Using machine learning to predict maximum waterheight for flash floods

A proof-of-concept study

Martijn Eduard Kuik
Enschede, The Netherlands, August, 2021

Thesis submitted to the Faculty of Geo-Information Science and Earth Observation of the University of Twente in partial fulfilment of the requirements for the degree of Master of Science in Geo-information Science and Earth Observation.
Specialization: Spatial Engineering

SUPERVISORS:
Prof. dr V.G. Jetten
dr. Marc van den Homberg (510 Global/Netherlands Red Cross)

Advisors:
dr. Jacopo Margutti (510 Global/Netherlands Red Cross)

510 AN INITIATIVE OF THE NETHERLANDS RED CROSS

ITC UNIVERSITY OF TWENTE.

Abstract

Flash floods are one of the most dangerous types of natural disasters. High velocities and the often short time between rainfall and flood occurrence make them hard to predict, especially in an early warning context.

For this, flash flood events can be modelled using a hydrodynamic flood model, but these are often complex and take a long time to develop. Moreover, simulations, depending on the study area, can take up to hours to complete. Therefore, machine learning (ML) is a good option, with significantly shorter runtimes and a less complex development process. Predicting flood extent with ML is currently being done in the form of (flash) flood susceptibility analyses. However, current methods are limited in the information they can provide, mainly due to limitations of the input data. The speed at which these flash floods occur and disappear can it difficult to collect detailed flood extent data. Therefore this research, in the Kyungu river basin in Malawi, assesses the feasibility of predicting maximum waterheight of flash floods with machine learning. This is done by developing a hydrodynamic flood model of the study area, after which the generated maximum waterheight maps of different return period events are used to train a Random Forest (RF), an Extreme Randomized Forest (EXRF) and an Extreme Gradient Boost algorithm (XGB). The ML model was based on a selection of thirteen fundamental for hydrodynamic modelling, predictors and was validated using 10 fold cross-validated R^2, MAE and RMSE. The training consisted of 10 year, 50 year and 100 year return period events, after which a separate 20 year and an 80 year return period event were predicted.

This research results show that the five most important features in predicting maximum waterheight were the DEM, Wetting front suction head (PSI), upstream cumulative (ups), saturated hydraulic conductivity (ksat) and mannings' N (n). Using these predictors, the EXRF algorithm was the most accurate for the training dataset, with an R^2 of 0.68. However, when predicting the 20 year return period and 80 year return period events, the accuracy decreased to an R^2 0.58 and 0.91, respectively. Evaluation of the results suggests that higher accuracy is caused by the significant similarity between trained and predicted events. Therefore this study shows that it is feasible to predict maximum waterheight with ML but also concludes that it has not been proven to be feasible for use in an early warning and early action context yet.

Foreword & Acknowledgements

The past year has been filled with plenty of highs and lows. When starting this topic, the goal was to prove a concept. But more importantly, to push my own boundaries/limits and challenge myself. And that I did, when starting all this with a very basic understanding of hydrological modelling, hydrology and even less knowledge about Machine Learning, it certainly proved to be a challenge. For this, I would like to thank my supervisors prof. dr. Victor Jetten and dr. Marc van den Homberg, my advisor dr. Jacopo Margutti and my mentor dr. ir. Janneke Ettema for their great insights and helpful advice. Without them, I would have been stranded somewhere or still only halfway. While the COVID-19 pandemic has definitely influenced me and this research, with the online meetings, we still managed to make it work, and therefore, this research is as much a result of my work as it is of yours.

I would also like to express my gratitude to my friends I have made at ITC. It has been so much fun working together, getting to know different cultures and having fun together! It has made it possible to keep working and keep moving forward in sometimes difficult times. I would like to thank my fellow student Luis Figueroa specifically, for always being willing to teaching me about Machine Learning and the patience you had when doing so.

