as "fluvial" are no longer used.

The reported pluvial flood events were finely combined and transformed into a presence absence raster (PA) with a resolution of 1x1km². The raster geometry is congruent with soil parameters and INCA precipitation data, as they provide the most variables. Every cell with at least one pluvial damage is coded with a value of one ("presence"), to be marked as a "pluvial flood damage" cell. All other cells are coded as zero ("absence"). There was also the option to use the count of flood reports as values, but due to the difficulties in distinguishing between pluvial and fluvial floods in some areas the weight would have been heavily shifted towards one or two small areas, and there were only few cells with a count greater than one.

3.2. Climate and catchment characteristics

3.2.1. Climate

Rain data was provided by the Central Institution for Meteorology and Geodynamics (Zentralanstalt für Meteorologie und Geodynamik, ZAMG). This data is part of the INCA-Dataset, which has a spatial resolution of 1x1 km² and a temporal resolution of 15 minutes. For the event analysis a time series is extracted for one representative cell and additionally aggregated to one hour and day. For the location analysis and modelling six different precipitation variables are extracted from this data, all representing average climate conditions of the seven-year observation period. We have also tested to use kriging interpolated rain data from the hydrological service network provided by eHYD (https://ehyd.gv.at) but this approach has been discarded because the interpolation errors were much higher compared to the INCA-Dataset and could not further improve the estimates. All rainfall variables of this study are therefore based on the INCA-Dataset.

The first group consists of mean annual maximum precipitation sums of different aggregation periods, including 15min (Max_rain_15), one hour (Max_rain_h), one day (Max_rain_d) and seven-day precipitation sum (Max_rain_7d). We further used the mean absolute deviation of annual 15min precipitation maxima (Max_rain_15_mad) as a measure of inter-annual variability of intense precipitation events. The second group considers the number of days with heavy rainfall intensities, determined by threshold exceedance of the 15min precipitation sums. For our analysis, we used the exceedance of exceedances of 20mm (Rain_20) and 30mm (Rain_30).

The INCA-Dataset stems from a combination of station data, radar data and elevation effects, which have been determined by the data providers using the following procedure: In a first step the rain measures from the stations are interpolated using inverse-distance-squared weights. The radar data, which has a higher spatial accuracy, is then aggregated to 15 minutes rain sums and rescaled using the station data at their positions. Then these two grids are combined with different weights depending on the radar return (cells, which are shielded by e.g. mountains get a lower weight). In the last step the elevation dependence is calculated by the interpolated topography of the stations and the interpolated precipitation in the valley. However this is only applied where the radar was ineffective. Unfortunately especially for high intensity rainfalls in summer at the 15min interval the mean relative analysis error is at around 50% (Haiden et al. 2011). For the event analysis this means that the shown amounts of rain can differ significantly from the actual rainfalls. Even variables that count how often a certain threshold is exceeded can become too high or too low.

3.2.2. Topography

The mean altitudes and mean slopes for each cell are calculated in ArcGIS 10.3 using a digital elevation model (DEM) with a spatial resolution of 10x10m and the "Spatial Analyst" toolbox for the altitude and the "3D Analyst" toolbox for the slope. The exposition was also calculated but dropped in the early phases of modelling, because of the missing explanatory power. The altitude and slope on the other hand are more likely to be significant, as they influence the movement of water directly. Additionally, a dataset on macro relief was available from the BORIS dataset of the Environment Agency Austria, but was not used due to the low spatial resolution. Adding it to our final dataset would have led to reduction of around 1000 grid cells of our dataset, which was not tolerable for a single variable.

3.2.3. Land use

For the land use the CORINE Land Cover data set, Version 2012 was used, which is a shapefile with 44 classes represented by their class number. With the "Conversion" toolbox it was converted into a raster, where each cell got the value of the class that covers most of the cell. The data were reclassified into seven main classes: sealed, forest, wetland, water, arable land, grassland and heterogeneous agricultural land. An alternative dataset of sealed land was not used, as preference was given to a more consistent classification as provided by the CORINE dataset. As the damage reports stem from agricultural land only, we expect that the agricultural classes will be more significant that other classes. Hence, the land use is expected to have a large discriminative power in the predictive models.

3.2.4. Soil

Each soil parameter used (except erosion) is taken from the eBOD dataset, which is freely available by the Austrian Federal Research Centre for Forests (BFW) as a grid with a spatial resolution of 1x1km². The dataset consists of 24 attributes but for this analysis only five of them are used. An additional soil variable could have been the soil moisture provided by the Technical University Vienna. However, its values only increased slowly after an intense rainfall and therefore were not a good indicator for soil moisture right before an event. In addition, the general water conditions have already been included in the eBOD data set. Another water-based variable, which was not used, is the groundwater from BORIS. Similar to the macro relief, it has a very low spatial resolution and therefore would have reduced the number of cells significantly. Generally, most of the soil variables might (at least for some part) indicate ideal agricultural land. However, they are expected to determine those agricultural areas with a higher risk of pluvial flood damage than others. This is supported by the observation, that soil types or soil water conditions differ.

Erosion

Soil erosion is understood as the natural process of removing soil through running water or wind (BFW 2013). The data for it was provided by the Austrian Federal Office for Water Management (Petzenkirchen) and was already aggregated to 1x1km² raster cells, which contain the average, yearly erosion in kg/ha. It was modelled with data of soil, precipitation, slope inclination, slope length, land cover and soil protection (BMLFUW 2007).

Soil type

The soil types were formed after many years of the exposure to climate, vegetation, humans, water and wind, among other things (BFW 2013). In the eBOD dataset the soil types were split into 38 categories which were reclassified to their nine main types of relict soil, alluvial soil, gley, atypical soil, brown soil, pseudogley, soil form complex, rendzina and rangier, bog and others. In the category of others are included podsol, black soil and raw soil. Bog, alluvial soil and gley are in groundwater area while rendzina and rangier, brown soil, podsol, pseudogley and relict soil are not. Gley and pseudogley are very similar as they both have solid bedrock which hinders water to seep away. The biggest difference between the two is that gley is under the influence of groundwater and pseudogley stores especially rainwater for a longer period. A soil form complex is categorized if many soil types are located in a single cell and therefore no unambiguous categorization is possible (BFW 2013).

Soil texture

The soil texture describes the particle size composition. There are three main soil separates: clay (particles <0.002mm), silt (0.002-0.06mm) and sand (0.06-2.0mm). Additionally there is one more texture called loam, which consists of relatively high proportions of all three size groups. The soil texture determines the behaviour of the soil in relation to water, its aeration, its nutrient supply and fixation and much more. Clay soils are densely packed and often collect water, while sandy soils are the exact opposite. Silt soils possess characteristics of clay and sand but are more a combination of the undesired characteristics than the desired (BFW 2013). The soil textures were subdivided into 12 different sub classes, which were reduced to the main four of silt, sand, clay, loam and cells with no unambiguous assignment.

Permeability

The soil permeability indicated how good the ground can hold water. High values mean that water will go through very quickly and low values that it will take a longer time to pass through (BFW 2013). The permeability was divided into ten classes, which were reordered so they can be used as continuous variable going from low to high.

Soil depth

The soil depth is the zone between the soil surface and the solid rock or extremely hardened horizon. In general the soil depth is classified into shallow (<30cm), medium (30-70cm) and deep soils (>70cm) but here we have 6 classes where 1 is strongly fluctuating and 2 – shallow going up to 6 – deep. With increasing soil depth more water can be stored, there are more nutrients and roots have more room to grow (BFW 2013).

Water conditions

The water condition of the soil depends on various factors, including precipitation and groundwater conditions, soil type, humus conditions. They are mostly classified from very dry to wet. In addition, there are soils that are fluctuating so much that they don't fit into these classes. They are referred to as alternating moisture, which can have more dry or wet phases (BFW 2013). For the analysis water conditions were resampled to 1 – inconsistent, 2 – alternating moisture with more dry phases to 4 – alternating with more wet phases and then 5 – dry to 17 wet. Even though being ordered, they can't be categorized as a continuous variable, because of the alternating moisture class.

3.3. Dataset structure

The final data frame is summarized in Table 1. It consists of 9405 grid cells vs. 15 predictor variables and one dichotomous dependent variable (PA), which separates the dataset into 8136 cells without flood reports and 1269 cells where pluvial floods have been reported.

Table 1: Summary of variables used in statistical modelling (for factor levels see Table 8)

Variable	Data type	Description	Unit
PA	Logical	Presence/absence of reported flood	0/1
Max_rain_15	Metric	Mean maximum 15 min rain sum	mm (or l/m²)
Max_rain_15_mad	Metric	Mean absolute deviation (MAD) of annual Max_rain_15	mm (or l/m²)
Max_rain_h	Metric	Mean maximum 1 hour rain sum	mm (or l/m²)
Max_rain_d	Metric	Mean maximum 1 day rain sum	mm (or l/m²)
Rain_20	Metric	Number of 15 min rain sums >20l/m²	1
Rain_30	Metric	Number of 15 min rain sums >30l/m²	1
Altitude	Metric	Mean altitude	m (or m.a.s.l.)
Slope	Metric	Mean slope	%
Erosion	Metric	Mean erosion	kg/ha
Land.use	Factor	Land use	7 classes
Soil.text	Factor	Soil texture	5 classes
Soil.type	Factor	Soil type	10 classes
Permeab.	Ordinal	Soil permeability	10 levels
Depth	Metric	Soil depth	

Water.con	Factor	Water condition	17 classes

4. Methods

4.1. Event analysis

Before going into more detailed analysis, it has to be determined which event should be used. By plotting all reports per year and month we get a general view when large events usually happened. In addition, we know in which years and months large events have taken place. After choosing one large event per year (in case a large event happened each year) they are further analysed and compared. First the rainfall in the whole study area is investigated, using a 26x26km window to represent the current event. From a selected cell three time series are extracted for a temporal window of five days (four days before the event to one day after), except the event of 2013 which is extended to four days after. For this cell, these time series show for each day the total rainfall sum and the maximum rain intensity reached in one hour and in 15 minutes. These are compared with the return periods provided by eHYD, which were calculated at the end of 2008. The primary purpose of the time series is to find out whether the reported damages were caused by a long lasting rainfall or a short but heavy one. It is also important to know if the soils were moist from rainfall in the previous days or dry.

4.2. Location analysis

The location analysis serves as a first visual assessment whether some location parameters might influence the risk for pluvial floods. As a continuation of the event analysis, the reported cells of each event window are compared to each other. This serves to answer the question of how the cells look like that got flooded in a specific event and if there are similarities between the different locations. This is then further expanded over the whole study area where the cells are separated in reports and no reports. Additionally the distribution of each location parameter over the study area is shown and analysed. At the end of this analysis comes a short summary where it is determined which classes will be excluded in the statistical models with aggregated classes.

4.3. Statistical modelling

Two statistical approaches are tested for the prediction: the linear logistic regression and the nonlinear random forest. Both of them are able to deal with a binary response variable and with continuous and factor explanatory variables. These two will be trained with two data sets and compared to each other and the other model using model quality assessment.

The first of the two tested data sets contains all classes from all variables; the second data set uses aggregated variable classes that have been determined during the location analysis to give a parsimonious description of the study area, which are determined at the end of the location analysis. For Chen, Liaw, and Breiman (2004) one way to deal with imbalanced data is to reduce the number of cells without reported damage, which was tested in a third model. There, a buffer of 1km was drawn around each cell with reported damage and only the cells inside this buffer are used for modelling. However, the model was analysed at first, but at the end it was dismissed because of lack of success.

4.3.1. Logistic regression

Usually, in linear regression models the response variable is continuous, but in this case this variable is a factor with two possible outcomes (0, 1). So, the first relevant difference is that the outcome variable can’t be bigger than 1 or smaller than 0 and represents the predicted probability. Another difference is that the estimates of the independent variables aren’t as easy to interpret as in the normal model. In the logit model they are transformed into log-odds ratios, which may be retransformed to odds-ratios after model fitting for easier interpretation of the results. The resulting values bigger than 1 are interpreted as increasing odds of presence with an increase of the variable and values between 0 and 1 decreasing odds (Hosmer and Lemeshow 2000). The variables do not need to be normalized as this is not an assumption for logistic regressions. An advantage of logistic regression is that weights can be added to deal with the imbalanced data. Here we chose weights of 1:7, because the imbalance of the data is roughly 7:1.

The analysis proceeded as follows: In an initial modelling step, each variable is separately tested in a single model to determine its individual value for explaining pluvial flood damage. At the same time a heatmap is used to identify correlations between the variables. For each single model three pseudo R^2-values are calculated, which indicate the explanatory power of a model among similar models (Mangiafico 2019). The calculated R^2 are from Nagelkerke, Cox and Snell and McFadden, with McFadden's index being often preferred for logistic regressions. It has usually lower values than the other R^2 but values between 0.2 and 0.4 are already considered really well. For the following forward stepwise regressions the variable with the highest R^2 is used in the first step.

After splitting the data into a training set and test set (80:20), variables are added to the model based on Akaike's "An Information Criterion" (AIC), where a lower AIC indicates a more parsimonious model. Variables that do not change the AIC or even increase it are not added to the model. That way, if two variables are highly correlated, one of them is left out, because it does not contribute much more to the model. Additionally, interaction terms can be added to account for the correlations between the variables. The resulting model is checked for multicollinearity using the generalized variation inflation factor (GVIF). If multicollinearity can be detected, the variable with the highest GVIF has to be excluded. This process is repeated until no more multicollinearity can be detected. At the end, a non-sequential analysis of variance ("Anova"-function of the "car"-package) is carried out that tests how much each variable contributes to the model (after all the others) and if they are significant to the model with a likelihood ratio chi-square test.

The final two models are used to predict the pluvial flood damage cells of the test set. The calculated prediction quality values are explained in chapter 4.3.3 and the predicted probabilities are then plotted over the whole study area.

4.3.2. Random forest

A Random forest (Breiman 2001) is an ensemble of decision trees, where each tree is randomized to a certain degree. Each tree is grown using a different bootstrap sample of the training data and the data that is not used is the test data. That way each data point would be in the test data around 36% of the times. For additional randomness at each split of the trees the best variable of only a subset of variables is chosen. With this method random forests perform very well in comparison to other machine learning methods like support vector machines and is robust against overfitting (Liaw and Wiener 2002). At the end each tree casts a vote for their most popular class and the class with the most votes is used by the random forest (Liaw and Wiener 2002; Wang et al. 2015).

For the random forests there are no weights added, instead a cut-off is chosen similar to the one for the GLM. As described above the classification forest uses votes to determine which class gets selected. The "cutoff" specification of the "randomForest" function would default to 0.5, which means that the class that has the majority of votes "wins". Here different cut-offs are tested where each gives the minority class (cells with reports) an advantage of different scales. When the type "response" is chosen at the prediction of the test set the model predicts a class for each data point. Choosing the type "prob" predicts class probabilities for each cell similar to the GLM.

In this study the R-package "randomForest" (Breiman et al. 2018) is used. It is used in the form of a classification forest, which then predicts the binary response variable. Even though the random forest of the "ranger"-package can be used with weights, which could be beneficial for the model, the cut-off cannot be chosen and for our study the cut-off is chosen to be more important.

For the random forests, there are no single-variable models that would make sense and there is no pseudo R^2 for classification forests. The relationships between the variables still require attention. Although they do not particularly harm the model, they have to be kept in mind in the interpretation of the model effects.

At the model optimization the variable significances are interpreted and checked for plausibility. In case of spurious effects the respective variables are excluded from the model. The split into a training and test set is not needed here, because of the bagging method of the random forest. For the predictions, all the pluvial floods have to be predicted. Other than that the same steps as for the logistic regression can be used, only the probabilities have to be predicted separately.

4.3.3. Model quality assessment

The basis for quality assessment is a confusion matrix, which shows how many data points of the test set are predicted correctly and incorrectly. From this matrix the following performance metrics can be obtained:

- Accuracy

The accuracy is defined as the percentage of correct predictions. It is the total number of true positive (flood damage cells) and true negative (no-flood damage cells) predictions divided by the number of all predictions (in our case the total number of cells within the study area)

- Sensitivity

The sensitivity is calculated by dividing the true positive predictions (in our case cells that got correctly predicted as flood damage cells) divided by the number of all positive predictions.

- Specificity

The specificity is the same as the sensitivity but with negatives. So for our case it is the number of correctly predicted no-flood damage cells divided by the total number of no-flood damage predictions.

- Balanced accuracy or area under curve (AUC)

Balanced accuracy or AUC is calculated as the sum of sensitivity and specificity divided by 2. The AUC refers to the area under the curve of the receiver operating characteristic (ROC), which is an easy way of showing how good a model performs. The benefit of using the balanced accuracy rather than the (simple) accuracy is that the latter is not a reliable metric for the real performance of a classifier for unbalanced data sets (imbalance of damaged/undamaged cells in our case).

- False alarm rate (FAR)

The FAR characterizes the percentage of false flood damage predictions. It is the number of false positive predictions divided by the total number of positive observations (flood-damage-cells). A good classified predictor should have a good balance between a high number of true alarms (sensitivity) and a low number of false alarms (FAR).

5. Event Analysis

The event analysis focuses on the largest events in terms of damaged cells. The question in which year and in which months these events took place is answered in Figure 1. What is immediately noticeable is that most pluvial flood damages were reported in August (4340), with the majority having taken place in a single year (2623 in 2008). With 1014 reported cells less the second most severe month is June. This time it consists mostly of two events, one in 2009 (1083) and one in 2013 (1777). In July 2112 cells got marked as pluvial flood, where most cells come from an event of 2009 (1415). Other reported damages refer to April, May and September, therewith covering the relevant vegetation period in the agricultural sector.

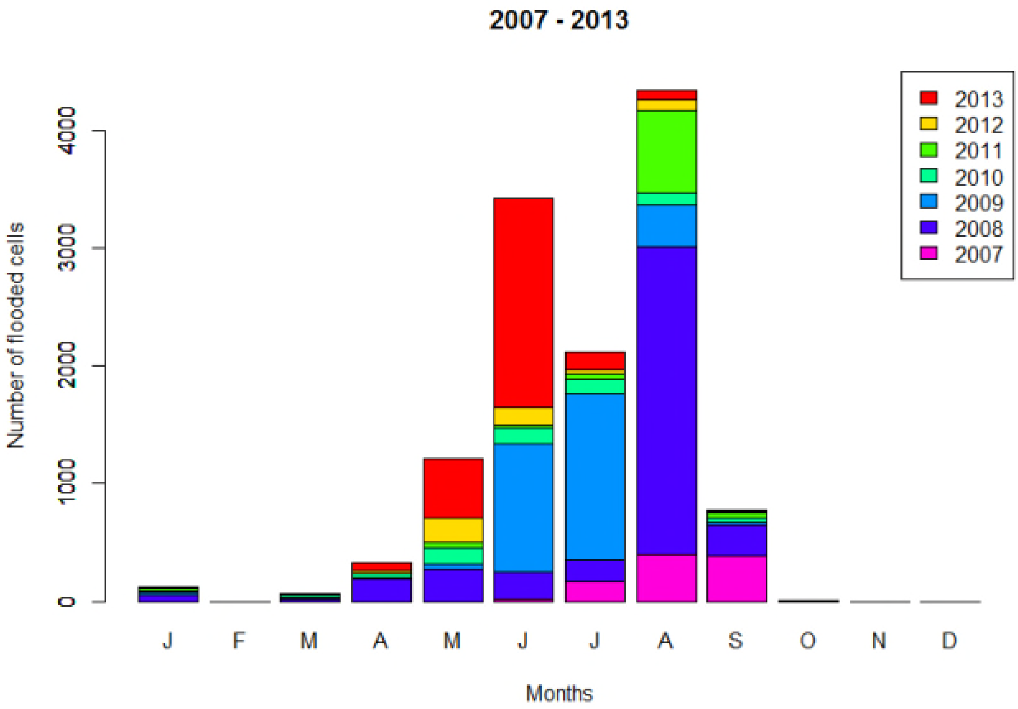


Figure 1: Number of reported cells per year and month

For comparison, mean annual monthly precipitation of 2007 to 2013 is illustrated in Figure 2. June is the month, where it rained the most in our study area, followed by July and May. The years that contributed most to the rain in June are 2009 and 2013 with 268l/m² and 218l/m² respectively (mean value for this month is 159l/m²). However, especially in August 2008, where over 2000 cells were reported and marked as pluvial flood damages, the precipitation added up to 120l/m² only and was not exceptional (mean value for this month is 116l/m²).

The highest annual precipitation can be found in the Alps with 1000-2500mm, in the lower altitudes it is between 700-1000mm (BMLFUW 2007). The lowest mean annual precipitation of our study area (which excludes most of the Alps) is in the years of 2008 (995mm) and 2011 (894mm), the highest in 2007 (1231mm) and 2009 (1231mm).

The highest 15 minute precipitation sum was observed in June 2011 with 69.9l/m² (Figure 3). When the highest intensities of each month are summed up, July is first with 345l/m², followed by June with 315l/m² and August with 310l/m². Overall, the years of 2009 and 2010 stand out with very low rain intensities and 2011 with relatively high values compared to the mean monthly precipitation. The years of 2007 and 2008 also stand out with high intensities and low sums.

Comparing Figures 1 and 2 shows that the reported flood damages do not directly correlate to the monthly mean rain sums in the respective years. There is some correlation with monthly rain intensity patterns in Figure 3, but again, the correlations between individual years are rather small. For a better understanding why pluvial flood damages occurred, the five biggest events are further analysed.

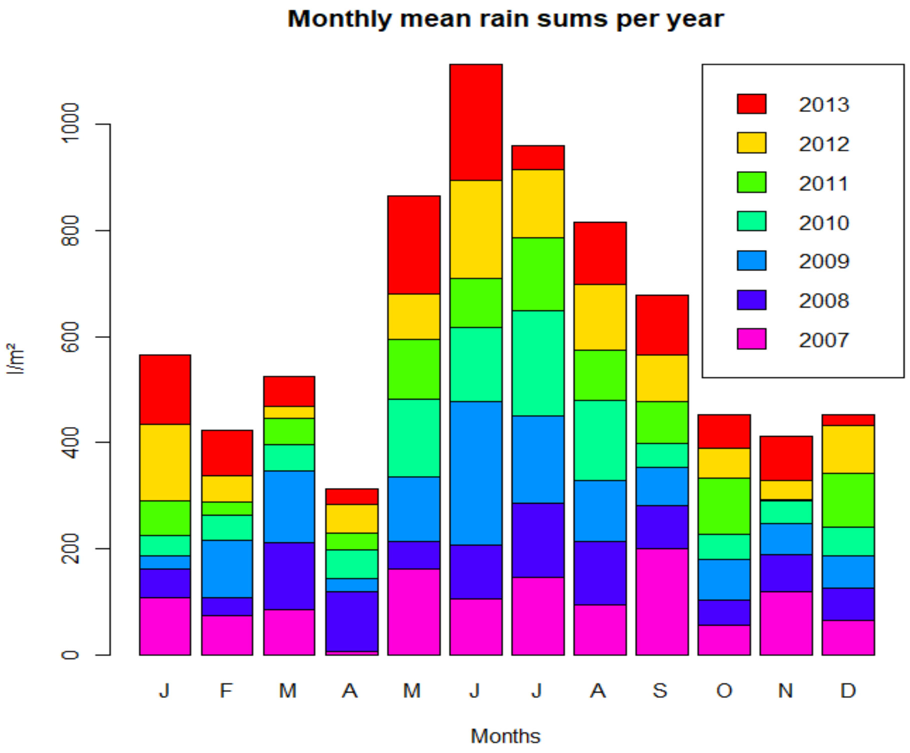


Figure 2: Mean annual monthly precipitation over the 2007-2013 period

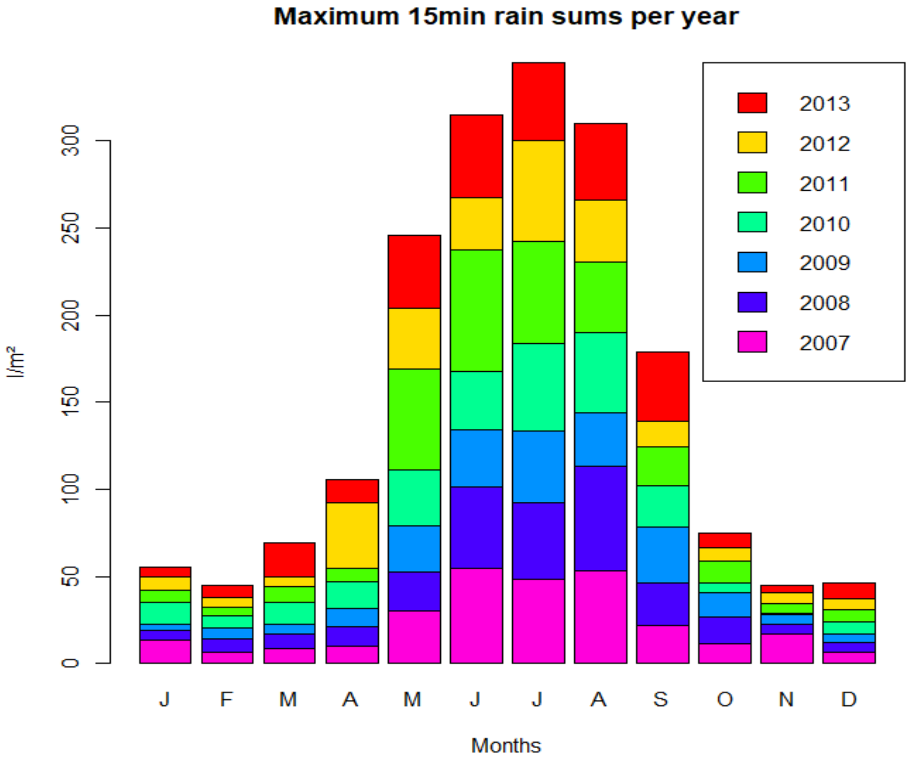


Figure 3: Maximum annual 15 minute precipitation over the 2007-2013 period

5.1. Event of 19.08.2007

The event of the 19th of August 2007 mostly led to reports in the districts of Linz-Land and Steyr-Land, more exactly south of Linz and west of Steyr (Figure 4). The rainfall of that day reached up to 60l/m², while most reported cells were under a direct rainfall of 30-45l/m². However, most of the rain fell with 75-90l/m² in the centre of the district Kirchdorf at the alpine border, which did not cause any flood reports.

In the time series of the centre cell of the 26x26km square (Figure 5), the rainfall on the day of the event reached 53l/m², which corresponds to a return period of just one year. However over half of this amount (28.4l/m²) fell in just 15 minutes and in the other 45 minutes of the hour only 2.6l/m² were added. The return periods for them are 30-75 years and 5-10 years respectively. The remaining amount of rain, which is 22l/m², fell during the other 23 hours of that day. Two days before there was a small raining period over a few hours that summed up to 7.8l/m². Even though the day before the event there has been no rainfall, it is not certain that the soil was very dry. The day of the event was not followed by any rain the day after. In the end, at the day of the event 279 cells were reported to be flooded and 7 more on the following day.

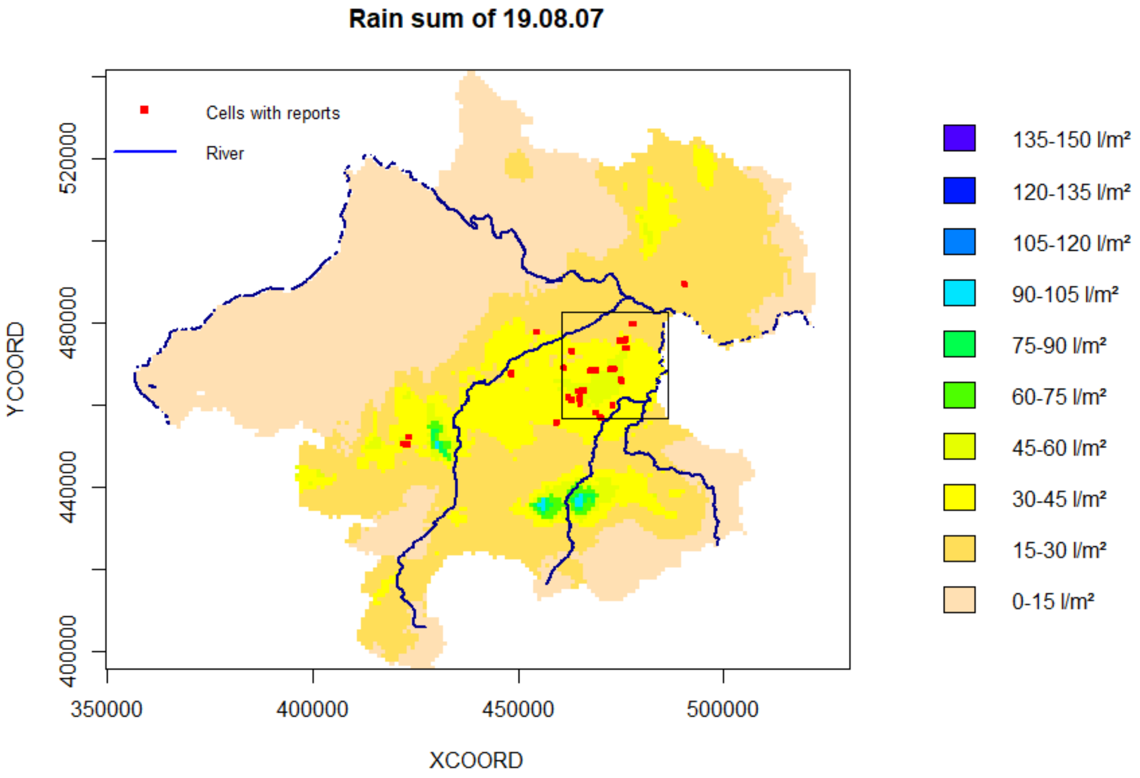


Figure 4: Spatial distribution of precipitation (daily rain sum) at the 19.08.2007

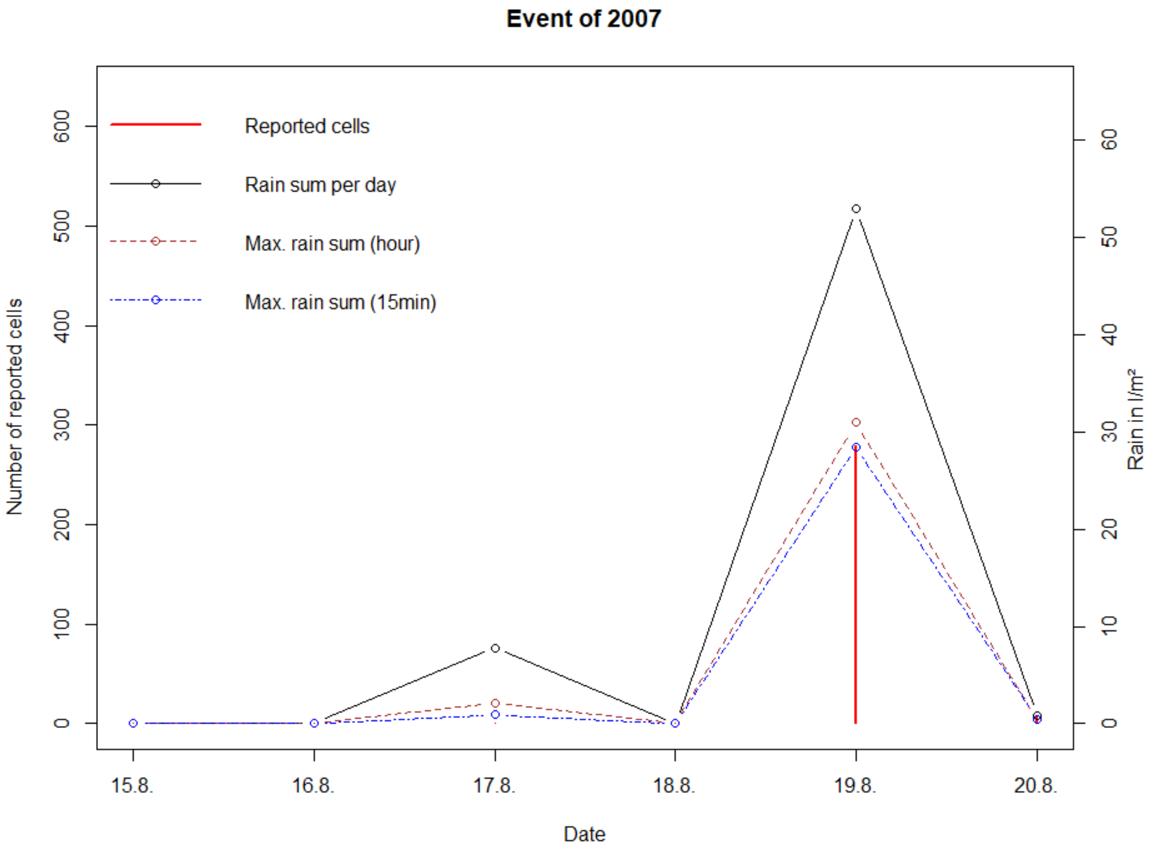


Figure 5: Temporal development of the event of 2007: Precipitation characteristics and reported flooding

5.2. Event of 22.08.2008

The event of 2008 caused the most reports during the investigated time period of this study (Figure 6). The precipitation of this day and the resulting pluvial floods stretched from the district of Ried to Urfahr-Umgebung. It had a smaller spatial extent compared to the one of 2007 but reached up to 90l/m² in one day. However there are also many reported cells where the amount of rain was lower than 30 or even 15l/m², for example the ones in Linz-Land.

Represented by the cell in the centre of the 26x26km window, Figure 7 shows the time series of the rain that caused the flood reports. With 44.5l/m² most of the rain at that day fell in a time span of 15 minutes, which even exceeds a 100-year event. The rest of the day contributed only 15.5l/m² on top, which occurs every one to two years. This short and very intense rainfall was followed by 635 reports of floods in this 26x26km square. Before that the soil was very dry as there was almost no rainfall in the previous four days. It also was not followed by more rain the day after. Why there are 26 reported floods the day before the event remains unclear.

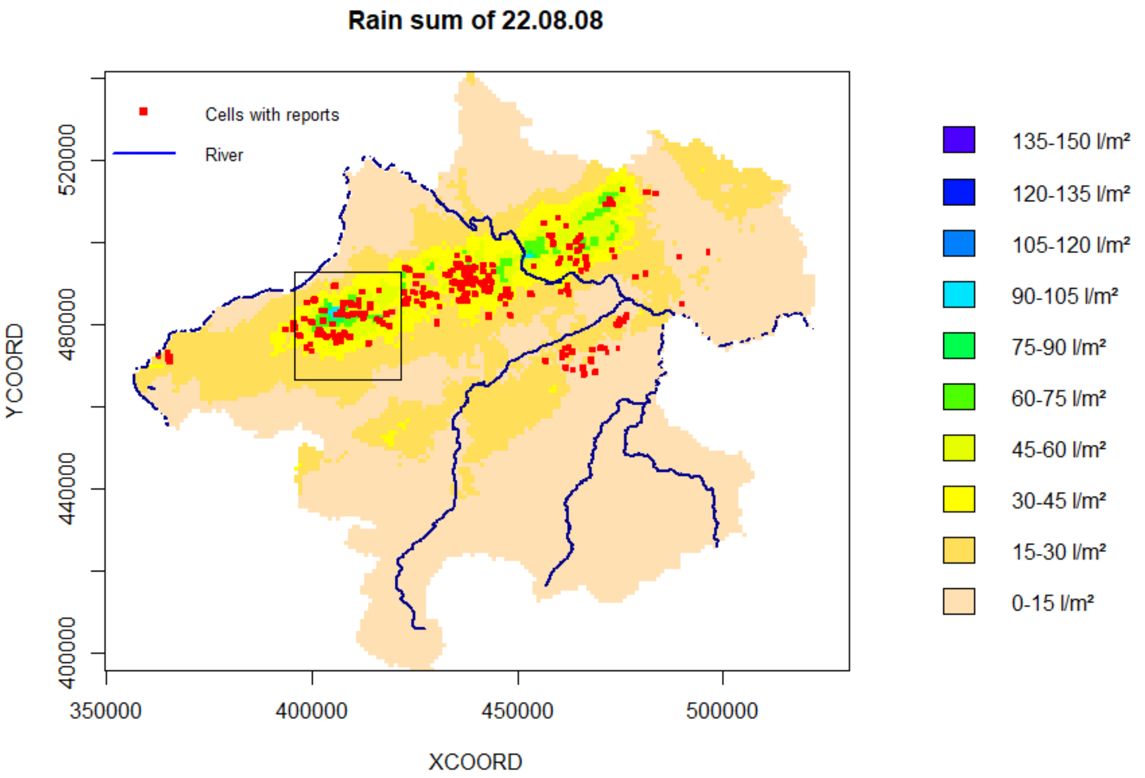


Figure 6: Spatial distribution of precipitation (daily rain sum) at the 22.08.08

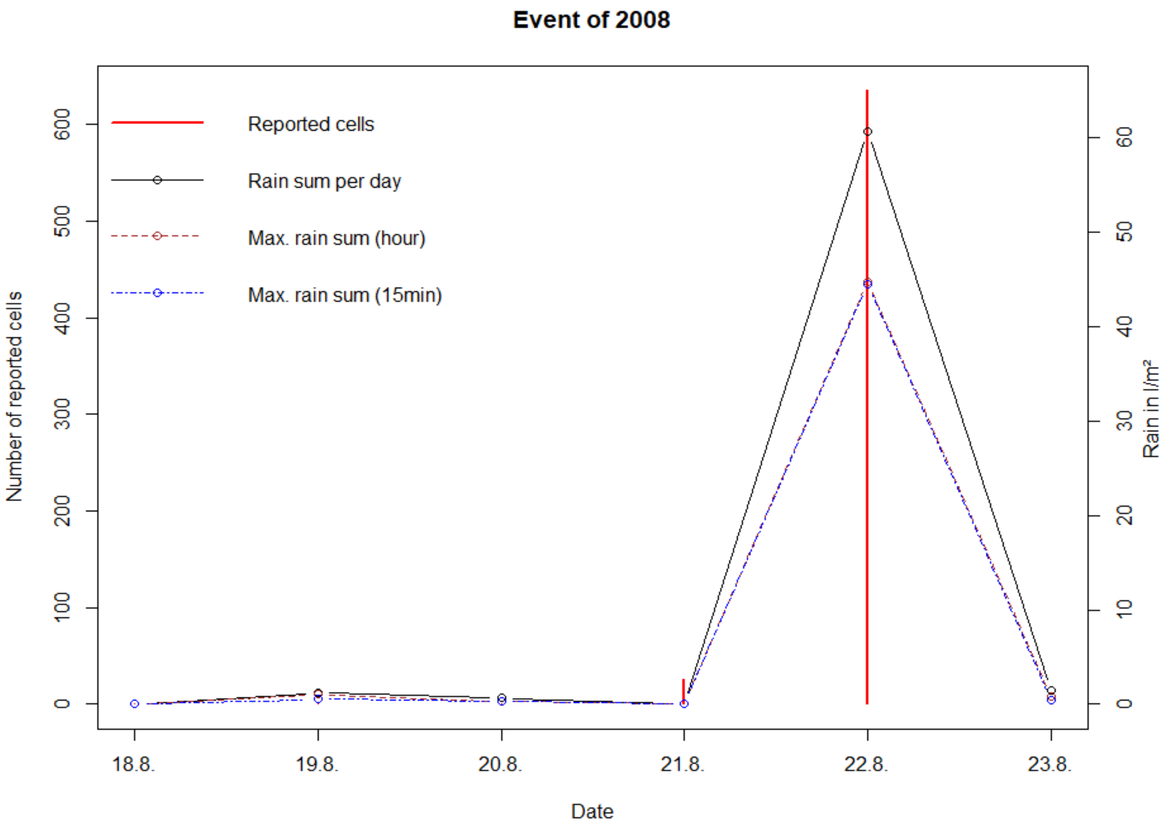


Figure 7: Temporal development of the event of 2008: Precipitation characteristics and reported flooding

5.3. Event of 07.07.2009

The event of the 7th of July 2009, shown in Figure 8, looks similar to the one of 2007. Most of the area experienced a rainfall of 30-45l/m² which has caused the reports and a few cells reached to the class of 60-75l/m² but they didn't lead to any reports (at least not directly). This time the affected district is Vöcklabruck which is located at the Alpine border.

This time, the precipitation time series does not represent the cell in the centre, but more to the south in the largest cluster of reported cells. Compared to the events shown before, the time series of the event of 2009 looks rather different (Figure 9). The rain sum of the day of the event almost reaches 48l/m² and half of it (23.9l/m²) fell in just 15 minutes but this time the soil was already moist from the days before. The return periods for these are less than one year and 3-5 years respectively. Although these rainfalls were not as long lasting as it seems because around half of it fell in 15 minutes just like the main event, only in a smaller scale between 1.2 and 12.7l/m². Even concerning these rainfalls, the values for the return period of one year are still far higher. These increasing heavy rainfalls over five days resulted in 304 reports of flood damaged cells.

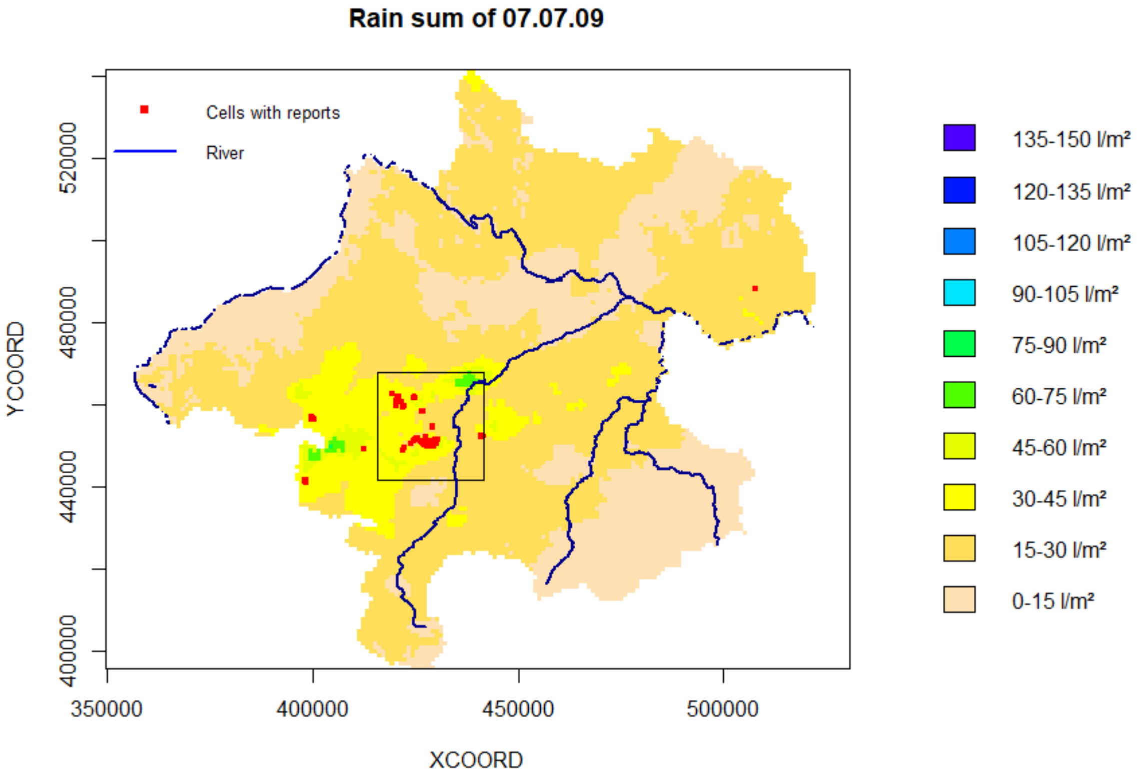


Figure 8: Spatial distribution of precipitation (daily rain sum) at the 07.07.09

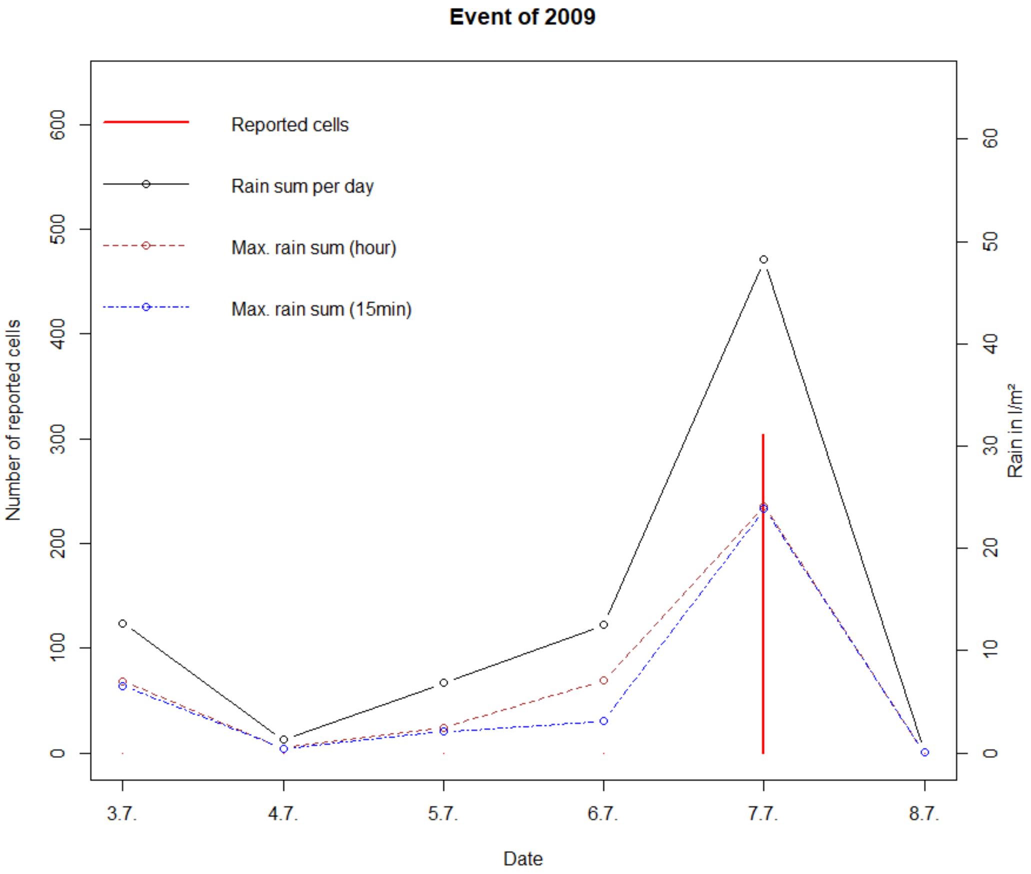


Figure 9: Temporal development of the event of 2009: Precipitation characteristics and reported flooding

5.4. Event of 24.08.2011

On the 24th of August 2011 there were actually two smaller events that caused damage reports at the same time (Figure 10). With only very few exceptions all the reported cells were located in the districts of Ried and Eferding under a rainfall of mostly 30-45l/m². These regions were already damaged at the event of 2008 but fortunately this time the rainfall was lower. For the time series analysis below the event in Ried is chosen because it covered a larger area.

The precipitation of the event of 2011 is represented by the centre cell of the window. It was the smallest in the amount of rainfall, compared to the others shown, but nevertheless important as also caused 263 reports of flood damages (Figure 11). Before the event the soil was rather dry because there was no rainfall the four days before. At the 24th of August the rainfall reached up to 33l/m², but this time the most intense 15 minute period “only” contributed 12l/m² to it. The return period for the daily sum is lower than one year and for the 15 minute sum one to two years. The sum of the most intense one hour period (16l/m²) is about half of the amount of the day, which also has a return period of less than one year. The other 17l/m² fell during the rest of the day. Another difference, compared to the events before, is that there were also flood damages reported the day after the event, when almost no rain fell. These could be the result of the late hours of these rains (9 p.m. and 10 p.m.), yielding that some of the flood damages were detected on the following day.

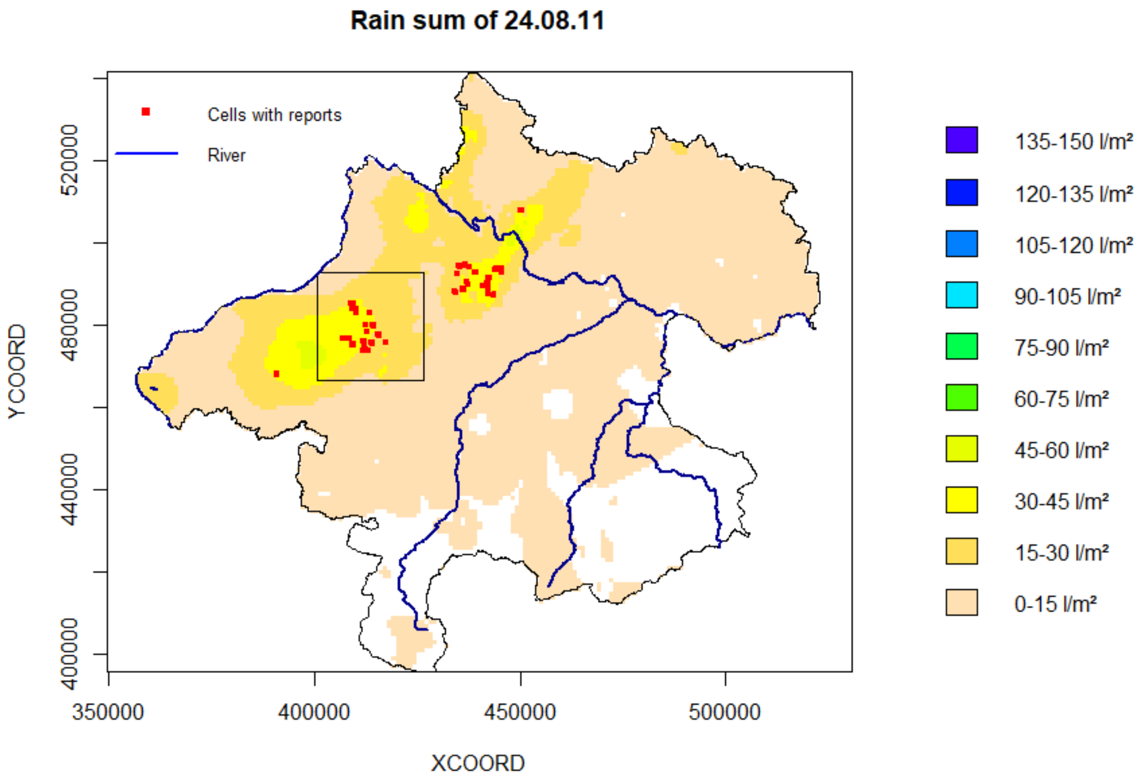


Figure 10: Spatial distribution of precipitation (daily rain sum) at the 24.08.11

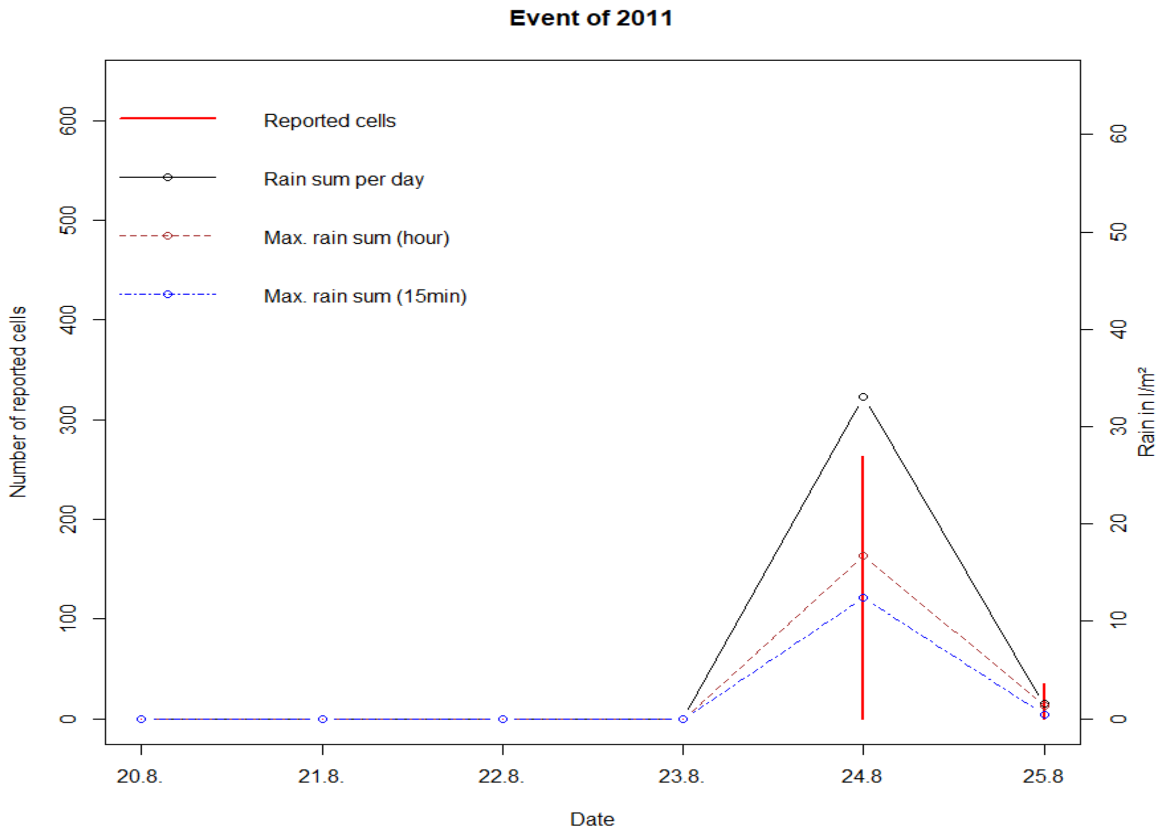


Figure 11: Temporal development of the event of 2011: Precipitation characteristics and reported flooding

5.5. Event of 02.06.2013

The event of the 2nd of June 2013 is unique compared to the others in several ways. First of all, it was raining in basically all of Upper Austria (Figure 12). The amount of rain over the regions north of the Alps was at around 30-60l/m², the same amount that caused many reports before. This time only a few flood damages got reported in comparison to the size of the other events. The biggest cluster of reported cells is in a basin in the district of Perg close to the Danube, where the maximum rain sum was up to 75l/m². The Eferding basin, which lies north-west of this, was also flooded but it was marked as fluvial flood due to many reports of the overflowing Danube in this region. In the Alpine part of the study area the amount of daily precipitation was more than 100l/m². These amounts however did not cause any direct damages to agricultural land.