Additionally, I would like to thank my roommate and good friend Philip Stobbelaar, my good friend Sander Ausems and my friends from Amsterdam, who made sure to also enjoy life a little bit, especially during the COVID-19 pandemic, and not to take everything too serious

Lastly, I would like to thank my parents and sister for always being there for me and always being willing to listen to me ramble about the various issues I was having with my thesis and giving good (life) advice even though the topics I have covered was rambling about were complex. Travelling to you always truly felt like coming home.

I am glad that I can conclude a year of research and I cannot wait to see what is next.

Inhoud

List of Appendices

List of Figures

List of Tables

List of abbreviations

UNDRR	United Nations Office for Disaster Risk Reduction
IFRC	International Federation of Red Cross and Red Crescent Societies
EM-DAT	Emergency Events Database from the Centre for Research on the Epidemiology of Disasters
DoDMA	Malawi Department of Disaster Management Affairs
DEC	The District Executive Committee
DCPC	District Civil Protection Committee
ACPC	The Area Civil Protection Committee
VCPC	Village Civil Protection Committees
MCA	Multi-Criteria analysis frameworks
OpenLISEM	(open) Limburg Soil Erosion Model
SRTM	NASA Shuttle Radar Topography Mission
NDVI	Normalized Difference Vegetation Index
ECHO	European Civil Protection and Humanitarian aid Operations
GUMBEL	Generalized Extreme Value distribution Type-I
IDF	Intensity-Duration-Frequency
GSMaP	Global Satellite Mapping of Precipitation dataset, from JAXA Earth Observation Research Centre
CHIRPS	Climate Hazards Group InfraRed Precipitation with Station data
ML	Machine Learning
DEM	Digital elevation map
RF	Random Forest
EXRF	Extreme Random Forest
XGB	Extreme Gradient Boost
SVM	Support Vector Machine
R^2	Coefficient of Determination
RMSE	Root Mean Squared Error
MAE	Mean Absolute Error
RFECV	Recursive feature elimination and cross-validated selection

1. Introduction

According to the United Nations Office for Disaster Risk Reduction, a hazard is defined as: *"A process, phenomenon or human activity that may cause loss of life, injury or other health impacts, property damage, social and economic disruption or environmental degradation."*, of which natural hazards are generally associated with natural processes (UNDRR, n.d.). These natural hazards are divided into five main subgroups: geophysical, climatological, biological, meteorological and hydrological. Geophysical referring to hazards such as earthquakes or volcanic activity, climatological hazards include extreme heatwaves or drought, biological referring to hazards such as plagues or epidemics (such as the current COVID-19 epidemic), meteorological consists of disasters such as cyclones or storms and lastly, hydrological hazards (IFRC, n.d.).

Hydrological hazards are defined as hazards or deviations in the water cycle caused by the occurrence, movement and distribution of surface and subsurface water (Guha-sapir et al., 2011; EM-DAT, 2020). These hydrological hazards include floods, wave-action and wet mass movements or avalanches and account for 36.9% of weather-related disaster response triggers between 2008 and 2017. Floods, in general, are and have been the type of hazard to most frequently trigger emergency responses (IFRC, 2018). Floods contributed to a total of 385 disaster response triggers in the same decade. The previous decade (1998-2007), in comparison, had 230 emergency response triggers. Moreover, in 2017, worldwide a total of 18 million people were displaced due to weather-related disasters, of which the majority was caused by floods, displacing 8.6 million people (IDMC, 2018). In 2019 this number rose to 10 million people, with a total of 23,9 million people displaced due to weather-related disasters (IDMC, 2020). Ruiter et al. (2020) describe that it is likely that the frequency of, primarily, non-tectonic disasters is going to increase due to climate change.
These numbers show an increase in natural disasters in general and, more specifically, hydrological disasters such as flooding (Hoeppe, 2016; Douben, 2003). With the potential for an increase in frequency, flood forecasting capability becomes more important.
Historically, protection from flooding and the monitoring of potential flood hazards has become more prevalent with the desire of governments and authorities to understand hazards better and be better prepared for them (Cloke, 2009; Douben, 2003).

Floods are generally divided into four types: Coastal, Riverine, Flash and Ice Jam flooding (EM-DAT, n.d), with each of them having different characteristics. Currently, the majority of flood forecasting research focuses on riverine floods, a type of flood characterised by larger flood extents and time scales (Bucherie, 2019; Douben, 2003). Flash floods, however, can be extremely destructive and deadly, but knowledge and data availability about this type of flood is and are limited (Gaume, 2009; Shresthra & Takara, 2008).
Flash floods are caused by excessive rainfall in one location. However, contrary to riverine floods, flash floods generally occur within several hours after the rain has fallen. This is because the location of the rainfall and the location of (flash)flooding are close in space. This, combined with relief in the surrounding area, where runoff flows quickly downhill into low-lying areas such as rivers or depressions, can cause rapid flooding (Marchi et al., 2010; Borga et al., 2014; Borga et al., 2008; Jonkman, 2005).
The definition of flash floods can be broad. However, the primary identifier of flash floods should be

the short time between rainfall and flooding. Considering this definition, this includes flash flooding in an urban context after exceeding the capacity of artificial drainage structures such as sewers. However, whether to separate flooding from artificial drainage structures from the flash flooding definition is a subject of debate (van Milligen de Wit, 2021).

While flash floods are, as mentioned above, caused by excessive rainfall, sediment can mix with the water and turn flash floods into mudflows and debris flows. Especially debris flows, which contain coarse sediment, can be extensively more destructive (van den Bout, 2020).
The short temporal footprint of flash floods makes predictability of location and size particularly difficult, which is why 510 Global, an initiative of The Netherlands Red Cross, is currently organizing and partaking in various research projects concerning the impact and predictability of flash floods in Northern Malawi (Bucherie, 2019; Broeken, 2018).
510 Global is an organisation that focuses on using and creating data and digital products that aid workers, decision-makers and victims can use to help them prepare and cope with disasters and crises on various scales. They focus on digital risk assessment before a disaster strikes, early warning and early action with predictive impact analysis when a disaster is imminent, offer products for predictive impact analytics during a disaster and offer products that can be used in all phases in terms of recovery to complete the disaster risk management cycle: preparedness, early warning, response and recovery (UNOCHA, 2013; 510 GLOBAL, n.d.).
Current research by 510 Global in Malawi is primarily focused on risk assessment and impact analysis, also known in disaster policy as “early warning action”. The focus here is on saving more lives and reducing possible damage by acting before a disaster strikes instead of only focussing on disaster relief after a disaster happens. This is in line with The International Federation of Red Cross and Red Crescent Societies (IFRC, 2008).

1.1 Flooding in Malawi

Flooding in Malawi due to heavy rainfall is generally a yearly occurrence. The month of January 2015 saw floods caused by extreme rainfall heavily influencing life in the country, with estimations of over one million people being affected (Malawi Government, 2015). This 1 in 500-year rainfall event displaced 230,000 people, killed approximately 200 people, destroyed infrastructure, buildings, cropland, and swept away livestock. This extreme event caused over $400 million in damages, and emergencies were declared in 15 separate districts. One was Karonga, one of the Northern districts where 510 Global is currently doing research (Simpson et al., 2019). At the beginning of 2020, more heavy rainfall caused flooding in Karonga and led to 450 households being displaced (Floodlist News, 2020).
Moreover, impacts due to flooding have increased in the last decade, with a 40% increase in affected households (Bucherie, 2019). Malawi, in general, is highly susceptible to


Figure 1: Flooding in Malawi early 2015. Photo: George Ntonya/UNDP (Floodlist news, 2015)

weather-related disasters, as most of Malawi's economy is based on agriculture. Extreme weather could, therefore, not only be damaging to the population but also to the country's economy (DoDMA, 2019).

Disaster risk management in Malawi is organised in a hierarchical structure. The Department of Disaster Management Affairs (DoDMA) is nationally oriented and focuses on directing and coordinating the implementation of disaster risk programs. The District Executive Committee (DEC) or District Civil Protection Committee (DCPC) is oriented on a district scale. The Area Civil Protection Committee (ACPC) is focused on a local scale, and Village Civil Protection Committees (VCPC) attend to the smallest, village size, scale.