The time series of the cell in the middle of the 2013 window looks very different compared to all the others before (Figure 13). There was no sudden, heavy rainfall, which caused many damaged cells. This was an extremely persistent rainfall that lasted up to a whole week, where over 19-56l/m² rain fell per day resulting in an event with a return period of 30-75 years. The highest rain intensities per hour and day were at around 2.5-6l/m² and the 15min intensities not even half of them (0.9-2.4l/m²), which is far less than a one-year event for this region. This continuous rainfall caused up to 288 reports of damaged cells per day, which summed up to 750 reported cells in six days resulting in the most severe event for the selected areas. Another interesting point is that most of the damages got reported, when the rainfall was declining. Some of the damages even got reported after the end of the rainfall. There are other events that follow a similar precipitation pattern, but they were not further analysed due to a small amount of pluvial flood reports. The only event that is relevant and comes close to the one of 2013 is the one of 2009.

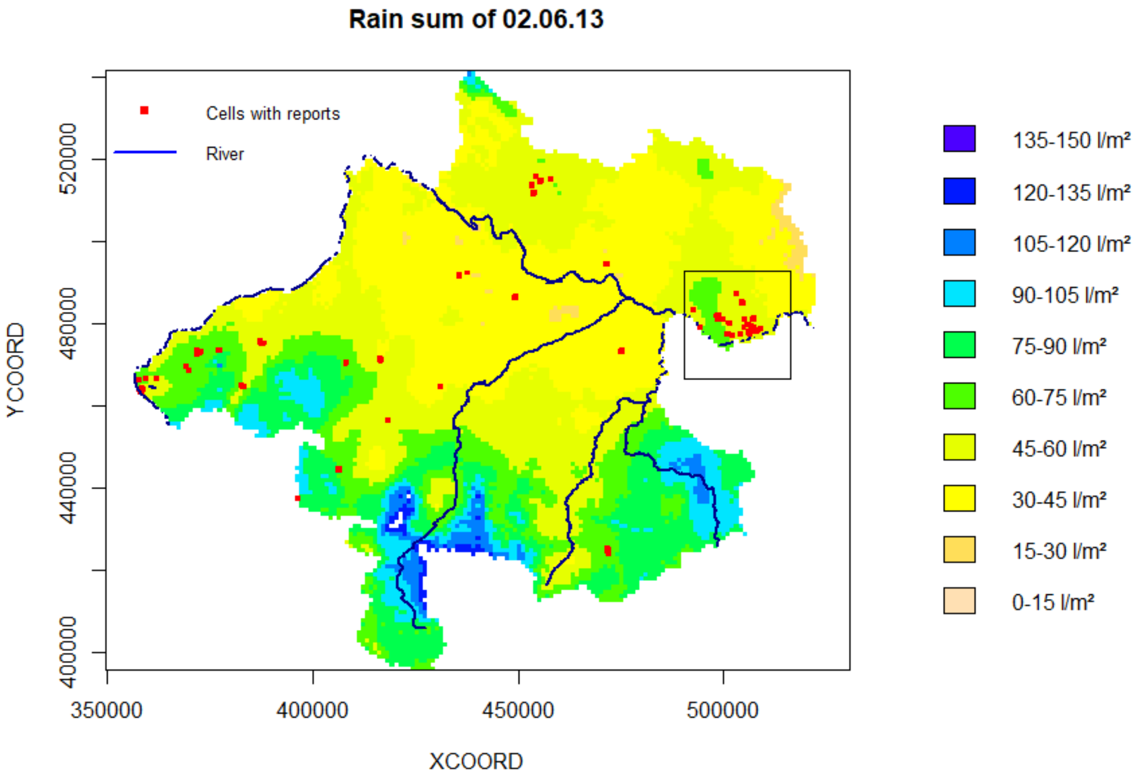


Figure 12: Spatial distribution of precipitation (daily rain sum) at the 02.06.13

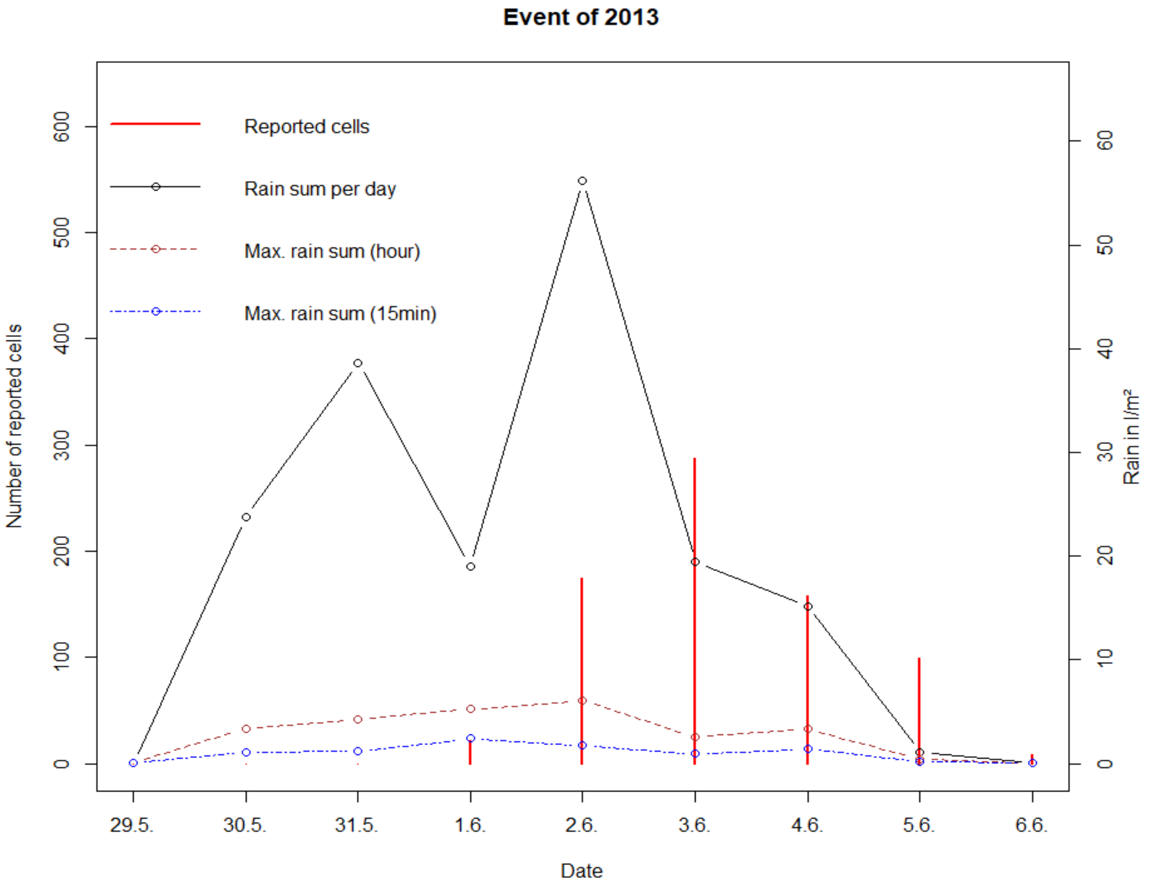


Figure 13: Temporal development of the event of 2013: Precipitation characteristics and reported flooding

6. Location Analysis

Up to this chapter the referenced event cells were based on the 100x100m grids from the original data. From now on, the cells of the 1x1km² presence/absence raster (PA) are used as a reference whether a cell was flooded or not.

In the first part, the locations of the flooded cells of each region, which were highlighted for the time series, are compared to each other in terms of precipitation characteristics. After that, all cells are split into cells with reported pluvial flood damages and cells without reports of pluvial flood damages based on PA.

6.1. Location analysis of flood damaged cells

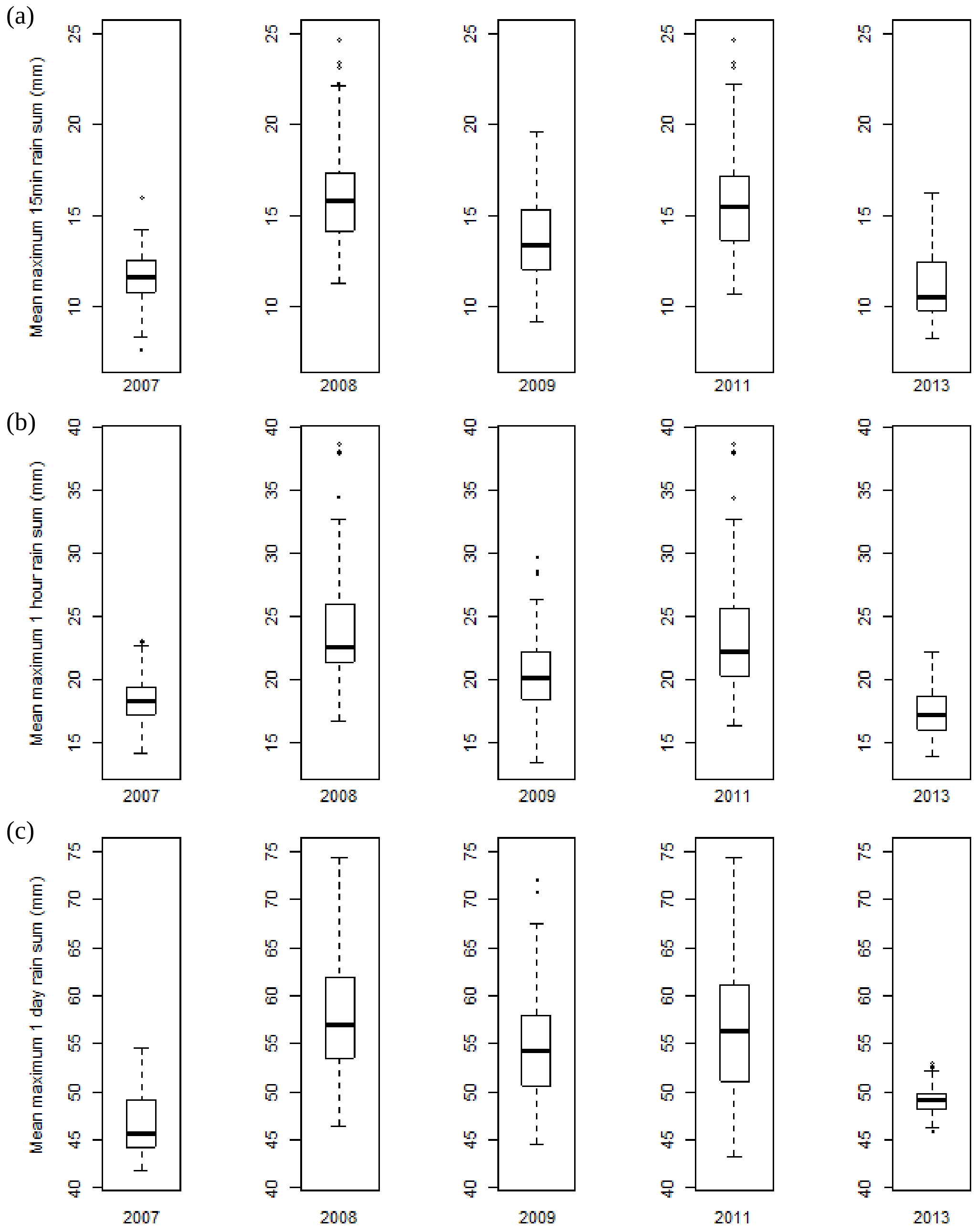
As a starting point, the spatial distributions of peak precipitation characteristics for flood damaged cells are assessed (Figure 14). The distributions of maximum 15 minutes rain sum (Figure 14a) differ substantially between the events. 2008 and 2011 have the highest values with their median at around 15.5l/m². The next highest value is in 2009 with a median of 13.3l/m². The lowest total values have 2007 and 2013 with a median of 11.6l/m² and 10.5l/m² respectively. The maximum rain sums of 1 hour (Figure 14b) and 1 day (Figure 14c) show the same inter-annual patterns as the maximum 15 minutes rain sums, but the values are ca. 7l/m² higher for the hourly values, and ca. 40l/m² higher for the daily aggregates. For the latter, the lowest median is now 45.8l/m² at 2007, and the highest 57l/m² at 2008. Also noticeable is the low variability in the year of 2013, where the values lie between 45 and 55 l/m². Altogether, the variability of peak rainfall increases with the magnitude of events, as it is indicated in Figure 14d by the MAD of the annual maximum 15 minute sums. The lowest MAD (which is a robust estimate of the standard deviation) can be observed in the year of 2013 with a value of 2.7l/m² followed by 2007 with 3.8l/m² and 2009 with 4.3l/m². The highest values can be found again in the years of 2008 and 2011 with 6.4 and 6.1l/m².

Figures 14d - 14f show the number of locations where peak rainfall exceeded some threshold level. The threshold of 20l/m² in 15 minutes was exceeded in the years 2008 and 2011 up to eight times, while their median is two. The next most frequent rain intensities over 20l/m² are found in 2009 with up to five and a median of one. 2007 and 2013 look quite similar but 2007 exceeds the threshold up to three times but its median is at zero. 2013 on the other hand exceeds it 0-2 times with its median being one. Not surprisingly the plots don’t change dramatically for the 30l/m² threshold. 2008 and 2011 are at maximum exceedances of two and a median of one, while the next highest is 2009 with up to three and a median of zero. The remaining two years also have a median of zero but some event of 2013 crossed the 30l/m² mark at least once, which is quite surprising as its mean 15 minute sums are even lower than 20l/m². These high intensity rainfalls in this region might be uncommon compared to the others and may not necessarily lead to pluvial flood damages.

It is now interesting to analyse the locations of reported cells in terms of topographic factors. The reported cells are located in an altitude of 226 to 696 meters above sea level (Figure 15). The damaged cells of the event of 2009 are located highest when we chose the median (bold black line in the box) as reference which is not a surprise as it is the most southern of the event windows. The median in a boxplot shows the values that separates the upper and lower half of the values. For 2009 it lies at 490 meters followed by 2008 and 2011 at around 460 meters. The lowest reported cells were caused by the event of 2007 and 2013 at around 340 meters. The location of the event of 2013 stands out again as it contains the highest and lowest values compared to the others. Nevertheless half of them are between 226 and 336 meters and the other half between 336 and 696 meters.

The box plots of the slopes in Figure 16 show a completely different picture, since most values of 0 to 15 % are quite low, but nevertheless many outliers reach up to 67 %. The medians of all selected years are between 4 and 9 %. 2013 looks different compared to the others because the values of the upper half are more spread and have no outliers. The year with the overall lowest average slopes is 2007 with the lowest median and outliers.

The erosion in Figure 17 shows bigger differences among the years. Basically they can be separated into two groups: 2007, 2008 and 2011 which cover a great range of 0 to over 8 or 10.000 kg/ha and 2009 and 2013 which have very low values but many outliers. It is not surprising that 2008 and 2011 are very similar to each other as they even partly overlap but 2007 is located in the east at lower altitudes and slopes. The biggest difference between 2007 and the other two is the upper limit which lies about 2.000 kg/ha lower. The locations of 2009 and 2013 didn't seem to be very similar until now.



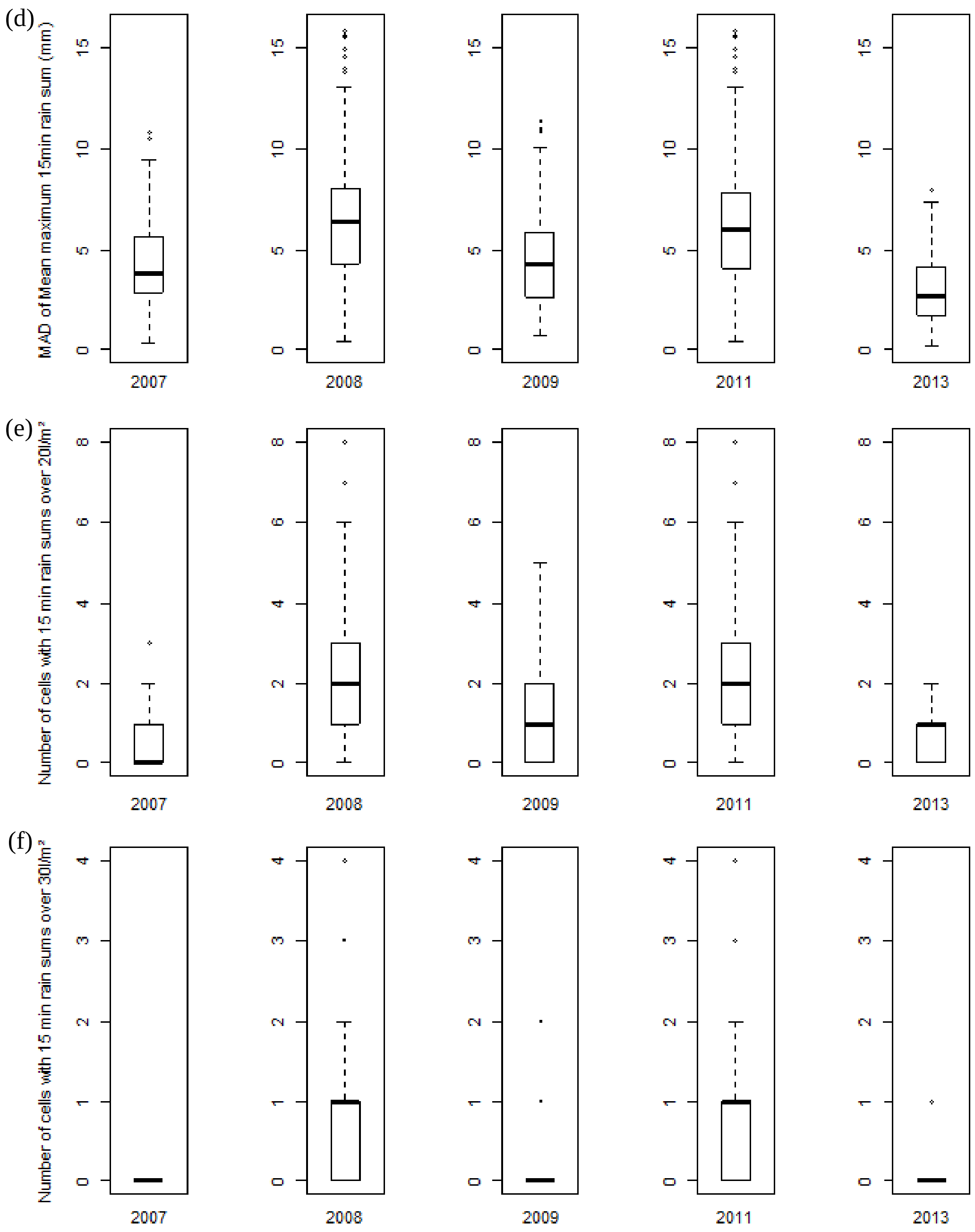


Figure 14: (a) Boxplots of mean maximum 15min rain sum (mm), (b) mean maximum 1 hour rain sum (mm), (c) mean maximum 1 day rain sum (mm), (d) MAD of annual maximum 15 min rain sum (mm), (e) number of cells with 15 min rain sums over 20 l/m², and (f) number of cells with 15 min rain sums over 30 l/m²

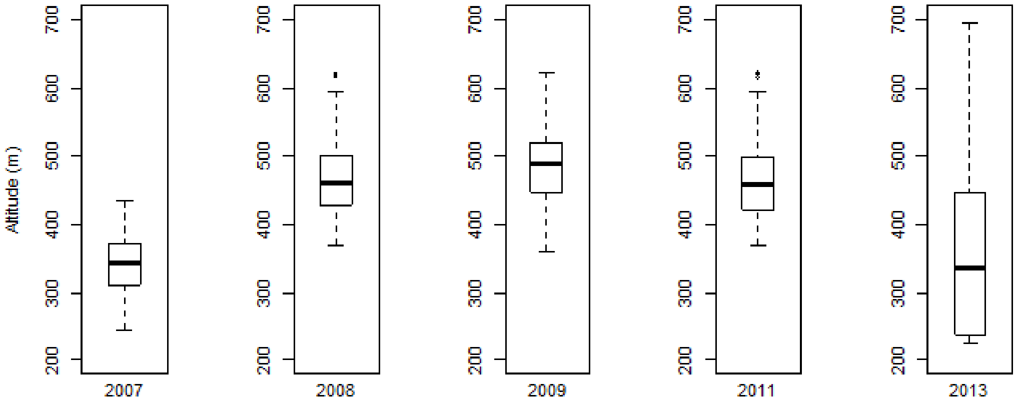


Figure 15: Boxplots of altitude of reported cells

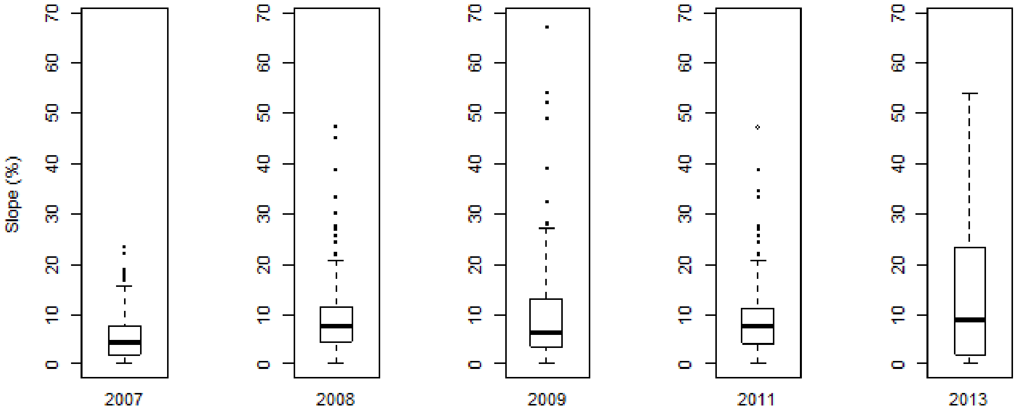


Figure 16: Boxplots of slope of reported cells

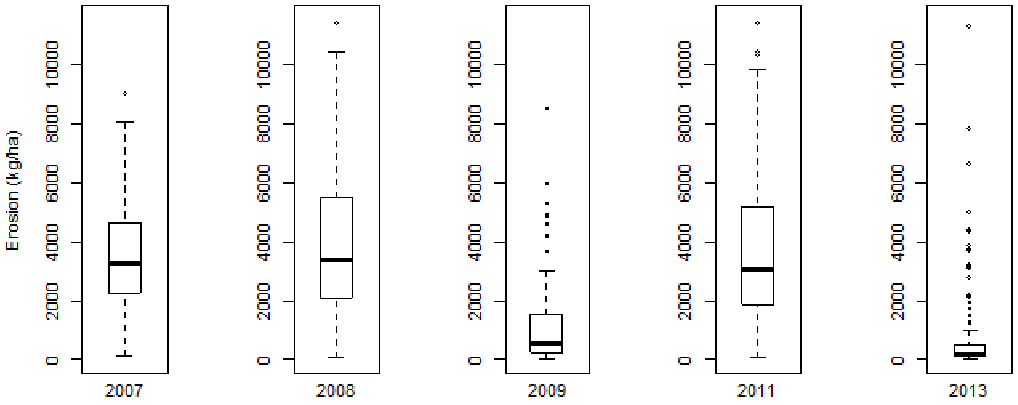


Figure 17: Boxplots of erosion of reported cells

Similar to the erosion box plots, land use can be used to divide the event years into a group of 2007, 2008 and 2011 and one with 2009 and 2013. Over 80% of the cells of the first group have arable land as their main land cover (Table 2). 2007 has 7% more but no grassland, less than 1% heterogeneous agricultural land and 6.67% forest while the other two have around 6, 8 and 2% respectively. The second group only has around 50% of arable land but much more forest, grassland and/or heterogeneous agricultural land. All years have the low values of sealed land and the missing categories of wetland and water in common.

The table of the soil textures draws a very similar picture to the one of the land use (Table 3). The locations of the events of 2009 and 2013 have a very high percentage of sand (59.83 and 41.23%) and a very low one of silt (33.33 and 36.84%) compared to the others. The location of 2007 has the highest relative amount of silt with 86.67%, the rest is almost exclusively loam and less than 1% is sand. 2008 and 2011 have the highest percentages of loam with 35.68% and 39.00%. Similarities between all years are the missing clay and the almost missing category of others.

Table 2: Relative percentages of land use per year of event analysis

	Sealed	Forest	Wetland	Water	Arable land	Grassland	Het.agr.land
2007	5.00	6.67	0.00	0.00	87.50	0.00	0.83
2008	3.24	2.16	0.00	0.00	80.54	5.95	8.11
2009	5.98	15.38	0.00	0.00	55.56	13.68	9.40
2011	3.50	2.00	0.00	0.00	80.00	6.50	8.00
2013	0.88	21.05	0.00	0.00	48.25	1.75	28.07

Table 3: Relative percentages of soil texture per year of event analysis

	Loam	Sand	Clay	Silt	Others
2007	12.50	0.83	0.00	86.67	0.00
2008	35.68	4.87	0.00	59.46	0.00
2009	4.27	59.83	0.00	33.33	2.56
2011	39.00	4.50	0.00	56.00	0.50
2013	18.42	41.23	0.00	36.84	3.51

The soil types are summarized in Table 4. The location of 2007 consists of almost one half of brown soil and the other half of pseudogley with small percentages of gley and alluvial soil. The areas of 2008 and 2011 have with 60% and 56% more brown soil but with 30% less pseudogley and almost no alluvial soil. 2009 has the highest percentage of brown soil with 79%, the other ~20% are split between pseudogley (10%), alluvial soil (5%) and gley (5%). 2013 stands out with the highest amount of alluvial soil and gley with 16% and 7% but the lowest of pseudogley with 7% and the other two third of the area consists of brown soil. Relict soil, soil form complex, rendzina and rangier and others were not represented in the event windows and only a minimal percentage of bog and atypical soil.

Across each year the soil permeability of the damaged cells was mostly medium. The only exception was the year of 2013 where it was higher. In case of the soil depth there is not much to learn apart from the fact that the reported cells had almost exclusively a high depth. The soil water conditions follow a similar pattern to the permeability and depth. Most cells are located at the centre values of the (alternating) water conditions. The only noticeable differences are the years 2007, which had a significantly greater amount of alternating soil with more wet phases, and 2013, which shows a good portion of its cells at drier conditions (Figure 18).