The early warning system in Malawi can be separated on three scales (Teule, 2019):

1. The national Operational Decision Support System (ODSS), an early warning system created in 2016 that includes multiple components, such as weather forecasts, drought monitors and crop calendars
2. Community-based forecasting, where the ACPC’s actively participate in risk identification, solution selection, solution implementation, monitoring and operation.
3. Indigenous knowledge-based forecasts, where the experience and knowledge of the local area and local weather patterns of residents is used in the forecasting and mitigation of hazards.

Most of the disaster risk reduction and early warning systems are the responsibility of the DEC/DCPC. However, often there is limited Disaster risk management or no early warning system in place (DoDMA, 2019; National Resilience Taskforce, 2018). For example, due to a lack of funding, only 15 disaster-prone districts out of 28 are monitored by the DoDMA. In 2018, a national multi-hazard risk assessment was yet to be developed (National Resilience Taskforce, 2018).

The early warning systems in place for flooding are hindered by outdated equipment, a limiting focus on only major rivers, communication problems, misunderstandings over early warning needs, and a lack of understanding of early warning by communities (National Resilience Taskforce, 2018). Regarding the data on flood hazards and the extent of these hazards, the National Resilience Taskforce describes that the database for flood hazard and impact data is incomprehensive and often only provides limited information, with no hazard specific information and a general lack of detail needed to do impact analyses. Progress is being made with disaster risk management and early warning systems. The progress, however, is slow and requires international funding and knowledge sharing with organisations such as 510 Global (DoDMA, 2019). One way to generate data and create insights on the impacts of floods with different return period events is a susceptibility analysis.

1.2 Susceptibility analysis

Susceptibility analyses are a way of evaluating the susceptibility of specific areas regarding disasters. They display how likely it is for an area to flood, often described as low to high susceptibility. This susceptibility is a subset of risk, which can be used to create hypothetical impact data. A susceptibility analysis regards historical evidence for different hazard types and analyses these with a set of independent landscape and climatic variables. If correlations can be found, they can be used to predict the susceptibility of a given location to a hazard. The field of susceptibility analyses gained popularity in the 1980s after the characteristics of a natural hazard were conceptualised in the context of landslides (Varnes, 1984). After which, in the following decades, efforts were made to understand, classify and possibly predict landslides with various quantitative and qualitative methods (Aleotti & Chowdhury, 1999; Corominas et al., 2003). On the contrary, the field of flood susceptibility mapping is relatively new, with an increase in research starting around 2010 in parallel with the increase in popularity of Machine Learning (ML) for various research applications (Chapi et al., 2017; Molnar et al., 2020).

The difference and difficulty of flood susceptibility mapping, compared to landslide susceptibility mapping, is the increase in uncertainty and complexity that comes with the upstream area of a catchment (Pappenberger et al., 2006; Chapi et al., 2017; Tehrany et al., 2015). The characteristics of the upstream area determine the severity of a flood and thus the amount of runoff generated because water flows from high to low (Murray-Hudson et al., 2006; Mati et al., 2008).

Nowadays, many different types of methods are used to perform susceptibility analyses for both landslides and floods. Examples are Multi-Criteria analysis frameworks (MCA), the Analytic hierarchy process (AHP) or ML methods. The common factor in all the different methods used is the conversion of all the input variables or features into one output: susceptibility.

From the more statistical oriented models, ML is a popular approach, whereas, with the multi-criteria (decision) analyses, the AHP method is most widely used (Khosravi et al., 2018; Brito & Evers, 2016; Muñoz et al., 2018). Generally, the ML approach provides more accurate results for flood susceptibility modelling compared to multi-criteria decision frameworks due to the option to use algorithms that prevent overfitting and distraction by noise in the datasets (Khosravi et al., 2019; Shafizadeh-Moghadam et al., 2018, Pham et al., 2016; Pham et al., 2017). For this reason and the fact that, generally, ML can produce results faster (Pourghasemi et al., 2020; Janizadeh et al., 2019), ML will be chosen as the method to create susceptibility maps. The speed at which a specific method can produce results is essential for the overarching context of this research project: early warning, early action for flash floods in Northern Malawi.

Within ML, many different algorithms and models are used for (flash)flood susceptibility mapping, popular examples are tree-based methods such as Classification And Regression Trees (CART) or Random Forest (RF), or even more complicated methods such as the Boosted Regression Tree (BRT) or Support Vector Machines (SVM) (Tehrany et al., 2015; Khosravi, 2018; Khosravi, 2019; Pham, 2017; Pourghasemi, 2020; Shafizadeh-Moghadam, 2018; Chapi et al., 2017).

ML algorithms mainly differ in the way they predict values, what each parameter represents and how they relate to the input. An important factor that is equal for all ML algorithms and approaches is that the importance of the selection of predictors for ML model performance is equal for all methods. Because in ML, the input data determines the quality of the output and, therefore, determines the model's performance.

For ML flash flood susceptibility analyses, these predictors are usually selected via a literature review of historical floods, flood-related geospatial databases, via previous flood susceptibility modelling literature or by using local knowledge (Janizadeh, 2019; Khosravi et al., 2019; Shafizadeh-Moghadam, 2018; Khosravi et al., 2019; Samanta et al., 2018; Bucherie et al., 2021; Tehrany et al., 2015; Lin et al., 2019). For most of the current literature, however, the study area size is large. With catchments often being thousands or tens of thousands square kilometres in size, flood locations are reduced to points in a vast area. Researching on a smaller catchment scale can give insights on how terrain features directly influence flood extent, where subtle differences in characteristics of the upstream area can be seen downstream (Hu et al., 2015). Considering this, having a smaller scale stimulates the selection of predictors that directly influence the hydraulic flow process on a fundamental level. Key hydrodynamic modelling parameters such as saturated hydraulic conductivity or the capillary suction of a soil directly determine the amount of water that runs off over the surface (Hu et al., 2015). The literature overall has shown a lack of inclusion of key hydrodynamic modelling parameters in ML susceptibility analyses. Considering the large scale of the study areas, the possible variation in landscape characteristics and the fact that flood susceptibility is reduced to points makes it hard to determine the accuracy of the prediction in the context of the hydraulic processes that take place in the real world to flood that pixel/point.

Literature has shown that, generally, the flood inventory databases created to train and test the ML models consist of historical flood events that all occurred within a broad timespan. Having such a mix of different events with possible different flood extents and rainfall patterns that caused them, leaves no possibility to do a susceptibility analysis for a specific event. In this way, the only thing to generally predict is whether an area can/will flood. A susceptibility analysis can be made more informative by adding population density or building information. It then becomes more of an impact analysis (Lin et al., 2019).

1.3 Problem statement

With an increase in disasters and, more specifically, weather-related disasters such as flash flooding, it is evident that the consequences for people living in risk areas will worsen. The nature of flash floods makes it hard to predict when it is happening, where it is happening, and what the impact of the flash flood will be. Current (flash) flood susceptibility analyses are, in this regard, limited in the information they can provide decision-makers on specific rainfall events and their flood extent. With limited resources, current risk management programs in Malawi are restricted, especially for flash flooding.

This leads to the following research question:

What is the feasibility of using machine learning to predict the maximum waterheight of flash floods in the context of early warning and early action in Northern Malawi?

This research will introduce key hydrodynamic modelling parameters into ML susceptibility analyses for flash floods by first simulating the study area in flash flood scenarios in openLISEM, a flood modelling software. Then the hydrodynamic modelling parameters will be selected based on their role and importance in openLISEM and flood modelling in general, which will then be added as features to the ML models. After which, an ML susceptibility analysis will be done with several algorithms. With the addition of hydrodynamic modelling parameters in the susceptibility analysis, the aim is to improve a susceptibility analysis' flood extent prediction capability for Northern Malawi by predicting waterheight for different return period events. Additionally, the overarching goal is to create and explore a method that can give as much relevant flood extent information as possible in a short time, assisting early warning, early action and improve impact data.