Table 4: Relative percentages of soil type per year of event analysis

	Relict soil	Alluvial soil	Gley	Atypical soil	Brown soil	Pseudogley	Soil form complex	Rendzina and rangier	Bog	Others
2007	0.00	2.50	4.17	0.00	49.17	44.17	0.00	0.00	0.00	0.00
2008	0.00	0.54	5.95	3.78	60.54	29.19	0.00	0.00	0.00	0.00
2009	0.00	5.13	5.98	0.00	78.63	10.26	0.00	0.00	0.00	0.00
2011	0.00	0.50	6.00	8.50	56.50	28.00	0.00	0.00	0.50	0.00
2013	0.00	15.79	7.02	0.88	67.54	7.02	0.00	0.00	1.75	0.00

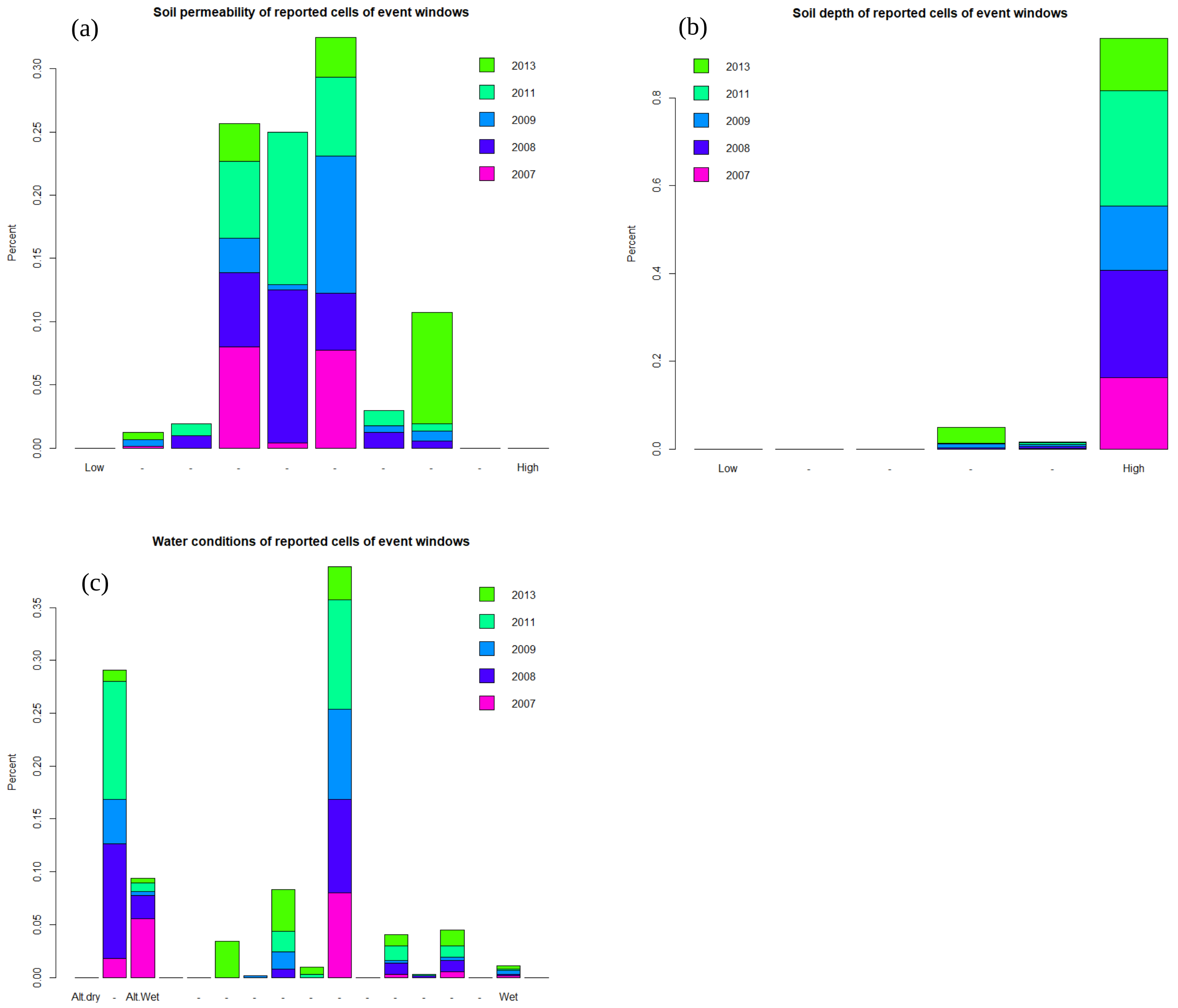


Figure 18: Soil permeability (a), depth (b) and water conditions (c) of reported cells of event analysis in percent

6.2. Location analysis of the study area

The number of 1x1km cells with reported pluvial flood damages in the time period of 2007 to 2013 is 1293 against 8297 cells without such reports. The resulting presence/absence raster for our study area can be seen in Figure 19.

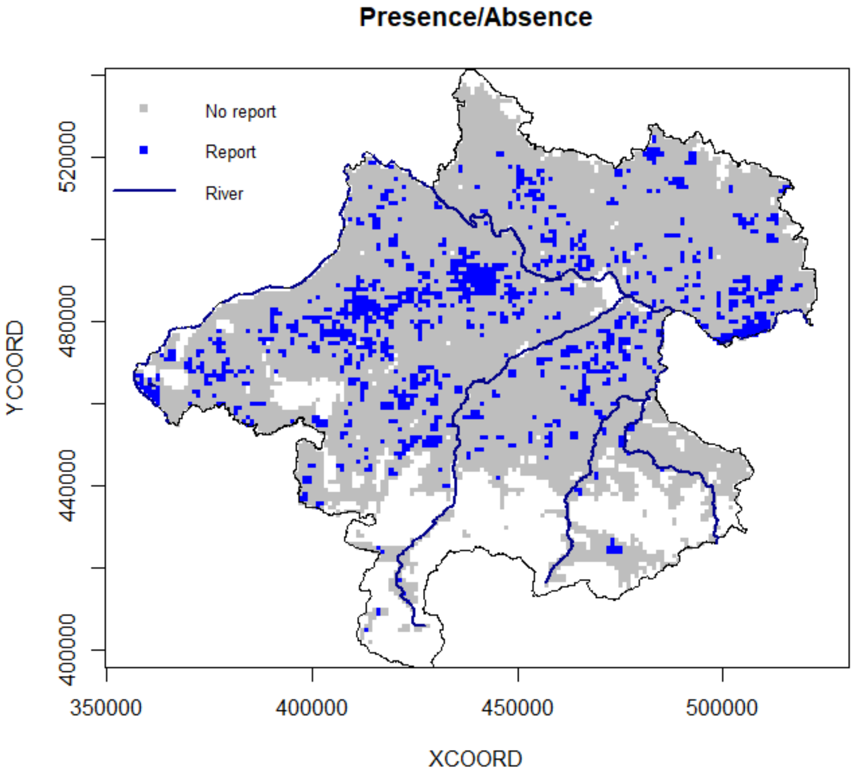


Figure 19: Observed pattern of flood reports in Upper Austria for the 2007 to 2013 period. Grid cells with reported floods are marked in blue, grey: no report, white: excluded areas.

The precipitation characteristics of the study area and of event vs. no event cells are displayed in Figures 20 – 26. The boxplots point out whether there is a general difference in the considered rainfall characteristic between cells, where no reports were sent (0) and cells where reports were sent (1), the right plot shows how these values are distributed over the study area and on the bottom the relationship between the values and pluvial flood damages is illustrated.

The boxes of the mean maximum 15 minute rain sums (Figure 20) look almost exactly the same with their medians at 12.4l/m^2 and 12.7l/m^2, which means that with respect to flood vulnerability, there is no trend visible. The kernel smoother shows a different picture. Actually there seems to be a positive linear relationship between higher 15 minute sums and pluvial flood damages. Most of the highest intensities are located in the south right in front of the Alps and in the region of the events of 2008 and 2011. Not surprisingly the mean maximum 1 hour rain sum looks very similar (Figure 22). Their medians are at 19.26l/m^2 and 19.31l/m^2 but the trend line again shows a mostly positive linear relationship between higher 1 hour sums and pluvial flood damages. In the study area the higher sums are located in the regions of the events of 2008 and 2011 and on the border of the Alps. The mean maximum 1 day sums (Figure 23) also have very similar medians at 51.2l/m^2 and 49.9l/m^2 but the variance of the values for the flooded cells is in comparison rather low. This apparently leads to a non-linear relationship between the values and pluvial flood risk. In the whole study area the highest values are now all located in the Alpine regions in the south at up to 90l/m^2. The median absolute deviations (Figure 21) are very similar for flood and no-flood cells, with their respective medians at 3.8l/m² and 4l/m². The highest values are in cells without pluvial flood damage reports, which leads to a sudden drop of the trend line after a positive relationship between pluvial flood damages and higher absolute deviations. On the map of the study area, most of the higher values are in regions that were flood damaged by the events of 2008 and 2011, but there are also a few cells with even higher values in regions that were not flood damaged.

A very similar problem can be seen at the number of 15 minute rain sums that exceeded 20l/m² (Figure 24). The medians for both boxes are one, but most of the flood damaged cells experienced more heavy rainfall than the not flood damaged ones. However, due to some cells in the west that reached the value of 10 and were not reported as pluvial flood damage, the kernel smoother suddenly sinks after a mostly positive linear relationship. The exceedances of 30l/m² (Figure 25) do not suffer from the same problem. The median for both boxes is at zero but the kernel smoother shows a positive linear relationship between pluvial flood damages and the number of 15 minute rain sums over 30l/m².

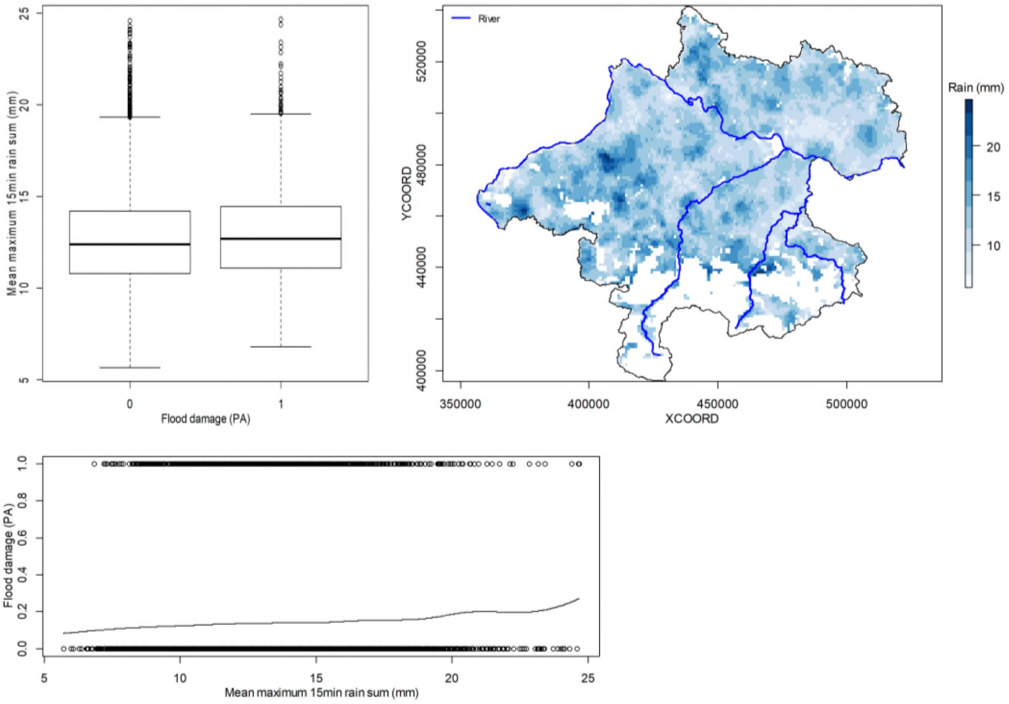


Figure 20: Mean maximum 15min rain sum (Max_rain_15 in mm). The upper panels show the areal distribution over Upper Austria (right) and empirical distributions of event vs. no-event cells (left). The lower panel plots flood damage vs. the number of events together with a locally weighted regression smoother.

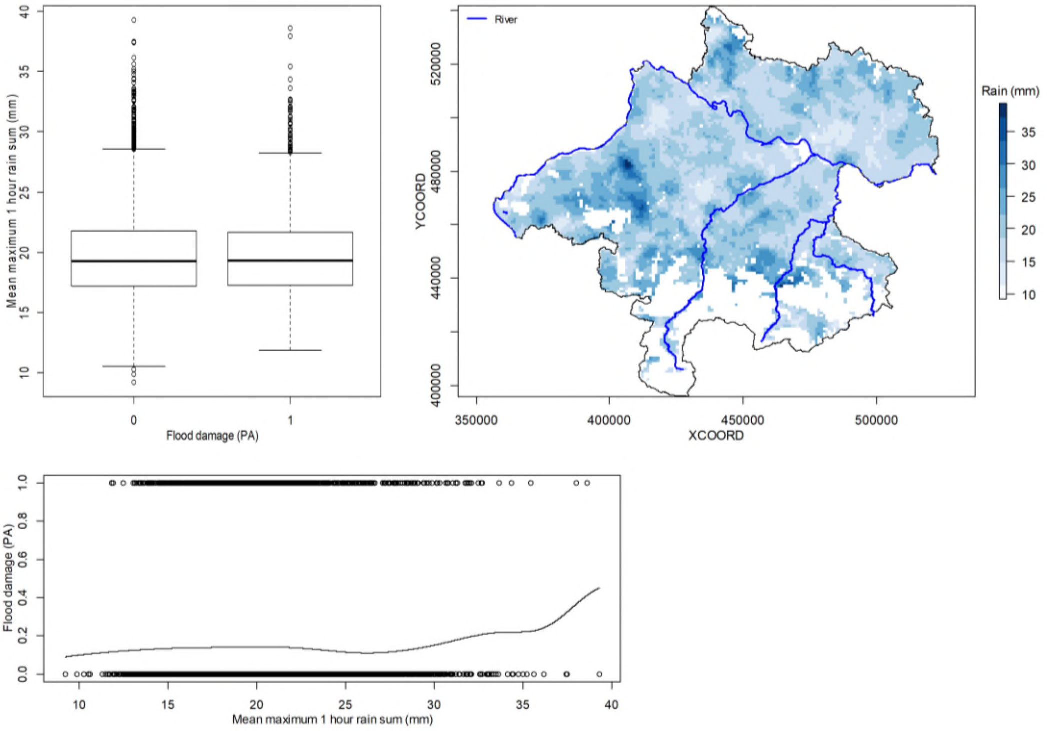


Figure 21: Mean absolute deviation (MAD) of annual Max_rain_15 in mm. The upper panels show the areal distribution over Upper Austria (right) and empirical distributions of event vs. no-event cells (left). The lower panel plots flood damage vs. the number of events together with a locally weighted regression smoother.

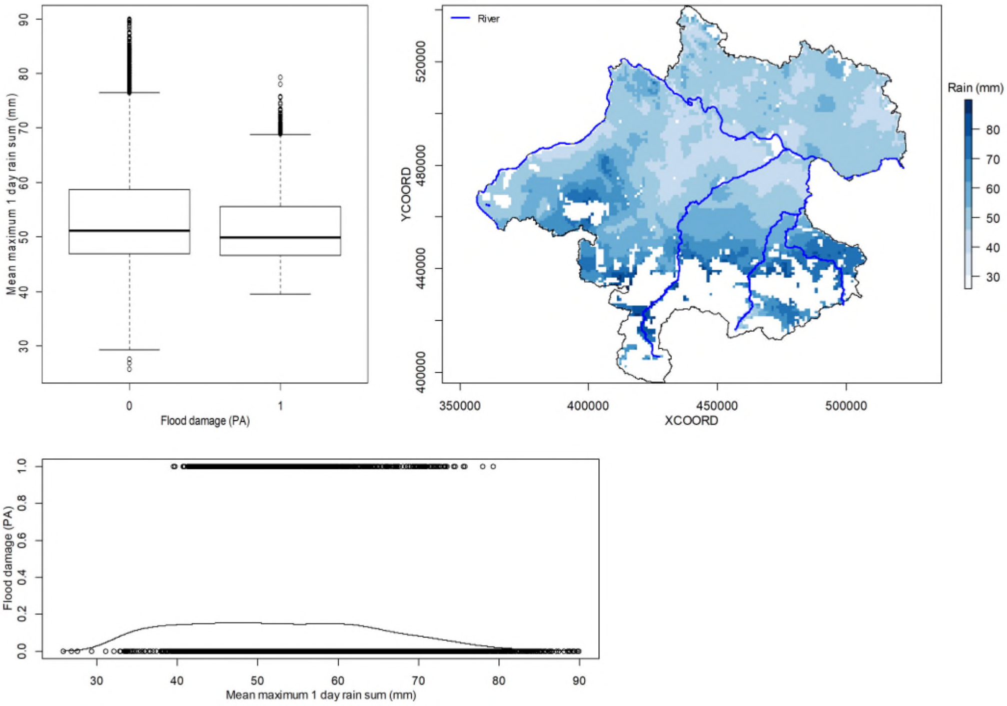


Figure 22: Mean maximum 1hour rain sum (Max_rain_h) in mm. The upper panels show the areal distribution over Upper Austria (right) and empirical distributions of event vs. no-event cells (left). The lower panel plots flood damage vs. the number of events together with a locally weighted regression smoother.

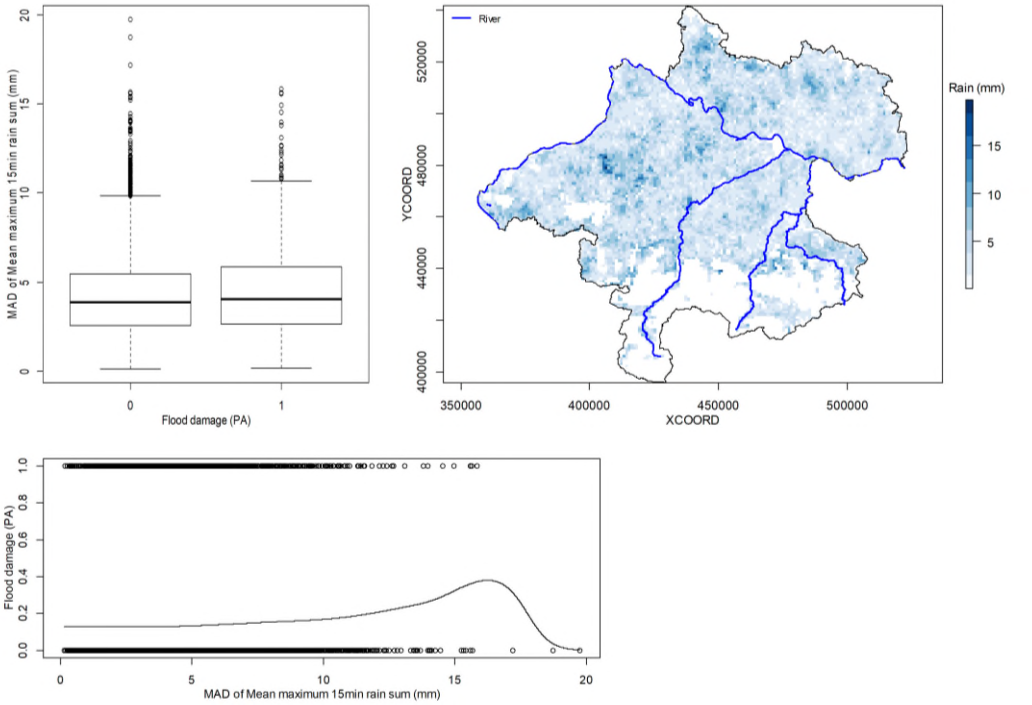


Figure 23: Mean maximum 1day rain sum (Max_rain_d) in mm. The upper panels show the areal distribution over Upper Austria (right) and empirical distributions of event vs. no-event cells (left). The lower panel plots flood damage vs. the number of events together with a locally weighted regression smoother.

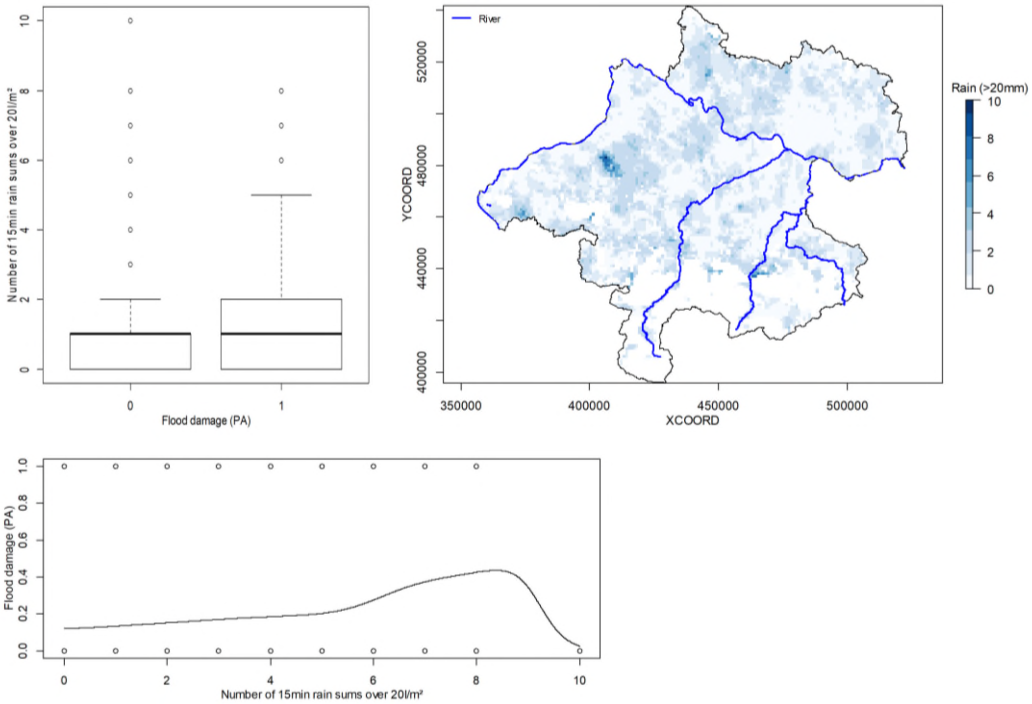


Figure 24: Number of events with 15min rain sum > 20mm. The upper panels show the areal distribution over Upper Austria (right) and empirical distributions of event vs. no-event cells (left). The lower panel plots flood damage vs. the number of events together with a locally weighted regression smoother.

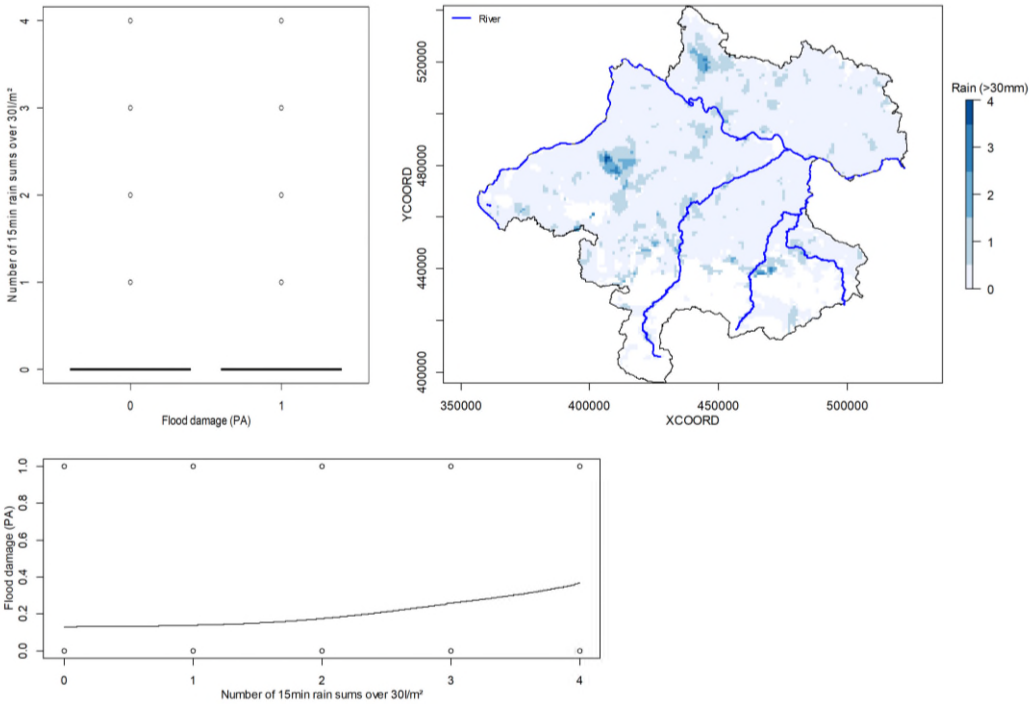


Figure 25: Number of events with 15min rain sum > 30mm. The upper panels show the areal distribution over Upper Austria (right) and empirical distributions of event vs. no-event cells (left). The lower panel plots flood damage vs. the number of events together with a locally weighted regression smoother.

In Figures 26 - 29 we can see that in general, cells with low heights and slopes look as if they are more likely to be flood damaged, while cells with high erosion may also be more susceptible. The relationships between altitude and slope to pluvial floods are mostly linear and negative, but erosion shows a nonlinear relationship. Looking at the graphs of the whole study area, it can be seen that at low altitudes the slopes are lower and the erosion higher, indicating high correlations between them.

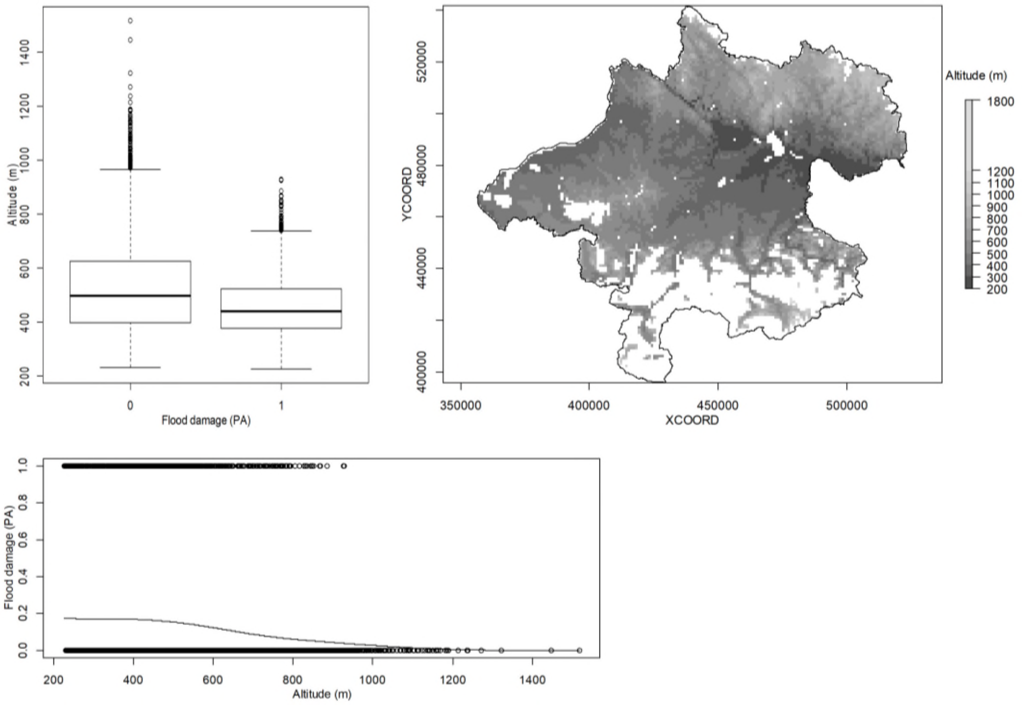


Figure 26: Altitude (in m.a.s.l.). The upper panels show the areal distribution over Upper Austria (right) and empirical distributions of event vs. no-event cells (left). The lower panel plots flood damage vs. the number of events together with a locally weighted regression smoother.

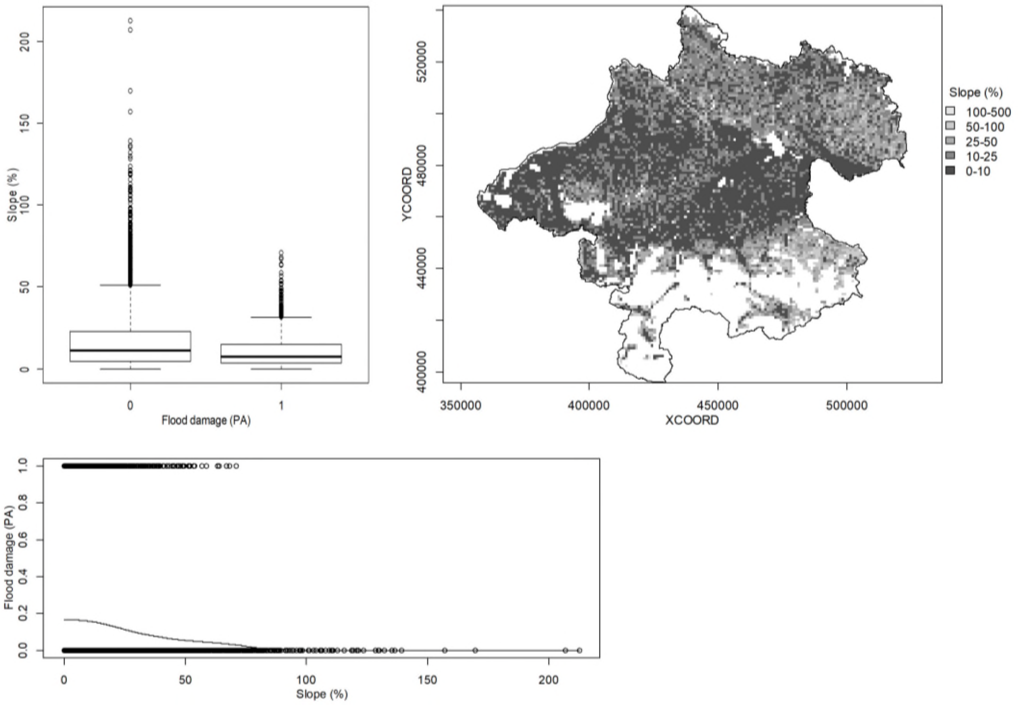


Figure 27: Slope (in %). The upper panels show the areal distribution over Upper Austria (right) and empirical distributions of event vs. no-event cells (left). The lower panel plots flood damage vs. the number of events together with a locally weighted regression smoother.

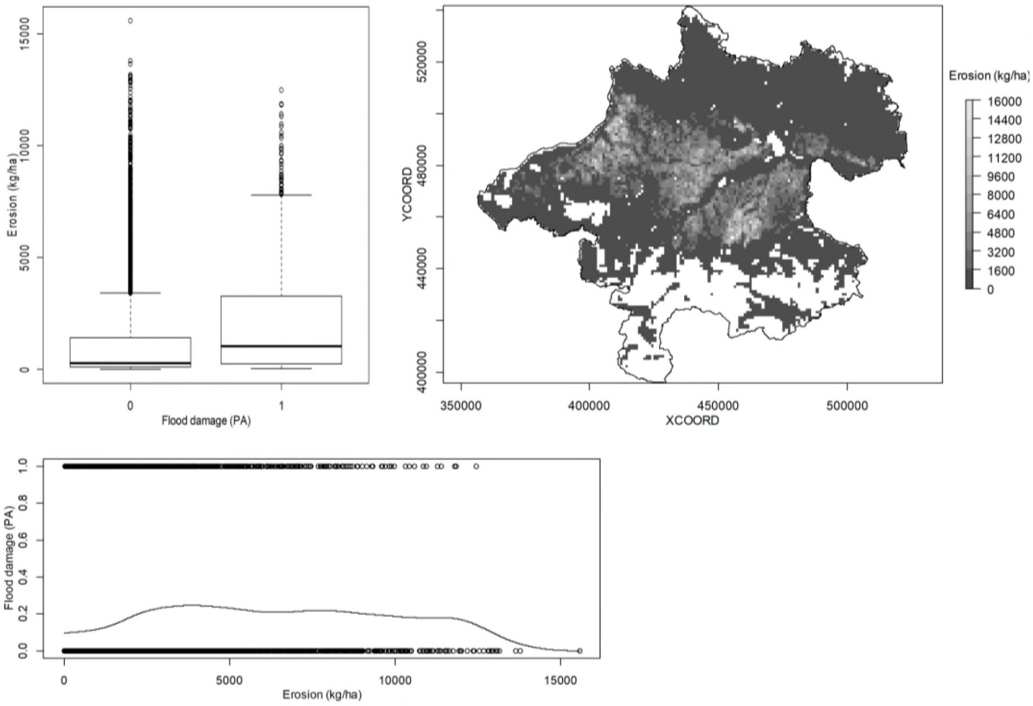


Figure 28: Erosion (in kg/ha). The upper panels show the areal distribution over Upper Austria (right) and empirical distributions of event vs. no-event cells (left). The lower panel plots flood damage vs. the number of events together with a locally weighted regression smoother.

In Figure 29 we see the land use of Upper Austria. In regions with lower altitudes most of the cells are arable land. The Bohemian Mass is characterized by the highest amount of heterogeneous agricultural land, especially in the eastern parts. With rising altitudes in the south the amount of grassland increases. Grasslands are mostly located in basins and areas with lower slopes (compared with the rest of the Alps). The regions with higher altitudes are characterized by high slopes, and thy are mainly used as forest. Cells with water and wetland as their biggest part are located around lakes or big rivers like the Danube, while sealed cells can be found around bigger cities. The whole study area consists mainly of agricultural land (37.22% arable land, 12.93% grassland and 11.78% heterogeneous agricultural land) and forest (31.71%), which can be seen in Table 5. When the area is separated by cells with and without reports, it becomes clear that most of the flood damaged cells were arable land (58.31%), which was expected because the reports refer to damages on agricultural land. Only 13.30% of the cells with reports were flood damaged agricultural land next to a forest.

Almost all of the Bohemian Mass has sand as the main soil texture (Figure 30). In the flatter regions it is mostly silt and at steeper slopes it is more loam. Only in the west of the Alps in Upper Austria there is clay as the main soil texture, which only covers 1.63% of the study area (Table 6). Out of the four different soil textures sand and silt are the most frequent in Upper Austria (41.29% and 37.38%) but half of the cells with reports were located on silt and one third on sand.

Three quarters of Upper Austria are covered by brown soil, irrespective of special regional characteristics (Figure 31). 10.21% are covered by pseudogley, which is located in lower regions and at the edge of the Alps (Table 7). The remaining 15% are covered by the other main soil types, the most noticeable being alluvial soil around rivers and basins. Out of all the soil types only pseudogley shows a bigger difference between cells with and without reports.

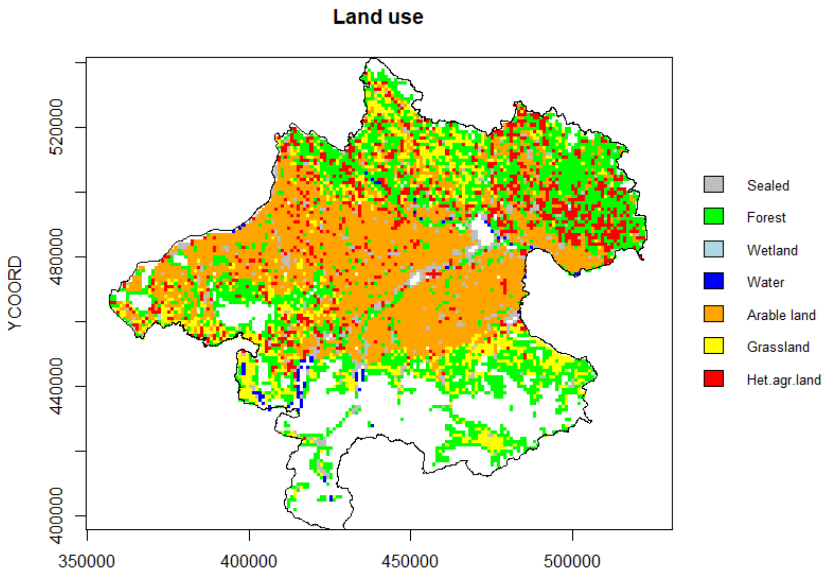


Figure 29: Plot of land use over Upper Austria

Table 5: Percentage shares of land use

	Sealed	Forest	Wetland	Water	Arable land	Grassland	Het.agr.land
All	5.56	31.71	0.03	0.76	37.22	12.93	11.78
No report	5.9	34.58	0.02	0.87	33.93	13.08	11.62
Report	3.40	13.30	0.08	0.08	58.31	11.99	12.84

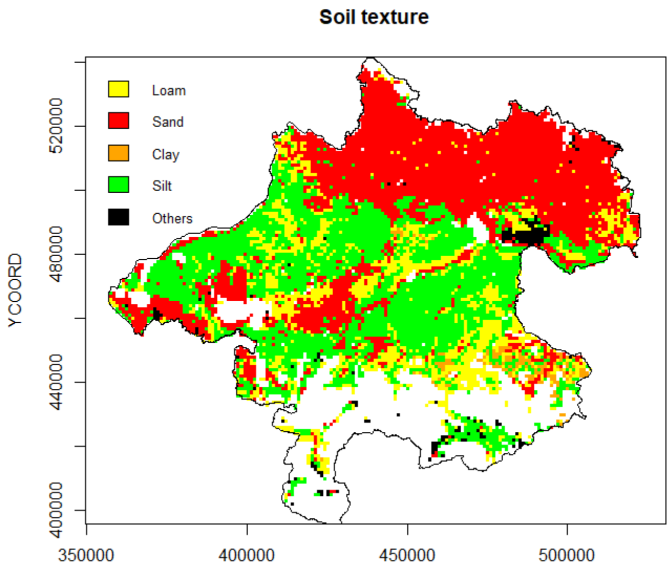


Figure 30: Plot of soil texture over Upper Austria

Table 6: Percentage shares of soil texture

	Loam	Sand	Clay	Silt	Others
All	17.86	41.29	1.63	37.38	1.85
No report	18.22	42.66	1.81	35.42	1.88
Report	15.55	32.48	0.46	49.88	1.62

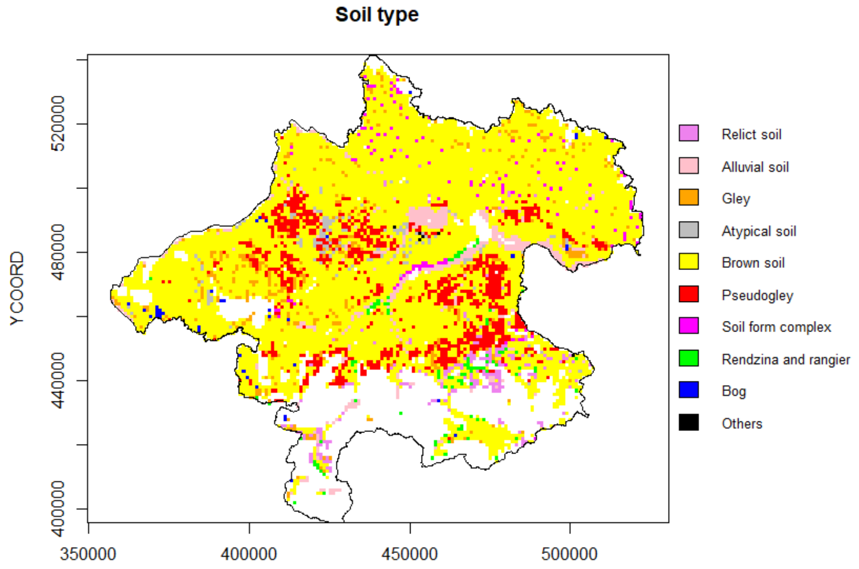


Figure 31: Plot of soil type over Upper Austria

Table 7: Percentage shares of soil type

	Relict soil	Alluvial soil	Gley	Atypical soil	Brown soil	Pseudogley	Soil form complex	Rendzina and rangier	Bog	Others
All	2.10	4.45	4.32	1.46	74.55	10.21	1.40	1.03	0.42	0.06
No report	2.36	4.45	4.27	1.27	74.98	9.51	1.53	1.18	0.37	0.07
Report	0.39	4.41	4.64	2.71	71.85	14.69	0.54	0.08	0.70	0.00

The soil permeability is typically greater in the Bohemian Mass and around rivers (Figure 32). With lowering altitude the permeability decreases too. Figure 33 shows a general trend of a higher vulnerability of soils with a low permeability.

The soil depth is in contrast to the permeability large at lower altitudes and more shallow at higher altitudes and around rivers (Figure 34). Figure 35 shows a higher vulnerability of soil with a high depth. However, as most of Upper Austria is characterized by soil with a high depth. it is questionable, if soil depth is an adequate indicator for a higher flood risk.

Low water conditions of the soils are typically located at hills and low mountain ranges higher altitudes with high slopes (Figure 36). Ravines and basins in these areas exhibit wet or medium water conditions. Medium water conditions are also typical for the regions in the centre and the west of the study area. The alternating soils are mixed in between the medium water conditions but rarely reach higher altitudes. The alternating soils with more dry phases are located around rivers, while the ones with more wet phases are rather in the east, west or south. Figure 37 shows a general trend towards more floods in wet soils, which applies to the alternating and the more consistent types. The decline of the PA in the wet conditions is actually related to a single class and is therefore not considered a downward trend.

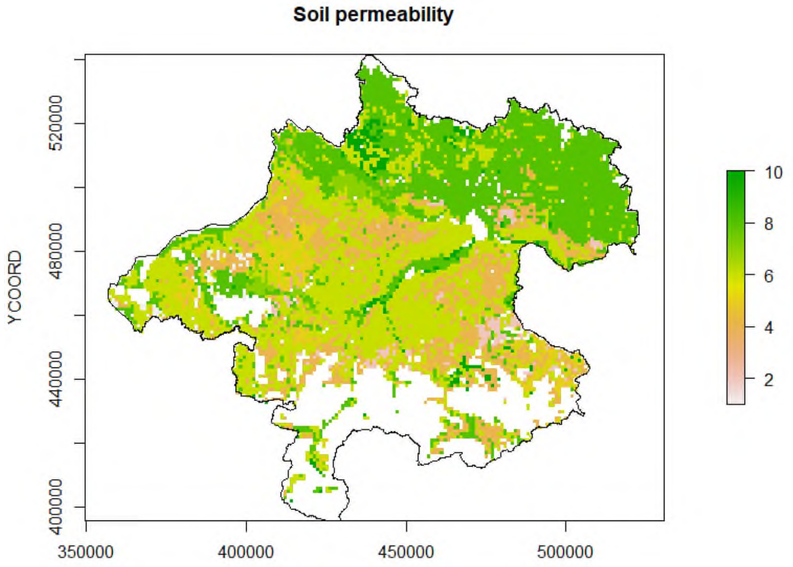


Figure 32: Plot of soil permeability over Upper Austria

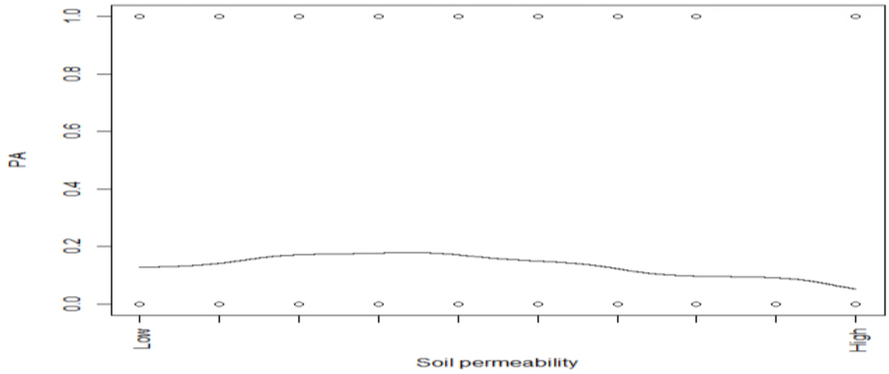


Figure 33: Soil permeability and reports with kernel smoother

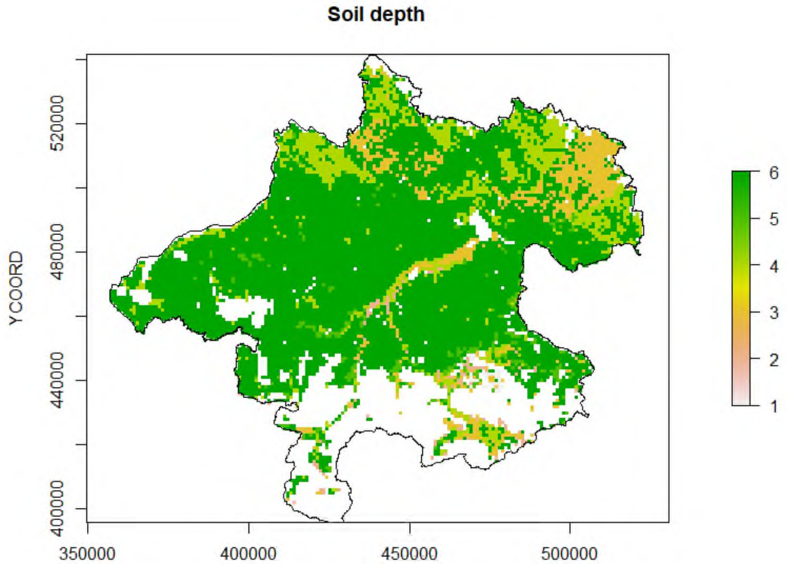


Figure 34: Plot of soil depth over Upper Austria

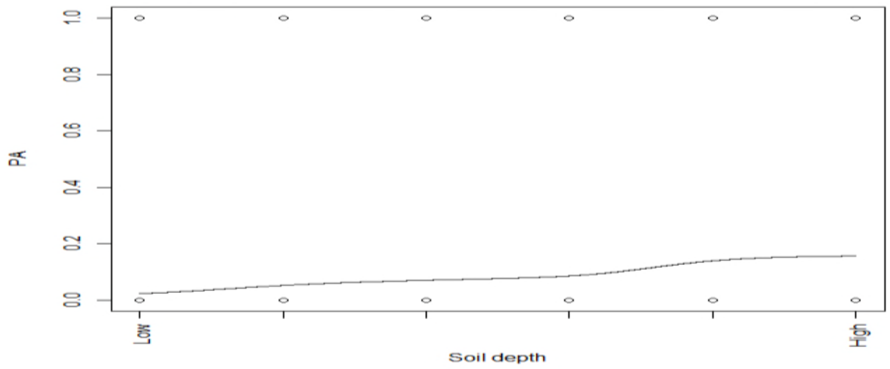


Figure 35: Soil depth and reports with kernel smoother

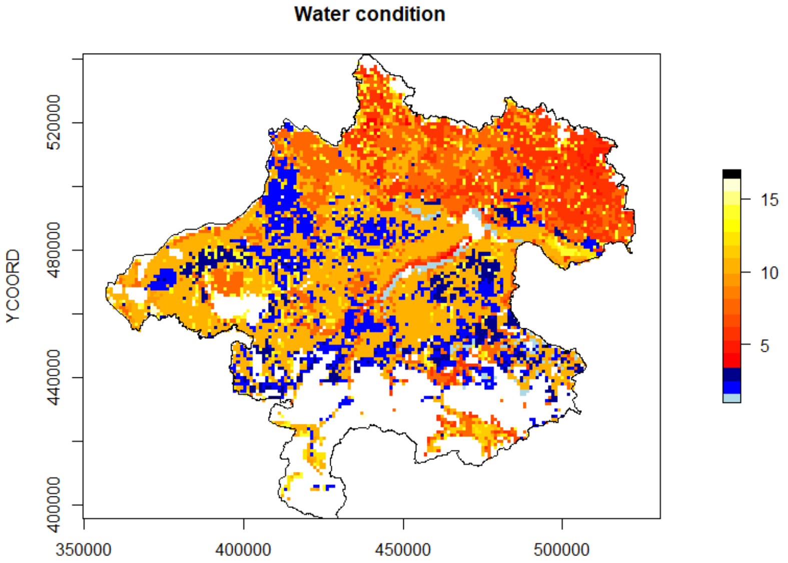


Figure 36: Plot of soil water conditions over Upper Austria

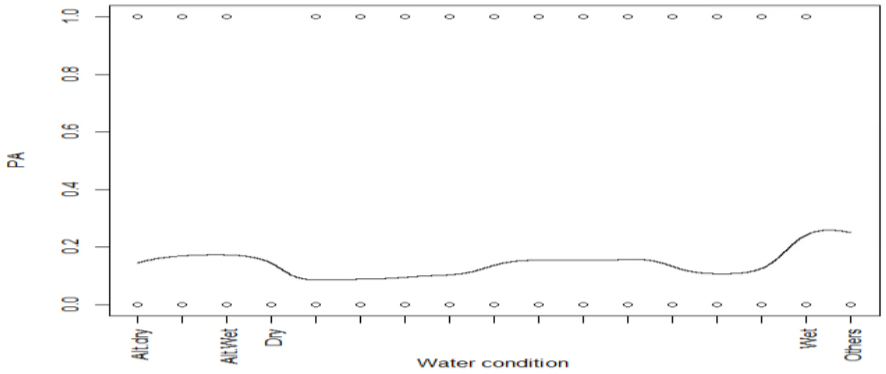


Figure 37: Water condition and reports with kernel smoother

6.3. Summary

- Mean maximum 15 minute rain sums

The mean maximum 15 minute rain sums doesn't show a clear separation between report and no report cells. Even the medians of the extracted event cells differ between 10.5l/m² and 15.8l/m² when the boxplot of the whole study area shows them at around 12.4-12.7l/m².

- Mean maximum 1 hour rain sum

The 1 hour rain sums don't differ much from the 15 minute sums. Their medians are around 7l/m^2 higher, but otherwise they are very similar.

- Mean maximum 1 day sums

Even for the 1 day sums the relationships between the medians didn't change. For the extracted event cells, the medians range between 45.8l/m^2 and 57l/m^2, while for the study area they are at 51.2l/m^2 and 49.9l/m^2. Due to the non-linearity shown by the kernel smoother, this variable will not be used in the logistic regression.

- Median absolute deviation of the 15 minute rain sums

The median absolute deviation behaves like the other rain variables, only with much smaller values. The range of the medians from the boxplots from the extracted cells is 2.7l/m^2 and 6.4l/m^2. On the whole study area they are again very close together at 2.8l/m^2 and 4l/m^2.

- Number of 15 minute rain sums that exceeded 20l/m^2 and 30l/m^2

As the numbers of exceedances of 20l/m^2 and 30l/m^2 are very low, their medians range only between zero and two. However, the maximum number for the extracted cells is eight for the 20l/m^2 threshold and three for the 30l/m^2. On the whole study area, there are a few cells, where the 20l/m^2 mark was exceeded ten times and four times at the 30l/m^2 mark. Differences in the medians between reported and not-reported cells are not present.

- Altitude

At the reported cells of the event windows the median of the altitudes was located between 336 and 490 meters. On the scale of the whole study area they were around 497 for the unreported cells and 439 for the reported ones. While the event of 2009 was located at a higher altitude, the other event years were at the level of the events of the whole area or lower.

- Slope

The median of the slopes for each major event is between 4.5 and 9 % while the medians for the whole study area are at 7.5 % for reported cells and 11 % for unreported. Most of the cells of the event years fit well around the 7.5 % of all reported cells. Only the event of 2013 is higher at 9 % but still closer to the reported cells than the unreported. Overall a flatter ground might indicate a higher pluvial flood risk.

- Erosion

The erosion separated the events into two groups: median above 3000kg/ha and median below 1000kg/ha. While the Figures for the entire study area give a clear indication that higher erosion increases the pluvial flood risk, the damage events of 2009 and 2013 happened in regions, where erosion was at a much lower level. Even the kernel smoother doesn't show a clear linear relationship between flood damaged and non-flood damaged cells, which means that the erosion will not be integrated in the logistic regression.

- Land use

First of all the land use data shows that most of the reported cells cover agricultural land (58.31%, 11.99% and 12.84%) and only a few were next to forests (13.30%). The cells without a report are more separated with around 1/3 of forest, 1/3 of arable land and 1/4 of grassland and heterogeneous agricultural land. With more than 80% the years of 2007, 2008 and 2011 lean heavily toward arable land and far less to the other categories. The years of 2009 and 2013 on the other hand contained a greater mixture of categories. Only about half of the cells damaged by these floods were arable land but the proportion of grassland or heterogeneous agricultural land was higher. Additionally the percentage of forest (15.38% and 21.05%) was higher compared to the other years and the whole study area. Due to the low percentages the classes of sealed, wetland and water will be added to the category of others in the models of aggregated classes.

- Soil texture

Flood damage cells are reported in areas with a higher proportion of sand (42.64% vs 32.48%) and a lower proportion of silt (35.42% vs. 49.88%), as compared to cells without flood observation. However, these proportions differ substantially between the events. The event of 2007 occurred on soil with a high amount of silt (59.46%) and some loam (12.50%). The soils of the events of 2008 and 2011 had around 30% less silt but with 35.68% and 39.00% far more loam and around 5% sand. Sand poses the main soil texture at the events of 2009 and 2013 with 59.83% and 41.23%, with about 1/3 silt and the rest is mostly loam. A synoptic view of Figure 29 and 30 suggests that soil texture is highly correlated with land use in the study area. This suggests that soil texture characteristics of cells with and without a report are likely influenced by the nature of the target variable (reported agricultural damage) and cannot be seen as an independent indicator of pluvial floods in the study area. In the models of aggregated classes the class of clay will be combined to the class of others due to the very low percentages.

- Soil type

Out of the ten different soil type classes only three look particularly interesting: alluvial soil, brown soil and pseudogley. While the alluvial soil is only a small factor for most event years, it represented 15.79% of soil types in the 2013 event, which is not surprising as this type is normally located near rivers and catchment areas. The brown soil of the event years vary between 48.33% and 78.63%. On the whole study area brown soil covers 3/4 of the area. So on the event level it could be an indicator for pluvial flood damages, however, it shows only a small difference of 2.65%. The most interesting type is pseudogley because on the whole area the cells without reports have 9.74% pseudogley but the event years reach up to 44.17% resulting in an overall difference of 4.88%. Based on the overall differences the only class that will remain besides “Others” will be pseudogley in the model of aggregated classes.

- Soil permeability

The soil permeability is usually higher in the Bohemian Mass and around rivers. For this reason, the event of 2013 had a higher percentage of damaged cells at a higher permeability. The events in the other years affected areas with medium permeability, which represents the most beneficial permeability for agriculture. Nevertheless the overall trend of a higher risk goes into the direction of lower soil permeability.

- Soil depth

Looking at the soil depth, the expectation of it being an influential factor to pluvial floods is quite low. The largest part of the study area is covered with a high soil depth, only areas at higher altitudes and around rivers have a lower depth. Even the event of 2013, which is located next to a river, shows a small percentage of a lower soil depth. At the separation of the presence/absence cells the trend indicates an increasing risk with increasing depth, which might be correlated rather with the altitude than with the actual soil depth.