1.4 Objectives and research questions

Main research objective:
To create a method to improve Machine Learning flash flood susceptibility analysis' accuracy and informative capability by predicting maximum waterheight in an early warning and early action context in Northern Malawi.

Main research question
What is the feasibility of using Machine Learning to predict the maximum waterheight of flash floods in the context of early warning and early action in Northern Malawi?

1. Sub-objective 1
To develop a realistic flash flood model of the study area in openLISEM, due to a lack of flood extent data, to establish a ground truth on which the Machine Learning algorithms can be trained.

Research questions sub-objective 1:
1. What are the different amounts of rainfall for different return period events for the study area?
2. What is the maximum waterheight and flood extent for different return period events in the study area?

2. Sub-objective 2
To develop a Machine Learning model for the study area using predictors created from a hydrodynamic modelling perspective to predict waterheight.

Research questions sub-objective 2:
1. Which combination of predictors is selected, and what is their importance in predicting maximum waterheight?
2. What is the performance during the training of the different machine learning algorithms

3. Sub-objective 3
To generate a predicted flash flood maximum waterheight for different return period events in the study area

Research questions sub-objective 3:
1. What is the prediction accuracy?
2. How does the ML prediction compare in flood extent?

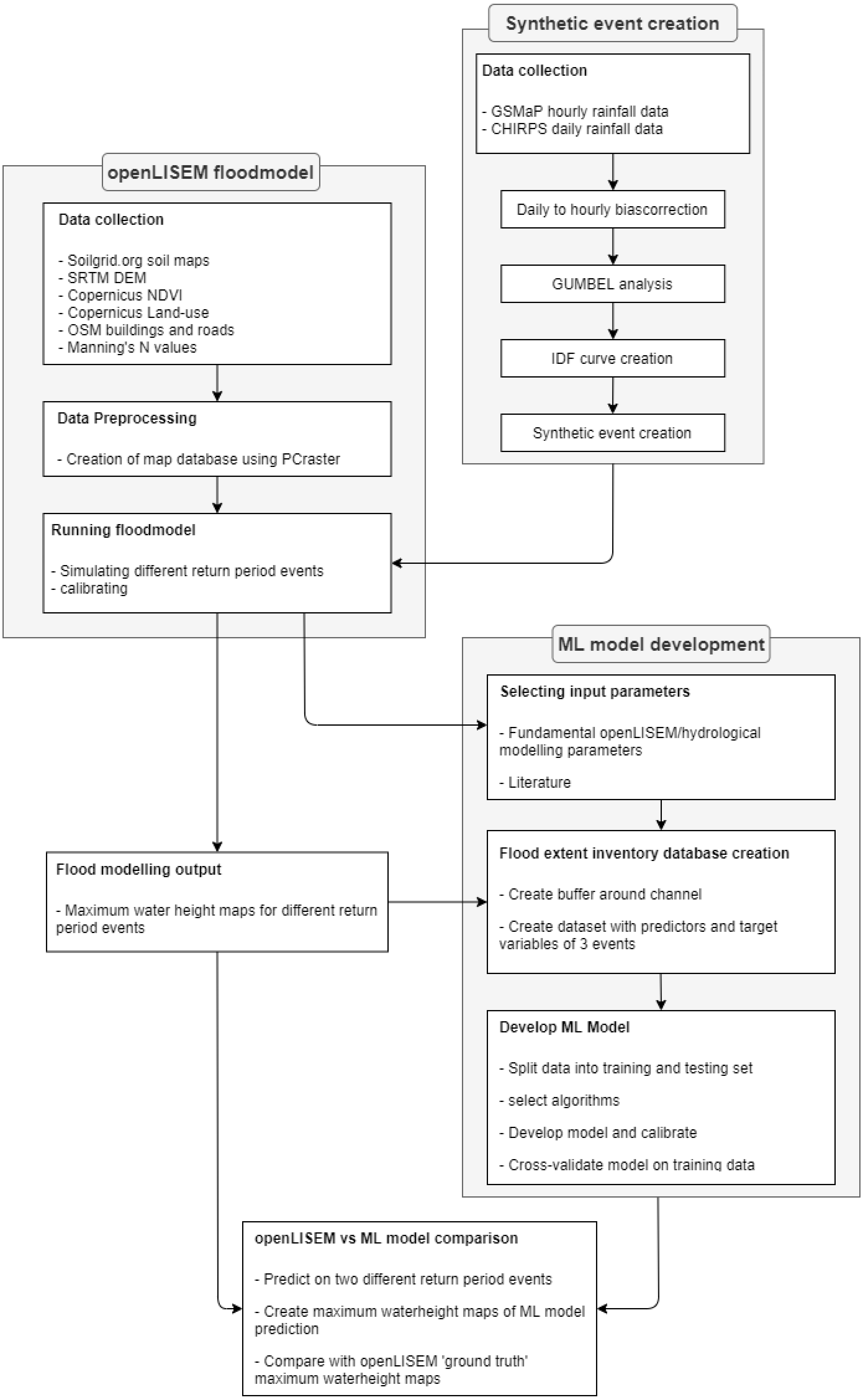


Figure 2: Flowchart of the different methods used for this research

2. Methodology

2.1 Case study area: Karonga, Malawi

Malawi is a country located in the south-eastern part of the African continent. The country is landlocked and surrounded by Mozambique, Tanzania and Zambia. Lake Malawi, also known as Lake Nyasa, is the main waterbody that borders the east side of Malawi, stretching about 580 kilometres long. The country has a total surface area of 118.484 km^2 and a population of 17.563.800 in 2019, with most people living in rural areas. Malawi has an agriculture-focused economy with a GDP per capita of $1200 in 2017, producing products such as tea, tobacco, cotton and sugar (National Statistical Office, 2019; Central Intelligence Agency, 2020).

Malawi's climate can be classified as sub-tropical, with warm wet seasons typically lasting from October to April and the dry cold season from May to September.

The study area is located in Northern Karonga in the Karonga and Chitipa districts and near the Tanzanian border, as shown in figure 3. The districts of Karonga and Chitipa have quite a broad range of average annual rainfall, as can be seen in figure 5. The town of Karonga itself annually receives about 1201 to 1400 mm, while the more elevated parts of the district northwest of the town receive up 2400 mm or more.