- Water conditions

The soil water conditions are usually drier in higher regions and around rivers. The dry soil next to water might be influenced by the lower soil depth and higher soil permeability. Similar to the soil permeability the reported cells from the event years are mostly located at medium water conditions (also alternating conditions without majority of wet or dry phases), the best condition for agriculture. The biggest exceptions are the event of 2013, which is drier, and 2007, which is also located at alternating soils with more wet phases. The overall trend points to a higher risk with increasing wetness, also for the alternating soils. However, water conditions have a similar problem to soil depth in terms of similarities with the altitude. For the modelling with aggregated classes the water conditions will be reduced to four: dry, medium, wet and others and furthermore tested as a continuous and a factor variable. For a list of factor levels of the aggregated classes see Table 10.

7. Modelling

In this chapter, the two data sets will be modelled with a logistic regression and random forests. The data set and resulting model with the original classes will also be referred to as "first", while the data and model with the aggregated classes will also be referred to as "second".

7.1. Logistic regression

For the logistic regression it is important to know, which variables correlate with each other and how important they are on their own. To better illustrate these correlations, a heat map is used, which shows negative (green), neutral (yellow) and positive (red) correlations. Finally, they group samples together using dendrograms, or tree diagrams, where the most similar variables are split at the bottom. Out of highly correlated variable groups only one will be chosen for the model. Additionally, variables that don't show significances in the single models are also excluded in further modelling. Then, for each single model the pseudo R^2 is calculated and compared to the others. It is used to determine, which variable out of a highly correlating group will be chosen. As explained in chapter 4.3, two data sets are compared with each other: one with the original data and one with aggregated classes, which are determined in chapter 6.3. These aggregated variables are analysed in the same way as the others before.

In the next step, a regression model for each data set with all predictors selected in the previous step is analysed. The interpretation for it is different to the usual regression. The estimated coefficients show the increase or decrease (depending on the sign) of the log odds for pluvial flood damages. In the next column the standard error is shown and next to it the z-statistics, which is calculated by dividing the estimate by the standard error. Corresponding to the z-statistics is the p-value, where a small value indicates a high significance. Variables that are not significant (p-value over 0.1) or have a sign that cannot be explained by the previous analyses will be excluded. At a last step the models are checked for multicollinearity using the generalized variation inflation factor (GVIF), where variables with a value over 10 have to be excluded. At the end, the two final models are interpreted and compared. Additionally, a forward stepwise regression is carried out to see, if it would choose variables differently.

These two models are split into training and test set in a ratio 80:20. First, the model is fitted using the training data and then used to predict the test set. The true and false predictions are taken from a confusion matrix and illustrated in a Table for three cut-offs. Finally, the probabilities are plotted over the whole study area to determine, where high risk areas are located.

7.1.1. Single models and correlations

This chapter provides information on how good each variable can explain pluvial floods individually and independent from other variables (Tables 8 and 9). Furthermore, the correlations between the variables are assessed and illustrated (Figure 38).

The mean maximum 15 minute rain sum shows a very high significance with a positive estimate, which indicates a rising pluvial flood risk with increasing rain intensity. It highly correlates with the mean maximum 1 hour sum (0.827) and the number of >20l/m^2 intensities (0.742). On the other hand the mean maximum 1 hour rain sum does not show any significance to the pluvial flood risk. As a result, it will not be used for further modelling, which is beneficial regarding the high correlation. The mean absolute deviation of the 15-minute sum is less significant than Max_rain_15, but still very significant with a positive estimate. It correlates mostly with the variables Max_rain_15 (0.512), Max_rain_h (0.395) and Rain_20 (0.456). Rain_20 has the highest significance and a positive estimate, indicating a higher risk of pluvial flooding with more rain intensities above 20l/m^2. As noted above it highly correlates with the Max_rain_15, Max_rain_mad and also Max_rain_h (0.582). Rain_30 also has the highest significance and a positive estimate and correlates the most positive with Rain_20 (0.448) and Max_rain_15 (0.474).

The altitude has a negative estimate and is very significant indicating a higher pluvial flood risk at lower altitudes. It positively correlates mostly with the slope (0.462). As this positive correlation shows, the slope is also very significant and has a negative estimate, indicating a higher risk in lower slopes.

The interpretation for factor variables is a little different. Here, the intercept is the sealed land and serves as the reference to the other classes. So, the estimate of forest is 0.38 lower than the estimate of the intercept and arable land is 1.11 higher. Consequently, the most significant land use class is the arable land followed by heterogeneous agricultural land. The classes of sealed and wetland are also significant but are too few to be significant for our study. Water is not significant and almost non-existent, while grassland is also very significant. From the most represented classes, forest indicates a decreasing pluvial flood risk, while each agricultural class increases the risk. However, it should be noted that our flood damage data originate from flood damaged agricultural areas. The most significant soil texture is silt, which indicates an increasing flood risk. However, as explained in the location analysis it is the typical texture for agricultural land. Clay is also very significant, but only present in a few cells. Loam and sand are slightly significant and indicate a decreasing pluvial flood risk, while others are not significant. With the exception of rendzina and rangier and others, every type of soil has a very high significance. They also increase the risk of flooding, except for relict soil and soil formation complexes. Nonetheless, the largest part of the study area is covered with brown soil and only pseudogley covers also more than 10%. The single model with the soil permeability indicates a decreasing pluvial flood risk with increasing soil permeability, which shows a very high significance. Its highest correlations are with the soil depth (-0.496) and erosion (-0.442) but agricultural land is more likely to be on soils with a lower permeability. The soil depth is also very important and indicates an increase in the flood risk, but is also increasingly located on cells with agricultural land. Beside its negative correlation with permeability, it correlates positively with erosion (0.402). Most soil water condition classes are highly significant, with overall wet soils indicating an increased risk. There are two classes between medium and wet, which alone would indicate a decrease in risk, but they are too few and are probably coincidentally located at cells that were not flood damaged. At the alternating soils only the ones with more dry phases indicate a decrease in flood risk. Nonetheless, many dry soils are located in the Bohemian Mass, where there is less agriculture and alternating soils with more dry phases are fewer compared to the other two alternating classes.

Over all pseudo R^2, the highest value is achieved by the land use, followed by altitude (Table 9). The rain variables do not perform very well, Max_rain_h, which was not significant, has basically a pseudo R^2 of 0. The other soil variables are at a value of 0.02 at the McFadden R^2 and 0.05 to 0.07 at Nagelkerke's.

Out of the first three correlating variables of Max_rain_15, Max_rain_h and Rain_20, Max_rain_h will not be included due to non-significance and Max_rain_15 has a lower pseudo R^2 (0.003) than Rain_20 (0.008), which means that Rain_20 will be chosen for modelling. Between Max_rain_15_mad and Rain_30, Rain_30 has a higher pseudo R^2 (0.006 vs. 0.002) and will therefore be chosen. The altitude and slope are also highly correlating and the former has a higher R^2, but the slope will be integrated into the model nonetheless because of its comparatively high R^2-value. The final pair is permeability and depth, where depth has a higher pseudo R^2 at 0.025 than permeability at 0.014.

Table 8: Summary of single models for the first model

Variables/Classes	Estimate	Std. Error	z value	Pr(>|z|)
Max_rain_15	0.046891	0.006381	7.348	2.01e-13 ***
Max_rain_h	-0.0009745	0.0046823	-0.208	0.835
Max_rain_15_mad	0.04742	0.01401	3.386	0.00071 ***
Rain_20	0.18409	0.01558	11.815	< 2e-16 ***
Rain_30	0.38593	0.03824	10.094	< 2e-16 ***
Altitude	-0.0033343	0.0001274	-26.18	< 2e-16 ***

Slope	-0.02893	0.00133	-21.75	< 2e-16 ***
Land use				
Sealed	-0.48268	0.08124	-5.941	2.83e-09 ***
Forest	-0.38308	0.08984	-4.264	2.01e-05 ***
Wetland	1.73544	0.80589	2.153	0.0313 *
Water	-14.08339	108.65823	-0.130	0.8969
Arable land	1.11494	0.08534	13.064	< 2e-16 ***
Grassland	0.48381	0.09418	5.137	2.79e-07 ***
Het.agr.land	0.69935	0.09501	7.361	1.83e-13 ***
Soil texture				
Loam	-0.10324	0.04187	-2.466	0.0137 *
Sand	-0.09198	0.05037	-1.826	0.0678 .
Clay	-1.68055	0.23952	-7.016	2.28e-12 ***
Silt	0.55375	0.04949	11.189	< 2e-16 ***
Others	0.11121	0.13300	0.836	0.4031
Soil type				
Relict soil	-1.9794	0.2328	-8.502	< 2e-16 ***
Alluvial soil	1.9921	0.2461	8.093	5.81e-16 ***
Gley	2.0640	0.2469	8.358	< 2e-16 ***
Atypical soil	2.6725	0.2659	10.052	< 2e-16 ***
Brown soil	2.0259	0.2336	8.671	< 2e-16 ***
Pseudogley	2.5810	0.2382	10.834	< 2e-16 ***
Soil form complex	1.0631	0.2959	3.593	0.000327 ***
Rendzina and rangier	-0.4055	0.4585	-0.884	0.376508
Bog	2.6374	0.3264	8.080	6.46e-16 ***
Others	-9.5867	88.0869	-0.109	0.913336
Permeab.	-0.16533	0.01019	-16.23	< 2e-16 ***
Depth	0.39256	0.01856	21.15	< 2e-16 ***
Water conditions				
Alternating dry	-2.1316	0.2827	-7.541	4.66e-14 ***
Alternating	2.5893	0.2859	9.055	< 2e-16 ***
Alternating wet	2.5982	0.2906	8.942	< 2e-16 ***
Very dry	-11.4344	99.4238	-0.115	0.908439
Very dry to dry	1.3695	0.4297	3.187	0.001438 **
Dry	1.6712	0.2874	5.814	6.08e-09 ***

Dry to moderately dry	1.7750	0.4487	3.956	7.63e-05 ***
Moderately dry	1.9006	0.2858	6.650	2.93e-11 ***
Moderately dry to medium	2.1680	0.3410	6.358	2.05e-10 ***
Medium	2.3715	0.2840	8.349	< 2e-16 ***
Medium to moderately wet	1.7918	0.3474	5.158	2.49e-07 ***
Moderately wet	2.4193	0.2970	8.145	3.81e-16 ***
Moderately wet to wet	2.2858	0.6240	3.663	0.000249 ***
Wet	1.9001	0.2957	6.426	1.31e-10 ***
Wet to very wet	2.9789	0.7457	3.995	6.48e-05 ***
Very wet	2.8919	0.3474	8.325	< 2e-16 ***
Others	-11.4344	535.4112	-0.021	0.982961

Table 9: Pseudo R^2 of the single models for the first model

Variables	McFadden	Cox and Snell	Nagelkerke
Max_rain_15	0.003	0.007	0.007
Max_rain_h	<0.001	<0.001	<0.001
Max_rain_15_mad	0.002	0.001	0.003
Rain_20	0.008	0.019	0.020
Rain_30	0.006	0.014	0.015
Altitude	0.039	0.094	0.102
Slope	0.029	0.069	0.075
Land.use	0.067	0.156	0.169
Soil.text	0.021	0.052	0.057
Soil.type	0.020	0.048	0.053
Permeab.	0.014	0.035	0.038
Depth	0.025	0.062	0.067
Water.con	0.026	0.063	0.069

The land use classes are very significant, with forest decreasing the flood risk and agricultural land increasing it (Table 10). Moving the class of clay to others changed almost nothing for the other soil textures. Pseudogley as only remaining soil type is still very significant, but has a lower estimate. However it still indicates an increase in pluvial flood risk. The class of others, which is now the intercept and consists mostly of brown soil, is not significant. The soil water conditions both show an increasing flood risk on wet soils and they are very significant. Concerning the factors there is a big increase in risk from dry to medium soils but between medium and wet soil there is not much of a difference. As a continuous variable it shows a high negative correlation with permeability (-0.571) and a positive one with depth (0.607).

The pseudo R² values of land use and soil texture have not changed much from the aggregation (Table 11). The soil type lost considerably over all R² measures. The water conditions reach better values when used as a factor, but by aggregation their pseudo- R² is halved.

Table 10: Summary of single models for the second model

Variables/Classes	Estimate	Std. Error	z value	Pr(>|z|)
Land use				
Forest	-0.86576	0.03835	-22.575	< 2e-16 ***
Arable land	1.49802	0.04641	32.281	< 2e-16 ***
Grassland	0.86689	0.06115	14.176	< 2e-16 ***
Het.agr.land	1.08243	0.06243	17.339	< 2e-16 ***
Others	0.25315	0.08712	2.906	0.00366 **
Soil texture				
Loam	-0.10324	0.04187	-2.466	0.013670 *
Sand	-0.09198	0.05037	-1.826	0.067840 .
Silt	0.55375	0.04949	11.189	< 2e-16 ***
Others	-0.42779	0.11205	-3.818	0.000135 ***
Soil type				
Others	0.01908	0.01814	1.052	0.293
Pseudogley	0.58255	0.05370	10.849	< 2e-16 ***
Water.con	0.42485	0.03508	12.11	< 2e-16 ***
Water conditions				
Dry	-0.58069	0.04921	-11.800	< 2e-16 ***
Medium	0.75804	0.05298	14.309	< 2e-16 ***
Wet	0.82821	0.07034	11.775	< 2e-16 ***
Others	-9.98534	119.46805	-0.084	0.933

Table 11: Pseudo R² of the single models for the second model

Variables	McFadden	Cox and Snell	Nagelkerke
Land.use	0.064	0.148	0.161
Soil.text	0.018	0.045	0.049
Soil.type	0.006	0.016	0.017
Water.con (continuous)	0.008	0.019	0.021
Water.con (factor)	0.012	0.029	0.032

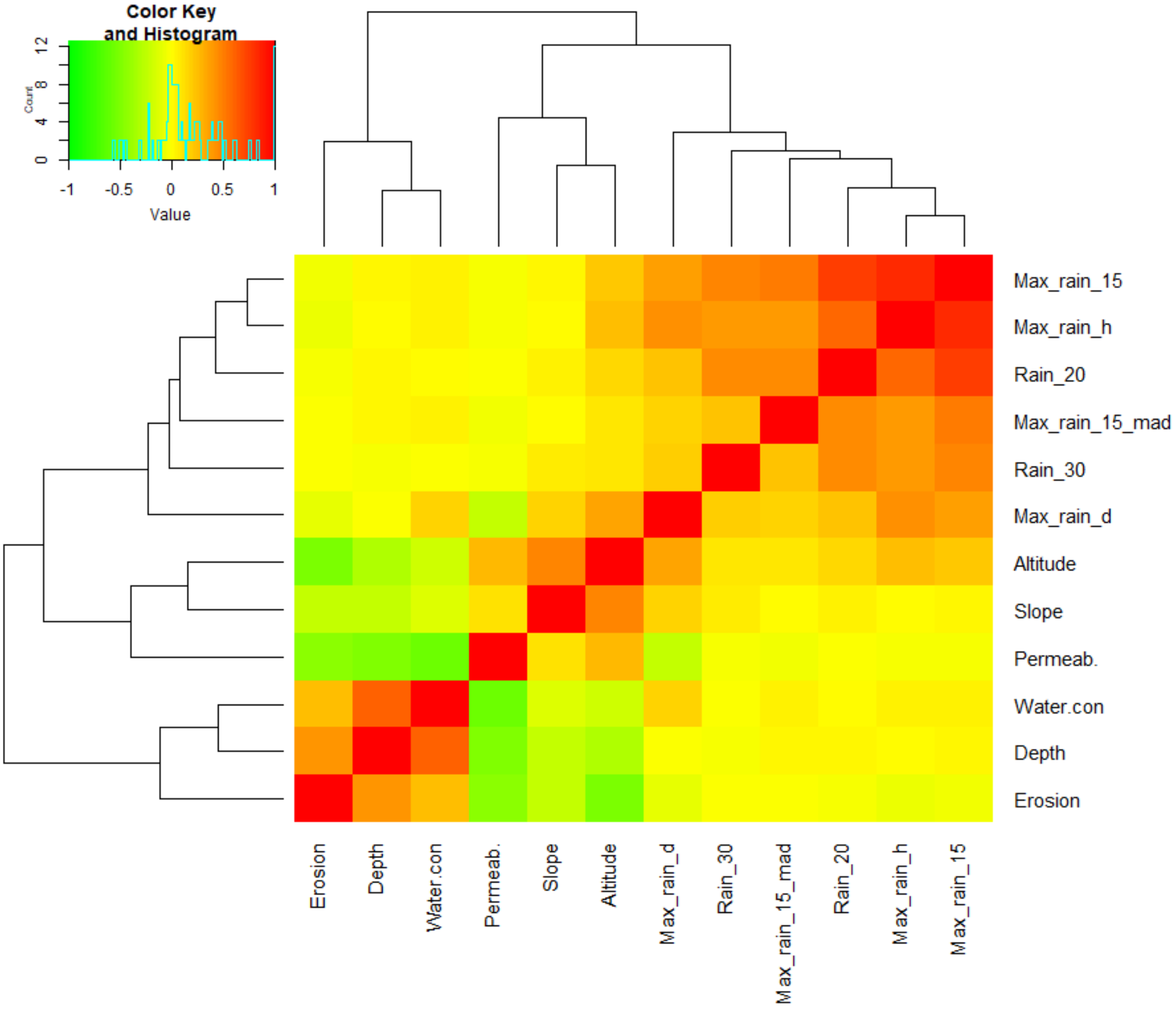


Figure 38: Heatmap of correlations between predictor variables

7.1.2. Model optimization

So far, the variables erosion and Max_rain_d were excluded because of non-linearity, Max_rain_h because of the non-significance and the variables Max_rain_15, Max_rain_mad and permeability because of high correlations. As these variables have also been excluded by the stepwise regression, only the latter is interpreted, because it provides more information. In oder to investigate if more variables can be excluded, we look at the summary of the resulting model in Table 12.

In the forward stepwise regression variables are added based on their importance to the model. So, the land use, altitude and Rain_20 are most important. There were even a few variables not included, which means that adding them would not benefit the model. These variables are the permeability, Max_rain_15 and Max_rain_15_mad. Due to the correlation structure, it cannot be concluded that they do not contribute to the model. What's more likely is that they do not contribute much more to the model than the other rain variables. For example, Rain_20 is one of the most important variables and is highly correlated with Max_rain_15. By including Rain_20 most of the explanatory power of Max_rain_15 is included in the model as well.

Usually in linear regression, we could calculate for each variable and class how much the probability (flood risk) would change, if it would rise by 1. However, with that many interactions, it would depend on other variables as well, so we would need interaction terms to calculate it. This on the other hand leads to a massive increase in multicollinearity, represented by the general variation inflation factor (GVIF) (Appendix 12.1.) and is therefore not analysed further.

Table 12: Summary of the stepwise logistic regression model with the original variables

```
Coefficients:
                         Estimate Std. Error z value Pr(>|z|)    
(Intercept)            -3.866e+00  4.339e-01  -8.911  < 2e-16 ***
Forest                  6.251e-02  9.923e-02   0.630  0.52871    
Wetland                 2.589e-01  9.607e-01   0.269  0.78756    
Water                  -1.526e+01  1.660e+02  -0.092  0.92674    
Arable land             9.471e-01  9.098e-02  10.410  < 2e-16 ***
Grassland               7.455e-01  1.022e-01   7.295 2.99e-13 ***
Het.agr.land            8.856e-01  1.020e-01   8.686  < 2e-16 ***
Altitude               -2.428e-03  1.815e-04 -13.374  < 2e-16 ***
Rain_20                 1.699e-01  2.106e-02   8.065 7.30e-16 ***
Alternating             2.235e+00  3.154e-01   7.085 1.39e-12 ***
Alternating wet         2.532e+00  3.234e-01   7.830 4.88e-15 ***
Very dry               -1.234e+01  2.448e+02  -0.050  0.95978    
Very dry to dry         2.857e+00  4.602e-01   6.208 5.35e-10 ***
Dry                     2.447e+00  3.165e-01   7.733 1.05e-14 ***
Dry to moderately dry   2.459e+00  5.008e-01   4.910 9.09e-07 ***
Moderately dry          2.001e+00  3.105e-01   6.442 1.18e-10 ***
Moderately dry to med.  1.785e+00  3.701e-01   4.822 1.42e-06 ***
Medium                  1.957e+00  3.111e-01   6.290 3.17e-10 ***
Medium to mod. wet      2.517e+00  3.768e-01   6.680 2.38e-11 ***
Moderately wet          2.209e+00  3.241e-01   6.816 9.33e-12 ***
Moderately wet o wet    1.954e+00  6.617e-01   2.954  0.00314 ** 
Wet                     2.079e+00  3.384e-01   6.145 8.01e-10 ***
Wet to very wet         3.290e+00  8.115e-01   4.054 5.03e-05 ***
Very wet                3.080e+00  4.226e-01   7.288 3.13e-13 ***
Others                  5.837e+00  1.623e+03   0.004  0.99713    
Alluvial soil           1.147e+00  2.762e-01   4.153 3.28e-05 ***
Gley                    9.251e-01  2.943e-01   3.143  0.00167 ** 
Atypical soil           1.661e+00  2.948e-01   5.633 1.77e-08 ***
Brown soil              1.102e+00  2.588e-01   4.257 2.07e-05 ***
Pseudogley              1.059e+00  2.628e-01   4.030 5.58e-05 ***
Soil form complex       6.313e-01  3.310e-01   1.907  0.05650 .  
Randzina and rangier   -3.546e-01  4.929e-01  -0.720  0.47179    
Bog                     1.354e+00  4.132e-01   3.277  0.00105 ** 
Others                 -1.523e+01  7.184e+02  -0.021  0.98308    
Sand                    2.841e-01  6.341e-02   4.480 7.47e-06 ***
Clay                   -1.469e+00  2.715e-01  -5.409 6.33e-08 ***
Silt                    1.872e-01  5.839e-02   3.206  0.00135 ** 
Others                 -1.152e-01  1.728e-01  -0.667  0.50480    
Slope                  -9.550e-03  1.582e-03  -6.036 1.58e-09 ***
Depth                   1.847e-01  3.443e-02   5.364 8.14e-08 ***
Rain_30                 2.580e-01  5.125e-02   5.033 4.82e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Before the optimization of the model is finished, we will look at the GVIF. As a rule of thumb the GVIF (second column) should not be higher than 4 to 10. However, according to Fox and Monette (1992), if the degrees of freedom (Df) are higher than 1 than the weighted GVIF (fourth column) should be used. In this case, the values there should not be greater than 2 to 3. While usually interaction terms have to be added to such a model, including them increases the GVIF far over the accepted threshold (see Annex 12.1.). By centring the variables (subtracting their mean value from them) and excluding variables that increase the GVIF too much, it is possible to include interaction terms in the model (Annex 12.2.). Nevertheless, the resulting model has too few variables, which would mean a loss of quality. So, only models without interaction terms are analysed.

In Table 13, we can see that the water condition and soil type have a very high GVIF with water.con at 52 but they also have more than 1 Df and in their fourth column the GVIF is actually below 2. With that, the pseudo R² of the model is at 11.42% for McFadden's to 27.10% for Nagelkerke's, which is still far from being good (Table 14).

Table 15 shows us the output of the "Anova"-function of the "car"-package (Fox et al. 2018). There, likelihood-ratio chi-square-tests are calculated, which test the goodness of fit of each variable to the model. A higher value at the column "LR Chisq" indicates a better fit and is based on the ratio of observed to expected frequencies (Colman 2009). These are then tested against the Null-hypothesis of no association between the explanatory and the target variable, which every single variable from our model clearly rejects by showing a high significance. With 463.80 the land use provides the best fit followed by the altitude (186.26) and water conditions (176.33). Rain_20 is now on the same level as the soil type and texture at 66 to 68. The slope, depth and Rain_30 provide the worst fit, but they prove to be still significantly associated with pluvial flood damages.

Table 13: GVIF of the stepwise logistic regression model with the original variables

```
                          GVIF Df GVIF^(1/(2*Df))
as.factor(Land.use)   2.010472  6        1.059924
Altitude              1.822650  1        1.350056
Rain_20               1.510185  1        1.228896
as.factor(Water.con) 52.149060 16        1.131524
as.factor(Soil.type) 17.210586  9        1.171265
as.factor(Soil.text)  3.030768  4        1.148667
Slope                 1.284897  1        1.133533
Depth                 2.965824  1        1.722157
Rain_30               1.473060  1        1.213697
```

Table 14: Pseudo R² of the stepwise logistic regression model with the original variables

```
                             Pseudo.R.squared
McFadden                             0.114206
Cox and Snell (ML)                   0.248890
Nagelkerke (Cragg and Uhler)         0.271001
```

Table 15: Anova of the stepwise logistic regression model with the original variables

```
                     LR Chisq Df Pr(>Chisq)
as.factor(Land.use)    463.80  6  < 2.2e-16 ***
Altitude               186.26  1  < 2.2e-16 ***
Rain_20                 66.04  1  4.411e-16 ***
as.factor(Water.con)   176.33 16  < 2.2e-16 ***
as.factor(Soil.type)    67.93  9  3.864e-11 ***
as.factor(Soil.text)    66.94  4  1.004e-13 ***
Slope                   37.53  1  8.981e-10 ***
Depth                   29.79  1  4.806e-08 ***
Rain_30                 25.80  1  3.787e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In our model with aggregated classes, the variables Max_rain_15 and Max_rain_15_mad were also excluded in the stepwise regression, but this time the permeability was included (Table 16). While land use remained the most important variable, water conditions and soil types lost importance (as they were simplified the most). As the remaining soil type, pseudogley was included by the stepwise regression, but does not show any significance as it is slightly higher than the largest significance code of 0.1.

The GVIFs of all variables are lower than the threshold, even the ones with more Dfs (Table 17). The resulting pseudo R² is a bit lower than the first model with 9.68% for the McFadden R² and 23.45% for Nagelkerke's (Table 18).

At the Table 19 of the Anova analysis it becomes clearer how much the water conditions and soil type suffered from the aggregation. From one of the best goodness of fit-values of 176.33 the water conditions dropped to 15.38 and lost a lot of significance. The soil type dropped from 67.93 to 2.58 and lost any previous significance. The newly added permeability shows a goodness of fit of only 11.64 (the second lowest model) but it is more significant than the water conditions. Overall, the aggregation of classes has led to a loss of information and performs worse than the model with original classes. This confirms that the general linear model is appropriate to handle continuous and categorical predictors and requires no aggregation of variables to improve model fitting.

Table 16: Summary of the stepwise logistic regression model with the aggregated variables

```
Coefficients:
                        Estimate Std. Error z value Pr(>|z|)    
(Intercept)           -2.158e-01  2.196e-01  -0.983 0.325812    
Arable land            9.175e-01  5.715e-02  16.055  < 2e-16 ***
Grassland              6.706e-01  6.440e-02  10.413  < 2e-16 ***
Het.agr.land           8.711e-01  6.606e-02  13.187  < 2e-16 ***
Others                -2.394e-01  9.462e-02  -2.530 0.011409 *  
Altitude              -2.053e-03  1.673e-04 -12.273  < 2e-16 ***
Rain_20                1.614e-01  2.058e-02   7.842 4.43e-15 ***
Slope                 -1.038e-02  1.520e-03  -6.831 8.41e-12 ***
Depth                  1.643e-01  2.983e-02   5.507 3.64e-08 ***
Sand                   3.691e-01  6.254e-02   5.902 3.60e-09 ***
Silt                   2.546e-01  5.430e-02   4.688 2.76e-06 ***
Others                -2.245e-01  1.249e-01  -1.797 0.072304 .  
Permeab.              -5.561e-02  1.631e-02  -3.411 0.000648 ***
Rain_30                2.529e-01  5.003e-02   5.056 4.29e-07 ***
Medium                -2.509e-01  8.574e-02  -2.927 0.003426 ** 
Wet                   -9.050e-02  1.097e-01  -0.825 0.409596    
Others                -8.029e+00  1.195e+02  -0.067 0.946415    
Pseudogley             1.035e-01  6.452e-02   1.604 0.108765    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Table 17: GVIF of the stepwise logistic regression model with the aggregated variables

```
                          GVIF Df GVIF^(1/(2*Df))
as.factor(Land.use)   1.727765  4        1.070744
Altitude              1.590844  1        1.261287
Rain_20               1.500480  1        1.224941
Slope                 1.266676  1        1.125467
Depth                 2.378056  1        1.542095
as.factor(Soil.text)  1.929295  3        1.115749
Permeab.              2.329354  1        1.526222
Rain_30               1.473223  1        1.213764
as.factor(Water.con)  3.046012  3        1.203987
as.factor(Soil.type)  1.322107  1        1.149829
```

Table 18: Pseudo R² of the stepwise logistic regression model with the aggregated variables

```
                             Pseudo.R.squared
McFadden                            0.0968373
Cox and Snell (ML)                  0.2154740
Nagelkerke (Cragg and Uhler)        0.2346170
```

Table 19: Anova of the stepwise logistic regression model with the aggregated variables

```
                     LR Chisq Df Pr(>Chisq)    
as.factor(Land.use)    443.23  4  < 2.2e-16 ***
Altitude               155.33  1  < 2.2e-16 ***
Rain_20                 62.32  1 2.914e-15 ***
Slope                   48.14  1 3.965e-12 ***
Depth                   31.24  1 2.284e-08 ***
as.factor(Soil.text)    49.89  3 8.436e-11 ***
Permeab.                11.64  1 0.0006453 ***
Rain_30                 25.95  1 3.497e-07 ***
as.factor(Water.con)    15.38  3 0.0015207 ** 
as.factor(Soil.type)     2.58  1 0.1081224    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

7.1.3. Prediction diagnostics

Before entering into the prediction diagnostic values, it is important to know how the predicted values are distributed. Figure 39 shows probability density plots of both models. It can be noticed that in the second model there are fewer cells with a very low risk but slightly more with a risk greater than 60%, but overall, both models are rather similar.

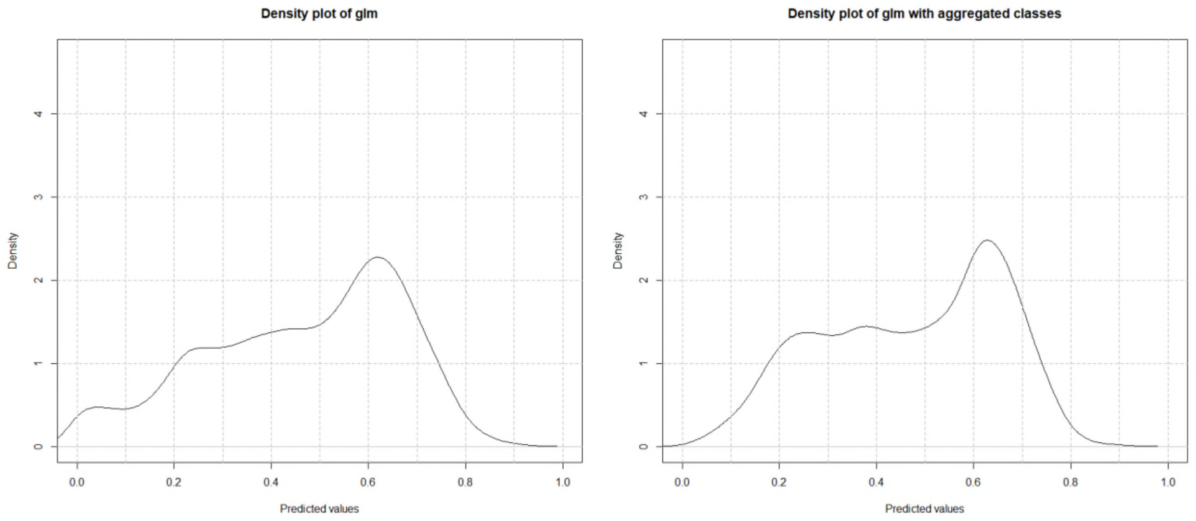


Figure 39: Density-plots of the logistic regressions (left first, right second model)

After the partition into training and test set, the test set is predicted and the results are shown in Table 20. Even with smaller pseudo R² values, the second model with aggregated classes reaches performance values very similar to the first model. The balanced accuracy (AUC) changes only slightly between cut-offs and models. Determining whether their AUCs of ~62% are good or bad depends considerably on the goal of the study. Baldwin (2009), who predicted the distribution of wildlife, classified an AUC lower than 70% as “uninformative”. Whether this classification applies to our study will be assessed with the risk maps a bit later. With a larger cut-off fewer cells are classified as flood damaged and so the sensitivity gets lower. At the same time, the specificity gets larger, because more cells are classified as no-flood damage. While the balanced accuracy stays the same, the accuracy grows with the specificity, because of the imbalance in the flood damage data.

Table 20: Prediction performance of logistic regression

	Logistic Regression				
Cut-off	Accuracy (in %)	Sensitivity (in %)	Specificity (in %)	FAR (in %)	AUC (in %)
0.4	44.97	85.27	38.70	61.30	61.98
0.5	55.45	71.71	52.92	47.14	62.31
0.6	68.28	55.04	70.34	29.90	62.69
	Logistic Regression with aggregated classes				
0.4	45.28	84.50	39.18	60.82	61.84
0.5	55.29	73.26	52.50	47.50	62.88
0.6	67.34	56.59	69.02	30.98	62.80

To better visualize the results, the balanced accuracy can be plotted as a typical ROC (Receiver Operating Characteristics) –curve (Figure 40). If the model would predict each cell correctly, the black line would reach the upper left corner, where the area under the curve (AUC) is 100%. The grey line symbolizes an AUC of 50%, corresponding to a prediction per chance.

Figure 41 shows the predicted risks of pluvial flood damages for the whole study area. The colour scale has been chosen in a way to reflect the rather uniform distribution of the predictor (Fig. 32), by using equally-spaced classes (that have a rather constant frequency as well) and red-shadings above the optimal cut-off-value of 0.5. The resulting probability maps are very similar, but as the density plots already indicated, in the second model there are fewer cells with very low risks. Additionally there seem to be more cells with a higher risk in the centre regions of Upper Austria than in the first model. In the west and east however the risk is calculated lower than in the first model. With this, it cannot be said that the results are "uninformative". Even when 62% is not a high AUC value, the model manages to identify high risk regions.

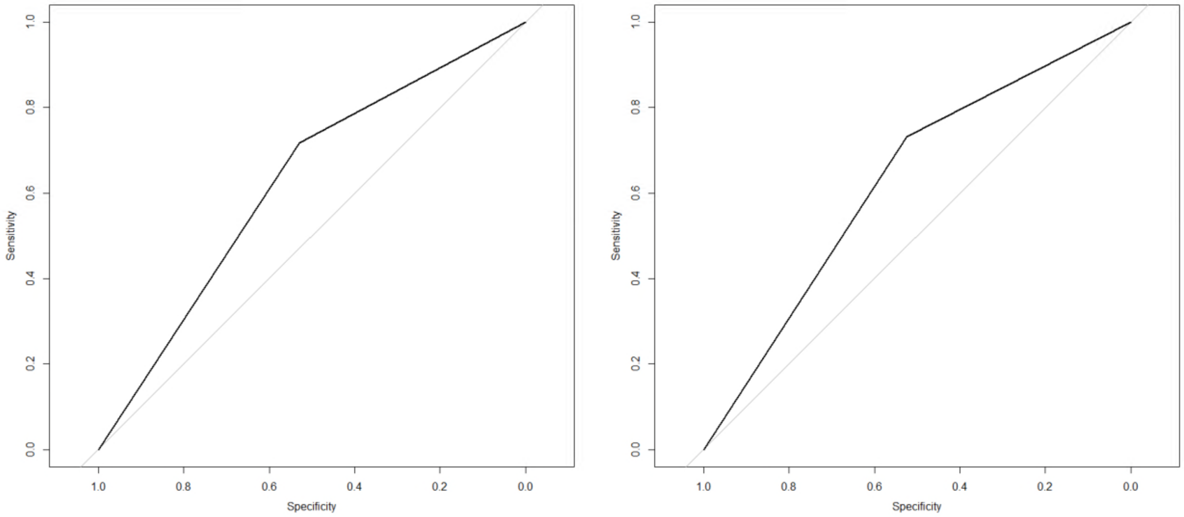


Figure 40: ROC-plots of the logistic regressions at the 0.5 cut-off. Left panel: first maodel, right panel: second model

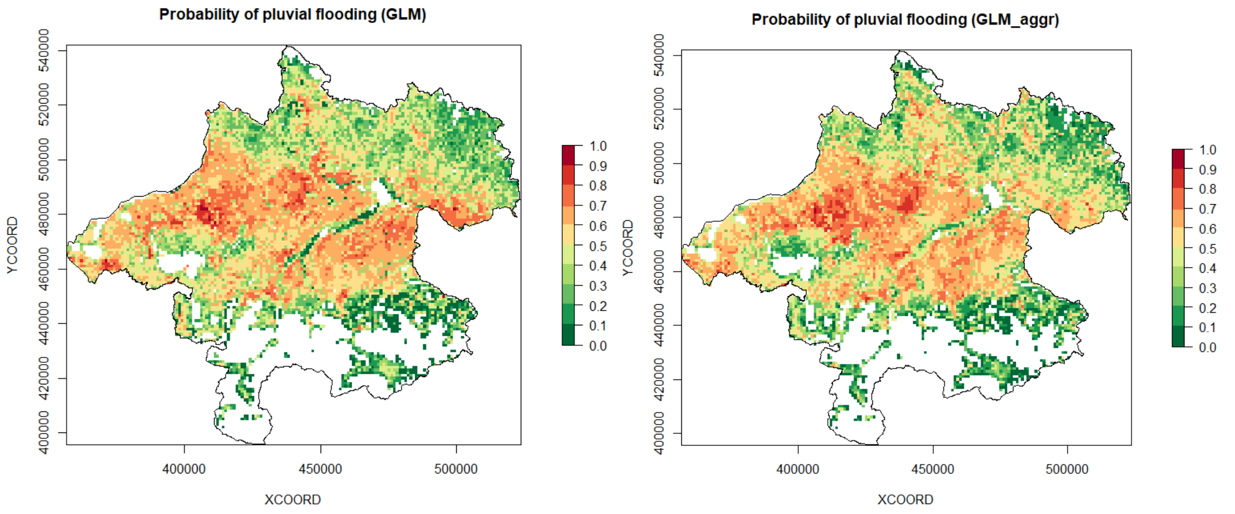


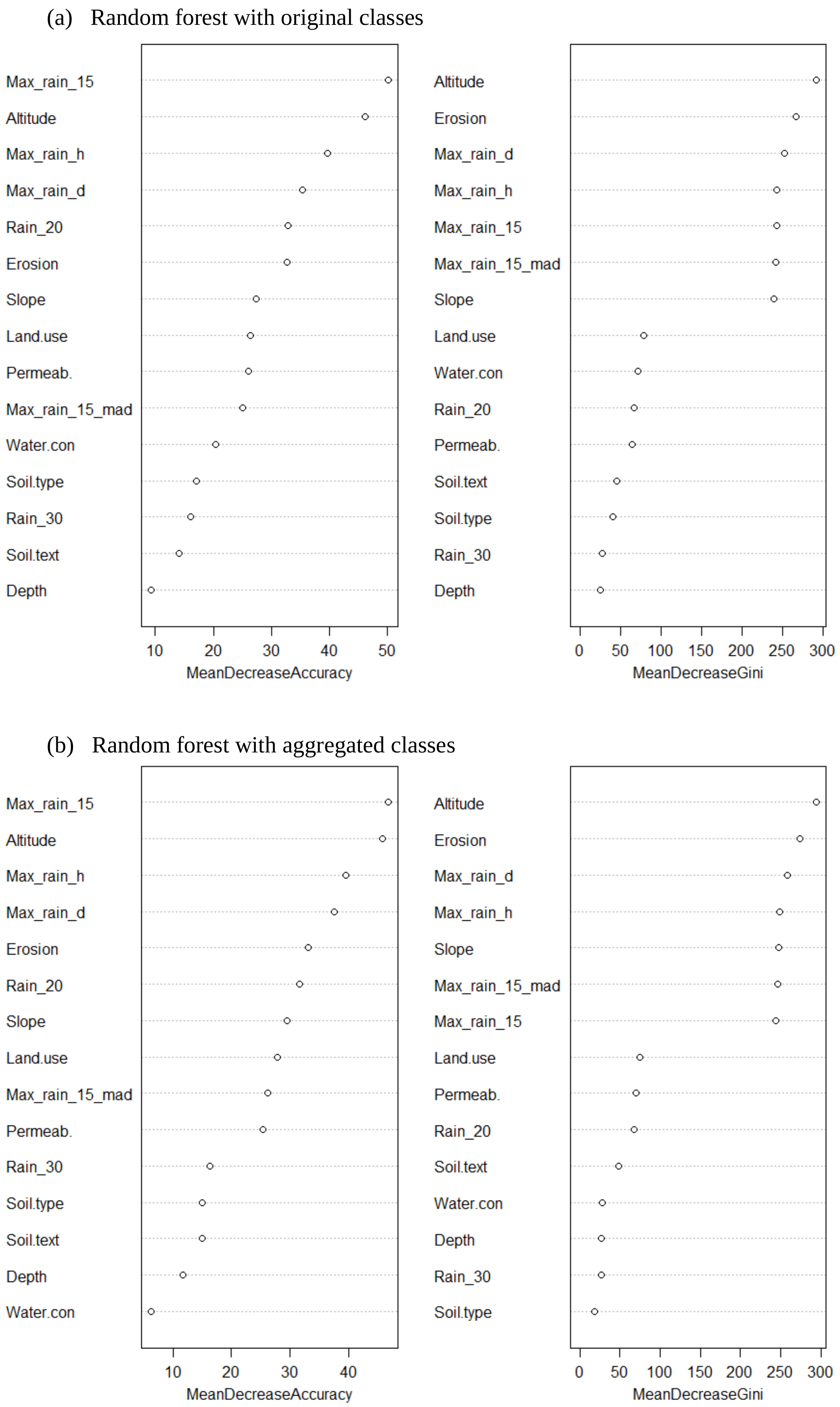
Figure 41: Pluvial flood probability map of the first (left) and second regression model (right).

7.2. Random forest

7.2.1. Model optimization

In random forests, the contribution of each variable to the output of the model cannot be determined as easily as in logistic regression. There are two methods in the randomForest-package to calculate the importance of each variable. Both of them have their pros and cons. The first measure is the permutation importance (Figure 42, first and third panel). It is calculated by permuting (i.e. randomly shuffling) the variable and measuring the difference in prediction accuracy before and after the permutation. The idea behind this approach is that the prediction accuracy must decrease greatly if the values of an important variable are randomly mixed. However, this importance measure might be biased in case of correlated predictors (Strobl et al. 2008). The second importance measure, shown in the right panels of Figure 3, is called impurity or "Gini" importance, and is based on the so-called "Gini" index measuring the mean of the individual trees' improvement in the splitting criterion produced by each variable. If one variable is important, the Gini importance index will decrease more than for other, less important variables. The main disadvantage of this approach is that it is biased in favour of continuous variables and variables with more categories (Strobl et al. 2007).

As it can be seen from Figure 42, four out of the six most important variables are precipitation variables. To lower the correlations among precipitation variables for the permutation importance (first and third panel), it was tested how the model behaves when excluding Max_rain_15 and Max_rain_15_mad as in the stepwise logistic regression model. But this had only marginal effects on permutation importance. Apart from precipitation indices, altitude, erosion and slope are highly ranked by both importance measures. Interestingly, land use is only in the middle of the ranking. Water conditions and soil parameters (permeability, type, texture, and depth) appear at the end of the ranking and seem to have little effect on the predictions. Finally, the second model based on aggregated classes (third and fourth panel) doesn't show many differences compared to the first model based on original classes (first and second panel).



(a) Random forest with original classes

(b) Random forest with aggregated classes

Figure 42: Predictor significances of the first RF-model with original classes (a) and the second RF-model with aggregated classes (b)

7.2.2. Prediction diagnostics

The prediction diagnostics for the random forest can be handled in the same way as in the logistic regression. At the density plots (Figure 43), the downside of not using weights is illustrated. Only very few cells get votes for the flood damage classification, so the cut-offs have to be chosen lower.

Without weights the balanced accuracies vary more between each cut-off (Table 21). The highest balanced accuracy (67.38%) can be achieved at the cut-off of 0.15, which is 4-5% higher than for the logistic regression, with a sensitivity of 67.90% and a specificity of 66.86%. For comparison, the FAR is 33.14%, showing that the risk of false alarms is rather low. The other prediction performance values are quite similar to the ones of the regression models: the sensitivity drops with higher cut-off, while specificity and accuracy grow. All in all, the first RF-model reaches slightly better values, but the differences are rather small.

The AUCs are illustrated in Figure 44. Again, there can be no real difference spotted between the two models. Compared to the regression models the AUC is slightly larger and the corner of the "curve" is located more at the bottom.

The risk maps also have to be adjusted to the unbalanced data, because without weights, a calculated probability of 30% is already considered high (Figure 45). A colour scale comparable to the one for logistic regression has been obtained by using equally-frequent classes and a red-shading above the optimal cut-off-value of 0.15. Due to the ability of including Max_rain_d, the region of the 2013 event has a high risk too. Apart from that, the risk maps are very similar. The highest risks are located at the locations of the events of 2008 and 2011 and the mountainous regions have a very low risk.

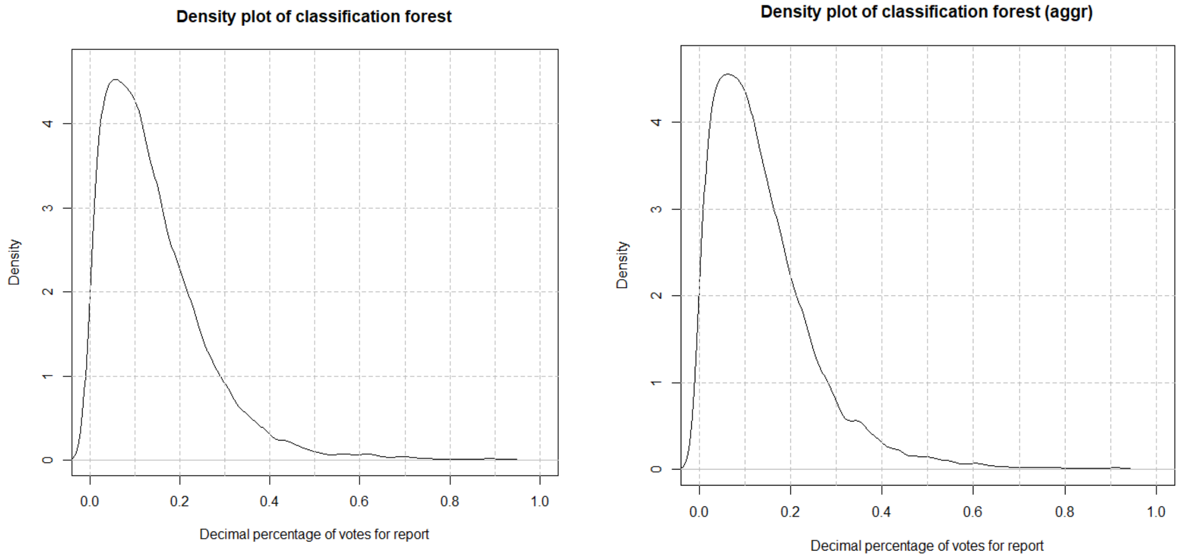


Figure 43: Density-plots of the first (left) and second RF-model (right)

Table 21: Prediction performance of classification forest. Best model marked in red.

	Classification forest				
Cut-off (in %)	Accuracy (in %)	Sensitivity (in %)	Specificity (in %)	FAR (in %)	AUC (in %)
10	52.75	83.45	47.96	52.04	65.71
15	67.00	67.90	66.86	33.14	67.38
20	76.66	52.20	80.48	19.52	66.34
30	84.78	26.06	93.94	06.06	60.00
	Classification forest with aggregated classes				
10	52.77	84.84	47.77	52.23	66.30
15	69.39	64.42	70.16	29.84	67.29
20	77.80	52.82	81.70	17.68	67.26
30	85.04	26.14	94.23	05.98	60.19

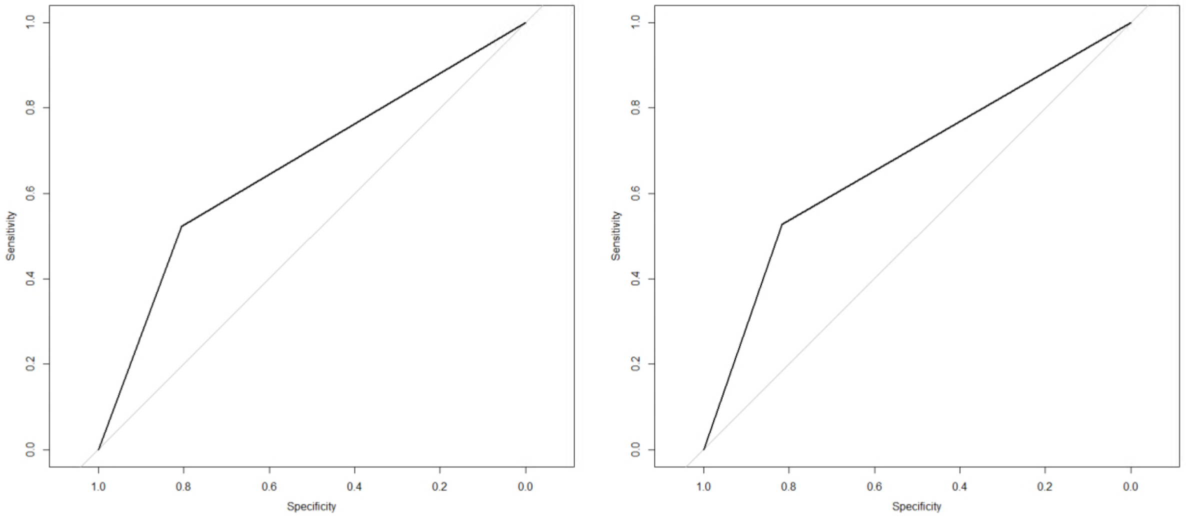


Figure 44: ROC-plots of the RF-models (left first, right second model) at the 0.2 cut-off

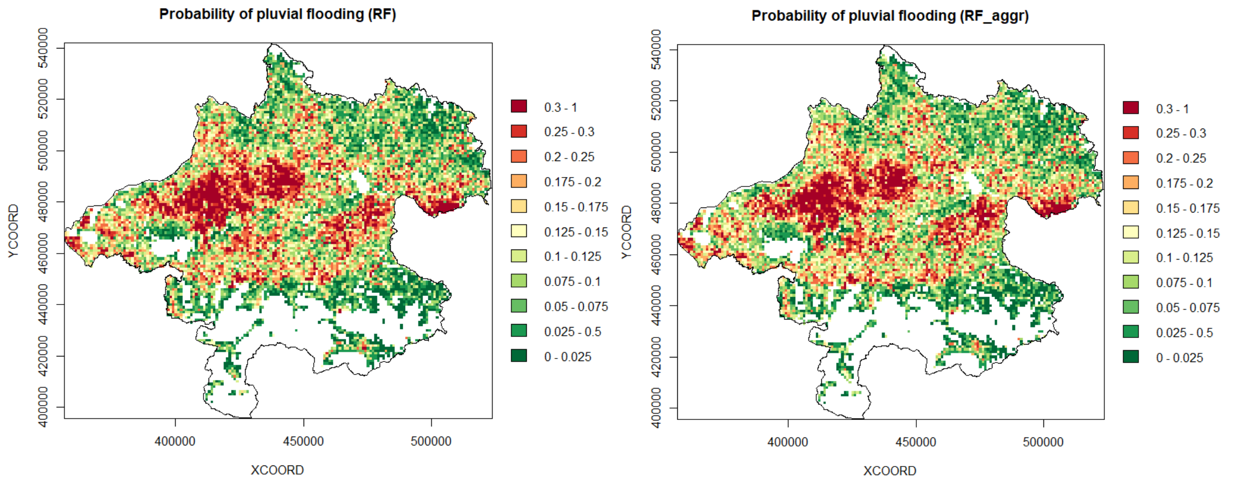


Figure 45: Pluvial flood probability map of the first (left) and second RF-model (right)

8. Discussion and conclusions

8.1. Discussion of the approach

Our analysis is based on insurance claims from the Austrian Hail Insurance. Zischg et al. (2018) tested insurance claims against inundation modelling and concluded that using these claims has a lot of advantages over inundation modelling. Damage claims were also used as target variable by Abebe, Kabir, and Tesfamariam (2018) and Bernet, Prasuhn, and Weingartner (2017). While the latter didn't find an optimal solution on dividing pluvial and fluvial floods and in the end used the distance to the next river or lake as indicator, we used a visual determination based on shape around rivers and mostly 'DORIS Atlas 4.0' (2018).

Importance of predictor variables

The best performing regression model consists of nine predictors, which have been automatically selected by the stepwise regression algorithm (Table 12). Their order in the regression equation, therefore, corresponds to the relative importance of catchment characteristics in terms of predictive performance. However, the importance for predictive performance may not be seen as a straightforward evaluation of process controls, because of intercorrelations between catchment characteristics, different accuracy of catchment characteristics and other effects on model fitting. We therefore included information of single effect models (Table 9) and heat map analysis (Figure 38) in the interpretation to partially account for these effects.

From the results, land use, which is represented by six classes, is the most important linear predictor. The most significant class is Arable land, followed by heterogeneous agricultural land (Het.agr.land) and Grassland. All correspond to agricultural areas and show an increased flood risk. As our pluvial flood data is based on reports of flooded agricultural land these predictors do not contribute much to the interpretation on how they occur, they rather lead the model on the right path. Altitude is the next important predictor. It has a negative effect on flood risk, indicating that elevated areas are less prone to pluvio-flood damage, as one would expect.

From the number of precipitation characteristics, the number of heavy, short-term (15 min) rainfall events exceeding 20mm (Rain_20) is the next important predictor. It is strongly correlated with the maximum 15 minute precipitation (Max_rain_15), which is masked in the stepwise model, and quite strongly correlated with the number of 30 mm events (Rain_30) as well. These variables indicate high rainfall intensities, which strongly increase the risk of pluvial flooding, as opposed to persistent low intensity rainfall, which are not selected in the model.

Water conditions (Water.con) are represented by 16 classes. While at the location analysis and single models it seem that wet soils increase the pluvial flood risk, the logistic regression gives every class a positive sign and similar estimate, except from the very dry soils which decrease the flood risk. By aggregating water conditions, it lost a lot of its significance, giving it to depth and permeability. So, even if it was excluded from the model, the other two variables would compensate some of the explanatory power and they are less ambiguous.

Soil type (Soil.type) consists of nine classes, but not all of them have a significant effect on flood risk. Likely, some of the classes are too rare in the study area to be considered by the model. From all classes, Alluvial soil, Gley, Atypical soil, Brown soil, and Pseudogley have a significantly positive effect and therefore tend to increase flood risk. Soil texture (Soil.text) consists of four classes. Sand and silt increase flood risk, while clay decreases flood risk. These effects do not correspond with our expectation from a process perspective and may be caused by their co-occurrence with other landscape characteristics. Hence, this variable needs to be interpreted with caution.

The remaining two significant variables measure catchment characteristics on a continuous scale. Slope has a strong negative effect on flood risk, indicating that flat areas are most prone to pluvio-flood damage, which corresponds well with hydrological expectation. However, soil depth has a positive effect on pluvio-flood damage. In the model equation, it seems rather an indicator of good quality agricultural areas than representing the effect of soil conditions on flood damage.

Besides a slightly different order of the variables, there is no big difference between the two models in the logistic regression. The main difference is that when aggregating water conditions the variable gets less significant and is replaced by permeability and depth. The second model has a slightly lower pseudo R^2-value, but at the pluvial flood prediction they are very similar. Overall, land use, topography, short-term (15 min) precipitation intensity and soil type are the most important characteristics in the logistic regression model. In addition, either soil water conditions or permeability and depth are significant predictors of pluvial flood risk.

The random forest model shows quite similar results. According to the permutation importance index, peak rain intensity (Max_rain_15) and altitude are the most important predictors. Hourly and daily rain sum (Max_rain_h and Max_rain_d) are the next important predictors, followed by erosion and slope. Interestingly, land use is only in the middle of the ranking and has a much lower effect than in logistic regression. Water conditions and soil parameters (permeability, type, texture, and depth) appear at the end of the ranking and seem to have little effect on the predictions. According to the Gini importance index, altitude and erosion are higher ranked as the precipitation indices. Apart from that, the importance of predictors is quite similar.

Predictive performance

We now assess to which degree our models are appropriate for performing a risk-mapping for the occurrence of pluvio-flood damage in Upper Austria. The prognoses maps of logistic regression and Random Forests have been shown in Figure 42 and 45. Both approaches lead to very similar patterns. High risk predictions are mostly located on arable land and in areas that were hit by more than one extreme rain event, which corresponds well to our expectations. Regions with high risks are located around clusters of reported cells. This means that on a large scale, the models are able to identify general regions with high risks, and predict these risks to surrounding areas with similar landscape characteristics. The location of risk-areas depends on rainfall characteristics, which constitute a climatic feature of the study area. By the event analysis and statistical modelling, intense or long-lasting rainfall events have been identified as the first-order controls of pluvio-flooding. Extreme precipitation events are necessary to trigger the genesis of pluvio-flood events. Landscape characteristics can be seen as a second-order control, which determines local infiltration and runoff. Locations with low infiltration and runoff are prone to a high pluvio-flood risk. The predictors used in both models reflect these controls, and can be well interpreted on hydrological grounds.

Predictions may also be sensitive to linearity / nonlinearity of predictors and scale of measurements. The finding that both models lead to similar prediction shows the robustness of the predictions. By comparing logistic regression and random forests, we find that the predictors are rather insensitive to linearity /nonlinearity of relationships. The non-linear random forests show a slightly better predictive performance and should be preferred. We also assessed the possible influence of variable scales on predictive performance. The models with aggregated classes performed very similar to the models with original classes, hence the predictors are rather insensitive to scales of the chosen classification. Overall, the prediction maps provide robust estimates of pluvio-flood risk for the observation period.

Despite these favourable characteristics, the predictive performance scores are rather modest. With a Nagelkerke R^2 of only 23% the logistic regression model explains only a small part of the variability of observed flood / non-flood damage cells in the cross-validation, and also its AUC performance of about 62% is rather modest. In a similar way, the random forest reaches an AUC of 67% which is also not huge. It is surprising that the models which yield such robust predictions and whose predictors are all well interpretable in accordance with process reasoning lead to relatively modest performance scores. We think that the reason for the low scores is rather due to the short observation period. With a longer observation period, critical rainfall events would have happened in other grid cells leading to an increased number of flood damaged cells. Using a subset of these cells corresponding to seven years of observation period constitutes an observational bias in the dependent variable that leads to two effects. First, the models are trained by a number of cells that are (on long-term) falsely classified as no-flood damage. This may make parameter estimates sub-optimal. The second effect is that true positive predictions could be in a cell which has not been flood damaged yet, but would be flood damaged in a few years. The theoretically true predictions are wrongly classified as false alarms, yielding to lower performance scores than being actually the case. Both effects are consequences of the observation bias which may explain the discrepancy of good prediction diagnostics and low performance scores. Because of that, we cannot say how well our risk map describes the actual risks. However, the magnitude of this bias can be roughly estimated from the effect of the number of flood damage cells on the coefficient of determination, defined as the ratio of the explained variance of the model by the total variance of the binary dependent variable (flood damage/no flood damage). When increasing the number of flood damage cells from 1/7 to 2/7 of cells within the study area, the total variance of the dependent variable increases from 0.122 to 0.204. This would increase the R^2 to a value of 56% when assuming that the model equation and the error variance (currently 0.089) remains unchanged. In reality, a more reliable classification of flood cells obtained by a longer observation period would also improve the parameter estimates, which would yield to a further increase of model performance.

An indication about the appropriateness of assuming twice the number of flood cells for long-term observation is given by the temporal development of the number of flood damage cells during the seven years of observation shown in Figure 46. From the curve, the number of flood damage cells has not stabilised and further increases can be expected in future years. To what extend the number of flood events will increase is, however, subject to speculation. The last value actually corresponds to the major 2013 flood event and could be interpreted as a peak at the end of the curve, thus representing an upper limit of the flood-affected area. The fact that the year 2013 was only moderate in terms of reported pluvial flood damage contradicts this interpretation and suggests that the seemingly stabilisation in the years before 2013 is simply caused by climate variability. In fact, the flat part of the curve corresponds to the three years with the lowest reported damages of the observation period. Overall, the observation period is certainly too short to safely judge about the future development of pluvial flood events and herby affected areas. This lack of data constitutes an observational bias that affects model fitting by both, an incorrect classification of presence/absence cells, and lacking information of process combinations of future pluvial flood events. Longer records are required to judge pluvial flood risks more safely.

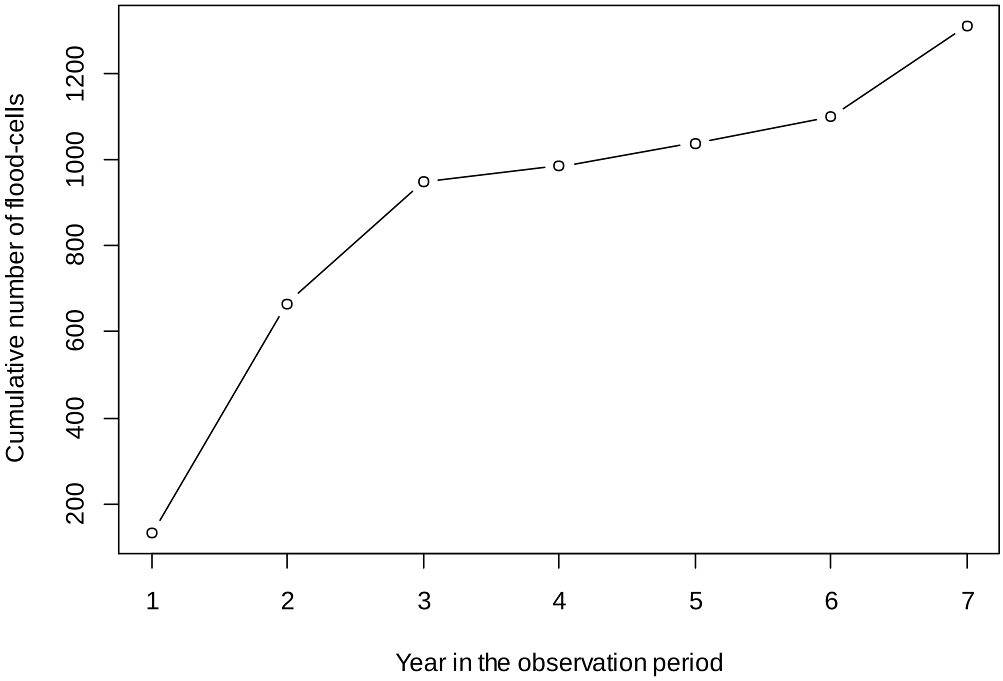


Figure 46: Development of flood-cells over the observation period

Models compared to the literature

It is now interesting to compare the models to the literature. In comparison with the logistic regression, the random forest performs slightly better and reaches a balanced accuracy of 66-67%. When a cut-of value of 0.15 is chosen, the model with aggregated classes reaches a sensitivity (true alarm rate) of 68% with a false alarm rate of 33%. For comparison, Bernet et al. (2018) reached a maximal sensitivity (termed hit rate in their paper) of 79% and Zischg et al. (2018) reached sensitivities of 61-91% with false alarm ratios of 20-34%, depending on the catchment. While these studies use physically-based models at the event scale, our study uses statistical models on a regional scale. But overall, the performances are quite similar. A possible explanation is that all studies depend on climate and catchment characteristics. These characteristics are prone to measurement errors, which limit the performance of statistical and process-based models.

In this study it was found that with the currently available data pluvial flood damages are more likely to occur on agricultural land with a low altitude and low slopes. These were also two of the main factors identified by Abebe, Kabir, and Tesfamariam (2018). As a general rule, these flood damages are caused by short, heavy rainfalls, however, long lasting rainfalls might as well lead to pluvial flood damages in basin locations. An example of a pluvio-flood damage triggered by a long-lasting rainfall event is the event of 2013.

While the study design seems to be well suited to analyse the pluvial-flood damage risk, the accuracy of the method suffers from the short observation period. Both models largely depend on the rain events that occurred in these seven years. With a longer time period rain events would most probably also happen in grid cells, for which our models now predicted a low risk. Because of that, we cannot say how well our risk map describes the actual long-term flood risks of the study area. Another point is that some possibly important variables could not be used, like the macro relief or groundwater data from the BORIS dataset, due to their low spatial resolution (Section 3.2.2). Adding it to our final dataset would have reduced our dataset by about 1000 grid cells, which was not tolerable. Additionally, a generalisation beyond the study area may be difficult as many of the variables used are specifically collected or modelled in Austria. Other countries might have other data available. The models themselves worked as intended. Even when some variables got rescaled and the level of detail got reduced, the risk models and plots didn’t change much, which indicates a high robustness of the models. High intercorrelations among climate and catchment characteristics posed a major challenge to modelling, which made interpretations of predictors difficult. Variable selection based on heat maps, stepwise regression and collinearity diagnostics and different types of predictor importance measures were used to address these problems and helped to identify robust predictors.

8.2. Summary and conclusions

In this study, we wanted to know how pluvial flood damages occur and if certain locations are more vulnerable than others. For the former we looked at the five biggest events and investigated the precipitation characteristics. For the latter a presence/absence raster was created and each location variable was parted in report and non-report. Finally, we used a linear and a non-linear statistical model to find out which variables are important and if it is possible to use them for creating a risk map.

The basis for the event analysis was a high-resolution precipitation raster with a mesh size of 1 km and a temporal resolution of 15 minutes. Despite of its high resolution, the measurements are still prone to considerable errors, especially for intense rain events in greater distance from a gauge (Section 3.2.1). This hinders an accurate appraisal of the severity of events in terms of their return periods or occurrence probabilities. Plotting the number of damage reports by year and month allowed us to identify the five biggest events, which caused the most damages during our investigation period. While some events occurred in a season with above-average precipitation, other events occurred in a season with average or below-average precipitation. It turned out that the events of 2007, 2008 and 2011 were based on short but very heavy rainfall events (10-45l/m^2 in 15 minutes), which happened during a month that was not particularly rainy. The events of 2009 and 2013 have shown some rainfalls prior to the highest number of damage reports. The damages of 2009 were also caused by short and heavy precipitation (23.9l/m^2 in 15 minutes) and the ones of 2013 by long-lasting, persistent but low intensity rainfall (0.9-2.4l/m^2 in 15 minutes). The rainfall of June 2013 seems to have left an impact on the mean rain sums, as the month was especially rainy. On the other hand, June 2009 was also very rainy, but the event, which caused to damage claims, occurred in July, which didn't stand out in monthly rain sums. So, most of the events consisted of short but heavy rainfalls, but especially the event of 2013 caused the most damage reports and consisted of a long lasting, low intensity rainfall.

The location analysis has shown us that the flood damaged cells on which the detailed event analysis is based can be very different. The event that varied most from the others was the one of 2013. There was the least amount of heavy rainfall and it showed large differences in altitude and slope, but only small differences in erosion. For these flood damaged cells, the variables of permeability, depth and water conditions more likely described an ideal location for agriculture under normal conditions rather than pluvial flood risks. On the whole study area there are many indicators for locations with an increasing pluvial flood risk. These are: low altitudes, low slopes, high erosion, arable land, silt, low soil permeability and high soil depth. At the end of that chapter, some classes were aggregated due to their low numbers or to give more focus on other classes.

Before the logistic regression, the variables of erosion and Max_rain_d were already excluded due to non-linearity. At the beginning of the logistic regression, each variable is tested on significance and pseudo R^2 of their single models. Additionally, with the heat map it was possible to identify strong correlations and dependencies between the variables. After excluding Max_rain_h due to non-significance and Max_rain_15, Max_rain_15_mad and permeability, because of high correlations, the stepwise regression model was fitted. At first, interaction terms were included, but they increased the GVIF far over the acceptable threshold and therefore could not be included in further modelling. This led to a broader interpretation, concentrating only on significances and signs and not on accurate calculations with the estimates. The model was then fitted without interaction terms, but then correlations had to be taken into account in the interpretation. The most important variables turned out to be land use (agricultural land), altitude (with negative sign) and Rain_20 (with positive sign). This order didn't change with the aggregated classes. Most noticeable was the drop of importance of the water conditions and the integration of the permeability. Still, both of these models showed very similar prediction performances at a balanced accuracy of 62-63%. On the risk map, the focus of high risks lies on arable land that suffered heavy rain events and was on a low altitude.

In comparison, the random forest, which is able to include non-linear variables, ranked Max_rain_d very high. However, the accuracy importance measure suffers under the interactions between the variables. Therefore, some highly correlated variables are ranked very high and close together, like the heavy rain variables. Due to these correlations the interpretation of significances is more difficult. Nevertheless, it can be said that high intensity rainfalls increase the risk as well as a low altitude, high daily rain sums, agricultural land and low permeability. Compared to the logistic regression, the daily rain sums are new and the permeability is ranked higher than the water conditions (even before aggregating). The result can be seen at the risk map, where the location of the event of 2013 is now highlighted as high risk, which was not the case at the logistic regression. The random forest also managed to get better prediction results at a balanced accuracy of 66-67%.

Nevertheless, the interpretation of the variables was still easier at the logistic regression due to the stepwise approach. The random forest has the big advantage that it can include all variables, but the high correlations made the interpretation more difficult than in the regression models. Both models calculated high risks in the same regions even with some variables aggregated, which indicates a high robustness of the models. The reason, why they scored such a low balanced accuracy, lies more at the small observation period and possibly wrong reports than the models.

The question whether statistical models are able to identify high risk areas cannot be answered definitely. Our models managed to predict certain regions with high risks, but didn’t manage to get a high prediction performance. With a wider observation period and additional predictors like groundwater or macro relief the answer will be more clearly. With a larger observation period, using count data to predict the risks can prove to be more beneficial as well.

8.3. Outlook

The analysis showed that the models appear well suited to represent the regional flood risk in Upper Austria. However, the models possibly suffer from the relatively short observation period, which could make predictions conditional to events in the observation period rather than presenting the long-term flood risk. In a future study, we would like to further assess the conditional dependence of predictors on the observation period, by segmented modelling or simulation. It would also be interesting to repeat the study once additional years of flood event data are available to analyse the sensitivity of the predictors to an augmented observation period. Knowing this sensitivity may be important to evaluate a possible risk mapping that may be obtained by applying the predictive models of this study to long-term climate indices, in order to characterise long-term pluvial flood risk in Upper Austria.

9. References

Abebe, Yekenalem, Golam Kabir, and Solomon Tesfamariam. 2018. 'Assessing Urban Areas Vulnerability to Pluvial Flooding Using GIS Applications and Bayesian Belief Network Model'. *Journal of Cleaner Production* 174 (February): 1629–41. https://doi.org/10.1016/j.jclepro.2017.11.066.

APA/Red. 2013. 'Hochwasser 2013 in Österreich: Alle Infos Und Bilder'. Vienna.At. 5 June 2013. https://www.vienna.at/hochwasser-2013-in-oesterreich-alle-infos-und-bilder/3595784.

Baldwin, Roger A. 2009. 'Use of Maximum Entropy Modeling in Wildlife Research'. *Entropy* 11 (4): 854–66. https://doi.org/10.3390/e11040854.

Bernet, Daniel B., Volker Prasuhn, and Rolf Weingartner. 2017. 'Surface Water Floods in Switzerland: What Insurance Claim Records Tell Us about the Damage in Space and Time'. *Natural Hazards and Earth System Sciences* 17 (9): 1659–82. https://doi.org/10.5194/nhess-17-1659-2017.

Bernet, Daniel B., Andreas Paul Zischg, Volker Prasuhn, and Rolf Weingartner. 2018. 'Modeling the Extent of Surface Water Floods in Rural Areas: Lessons Learned from the Application of Various Uncalibrated Models'. *Environmental Modelling & Software* 109 (November): 134–51. https://doi.org/10.1016/j.envsoft.2018.08.005.

BFW. 2013. 'Einführung in die bodenkundlichen Grundlagen'. https://bfw.ac.at/300/pdf/Einfuehrung_Bodenkartierung.pdf.

BMLFUW. 2007. *Hydrologischer Atlas Österreichs*. 3rd ed. Wien: Bundesministerium für Land- und Forstwirtschaft, Umwelt und Wasserwirtschaft.

Breiman, Leo. 2001. 'Random Forests'. *Machine Learning* 45 (1): 5–32. https://doi.org/10.1023/A:1010933404324.

Breiman, Leo, Adele Cutler, Andy Liaw, and Matthew Wiener. 2018. *RandomForest: Breiman and Cutler's Random Forests for Classification and Regression* (version 4.6-14). https://CRAN.R-project.org/package=randomForest.

Chen, Chao, Andy Liaw, and Leo Breiman. 2004. 'Using Random Forest to Learn Imbalanced Data', January, 12.

Colman, Andrew M. 2009. *A Dictionary of Psychology*. Oxford University Press. http://www.oxfordreference.com/view/10.1093/acref/9780199534067.001.0001/acref-9780199534067.

'DORIS Atlas 4.0'. 2018. 2018. https://www.doris.at/viewer/(S(2n4cgwgjne4q3xikogly1qfo))/init.aspx?ks=alk&karte=wage&t=636678803443863171.

ESRI. 2016. *ArcGIS* (version 10.3). United States of America. https://www.arcgis.com/index.html.

Falconer, R.H., D. Cobby, P. Smyth, G. Astle, J. Dent, and B. Golding. 2009. 'Pluvial Flooding: New Approaches in Flood Warning, Mapping and Risk Management'. *Journal of Flood Risk Management* 2 (3): 198–208. https://doi.org/10.1111/j.1753-318X.2009.01034.x.

Fox, John, and Georges Monette. 1992. 'Generalized Collinearity Diagnostics'. *Journal of the American Statistical Association* 87 (417): 178–83. https://doi.org/10.1080/01621459.1992.10475190.

Fox, John, Sanford Weisberg, Brad Price, Daniel Adler, Douglas Bates, Gabriel Baud-Bovy, Ben Bolker, et al. 2018. *Car: Companion to Applied Regression* (version 3.0-2). https://CRAN.R-project.org/package=car.

Guerreiro, Selma B., Vassilis Glenis, Richard J. Dawson, and Chris Kilsby. 2017. 'Pluvial Flooding in European Cities—A Continental Approach to Urban Flood Modelling'. *Water* 9 (4): 296. https://doi.org/10.3390/w9040296.

Haiden, T., A. Kann, C. Wittmann, G. Pistotnik, B. Bica, and C. Gruber. 2011. 'The Integrated Nowcasting through Comprehensive Analysis (INCA) System and Its Validation over the Eastern Alpine Region'. *Weather and Forecasting* 26 (2): 166–83. https://doi.org/10.1175/2010WAF2222451.1.

Hosmer, David W Jr, and Stanley Lemeshow. 2000. *Applied Logistic Regression - Second Edition*. 2nd ed. New York: John Wiley & Sons.

'Land Oberösterreich - Administrative Gliederung'. n.d. Accessed 11 July 2018. https://www.land-oberoesterreich.gv.at/147155.htm.

‘Land Oberösterreich - Klima in Oberösterreich’. n.d. Accessed 11 July 2018. https://www.land-oberoesterreich.gv.at/18479.htm.

Liaw, Andy, and Matthew Wiener. 2002. ‘Classification and Regression by RandomForest’ 2: 6.

Mangiafico, Salvatore. 2019. *Rcompanion: Functions to Support Extension Education Program Evaluation* (version 2.1.1). https://CRAN.R-project.org/package=rcompanion.

R Core Team. 2017. *R: A Language and Environment for Statistical Computing* (version 3.4.3). Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org.

Rainman. n.d. ‘Rainman Leaflet’. Accessed 20 August 2018. https://www.interreg-central.eu/Content.Node/Rainman-Leaflet-en.pdf.

Skougaard Kaspersen, Per, Nanna Høegh Ravn, Karsten Arnbjerg-Nielsen, Henrik Madsen, and Martin Drews. 2017. ‘Comparison of the Impacts of Urban Development and Climate Change on Exposing European Cities to Pluvial Flooding’. *Hydrology and Earth System Sciences* 21 (8): 4131–47. https://doi.org/10.5194/hess-21-4131-2017.

Sörensen, Johanna, and Shifteh Mobini. 2017. ‘Pluvial, Urban Flood Mechanisms and Characteristics – Assessment Based on Insurance Claims’. *Journal of Hydrology* 555 (December): 51–67. https://doi.org/10.1016/j.jhydrol.2017.09.039.

Spira Yvonne. 2018. ‘Daten Hagelversicherung’, 24 May 2018.

Strobl, Carolin, Anne-Laure Boulesteix, Thomas Kneib, Thomas Augustin, and Achim Zeileis. 2008. ‘Conditional Variable Importance for Random Forests’. *BMC Bioinformatics* 9 (1): 307. https://doi.org/10.1186/1471-2105-9-307.

Strobl, Carolin, Anne-Laure Boulesteix, Achim Zeileis, and Torsten Hothorn. 2007. ‘Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution’. *BMC Bioinformatics* 8 (1): 25. https://doi.org/10.1186/1471-2105-8-25.

Uibner Florian. 2018. ‘5 Jahre danach: Hochwasserschutz im Eferdinger Becken’. meinbezirk.at. 6 June 2018. https://www.meinbezirk.at/grieskirchen-eferding/c-lokales/5-jahre-danach-hochwasserschutz-im-eferdinger-becken_a2664494.

Wang, Zhaoli, Chengguang Lai, Xiaohong Chen, Bing Yang, Shiwei Zhao, and Xiaoyan Bai. 2015. ‘Flood Hazard Risk Assessment Model Based on Random Forest’. *Journal of Hydrology* 527 (August): 1130–41. https://doi.org/10.1016/j.jhydrol.2015.06.008.

Zahnt, Nina, Markus Eder, and Helmut Habersack. 2018. ‘Herausforderungen durch pluviale Überflutungen – Grundlagen, Schäden und Lösungsansätze’. *Österreichische Wasser- und Abfallwirtschaft* 70 (1–2): 64–77. https://doi.org/10.1007/s00506-017-0451-7.

Zischg, Andreas Paul, Markus Mosimann, Daniel Benjamin Bernet, and Veronika Röthlisberger. 2018. ‘Validation of 2D Flood Models with Insurance Claims’. *Journal of Hydrology* 557 (February): 350–61. https://doi.org/10.1016/j.jhydrol.2017.12.042.

10. List of figures

11. List of tables

12. Annex

12.1. GVIF of logistic regression with interaction terms

```
                            GVIF Df GVIF^(1/(2*Df))
as.factor(Land.use)     2.030289  6        1.060791
Altitude                3.302706  1        1.817335
Rain_20                 1.508753  1        1.228313
as.factor(Water.con) 53.078285 16        1.132149
as.factor(Soil.type) 17.768493  9        1.173343
as.factor(Soil.text)  3.089783  4        1.151439
Slope                16.146844  1        4.018314
Depth                 2.972931  1        1.724219
Rain_30               1.470836  1        1.212780
Altitude:Slope       19.065746  1        4.366434
```

12.2. GVIF of logistic regression with successful interaction terms

```
                                GVIF Df GVIF^(1/(2*Df))
as.factor(Land.use)         2.830919  4        1.138914
Rain_30                     7.859832  1        2.803539
Slope                       2.452103  1        1.565919
Permeab.                    3.860027  1        1.964695
Max_rain_15                 1.521284  1        1.233403
as.factor(Land.use):Slope   4.766032  4        1.215541
as.factor(Land.use):Permeab. 5.136856  4        1.226979
as.factor(Land.use):Rain_30 4.810234  4        1.216945
Rain_30:Permeab.            1.576794  1        1.255705
Slope:Permeab.              1.223370  1        1.106060
Permeab.:Max_rain_15        1.495757  1        1.223011
Rain_30:Max_rain_15         3.217693  1        1.793793
```

```
                             Pseudo.R.squared
McFadden                            0.0990865
Cox and Snell (ML)                  0.2198840
Nagelkerke (Cragg and Uhler)        0.2394180
```

RAINMAN Key Facts

Project duration:	07.2017 – 06.2020
Project budget:	3,045,287 €
ERDF funding:	2,488,510 €
RAINMAN website & newsletter registration:	**www.interreg-central.eu/rainman**

Lead Partner

LANDESAMT FÜR UMWELT, LANDWIRTSCHAFT UND GEOLOGIE | Freistaat SACHSEN

Saxon State Office for Environment,
Agriculture and Geology

✉ rainman.lfulg@smul.sachsen.de

Project Partner

Saxon State Ministry of the Interior

Environment Agency Austria

Office of the Styrian Government

T. G. Masaryk Water Research Institute, p.r.i

Region of South Bohemia

Croatian Waters

Middle Tisza District Water Directorate

Institute of Meteorology and Water Management National Research Institute

Leibniz Institute of Ecological Urban and Regional Development

Project support

INFRASTRUKTUR & UMWELT
Professor Böhm und Partner

✉ RAINMAN@iu-info.de