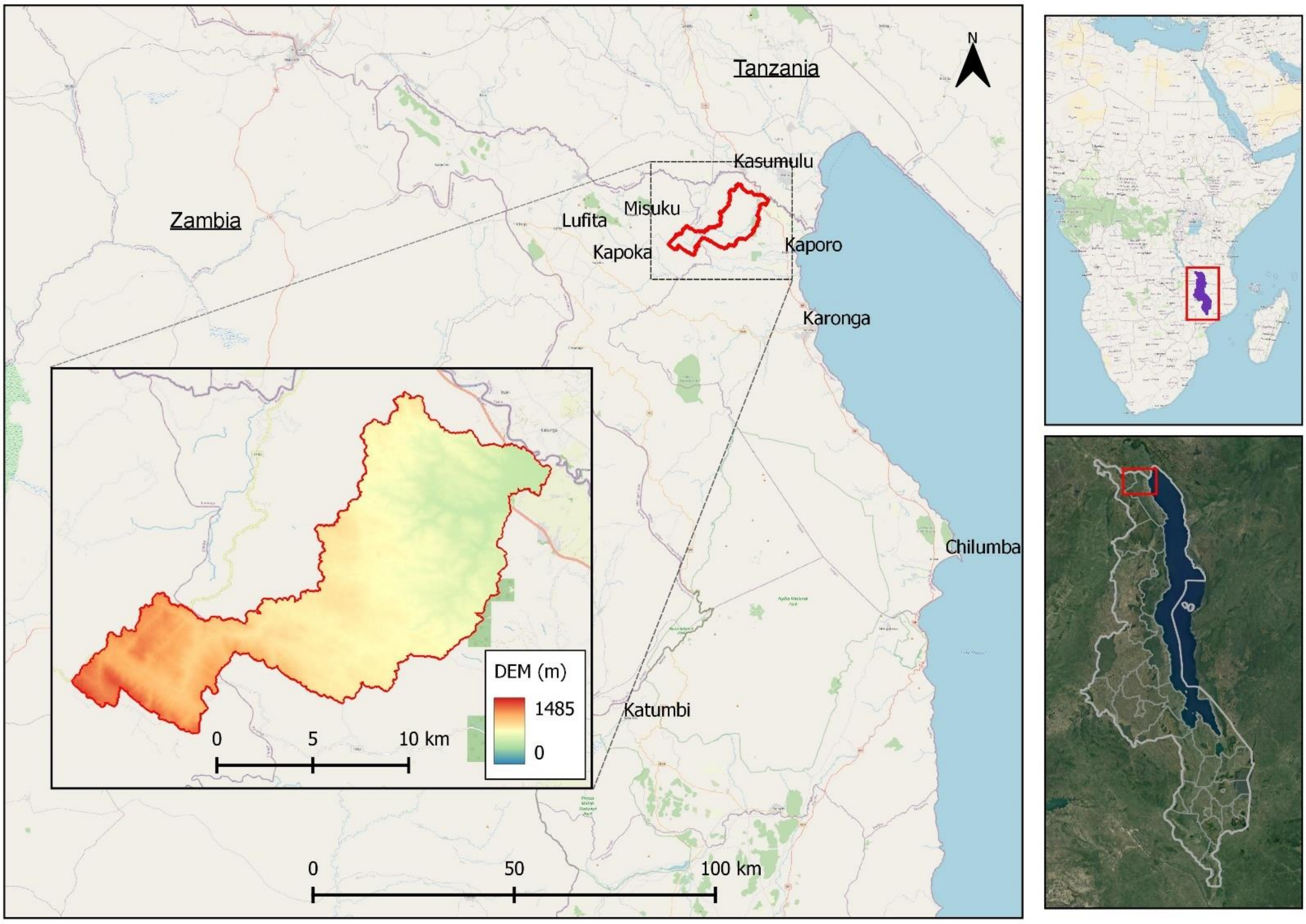


Figure 3: Study area of the Kyungu river catchment in the Karonga and Chitipa districts in Northern Malawi

These rain patterns have caused various types of floods in Karonga town, such as flash floods and riverine floods of the North Rukuru River and Lake Malawi itself (Manda et al., 2016). The specific catchment is selected out of the PCA based catchment susceptibility analysis done by Bucherie (2019), as shown in figure 5.

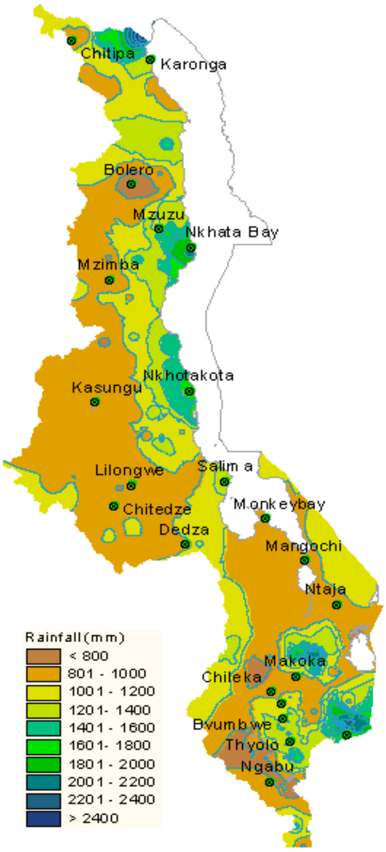


Figure 4: Malawi Annual average rainfall (DCCMS, 2020)

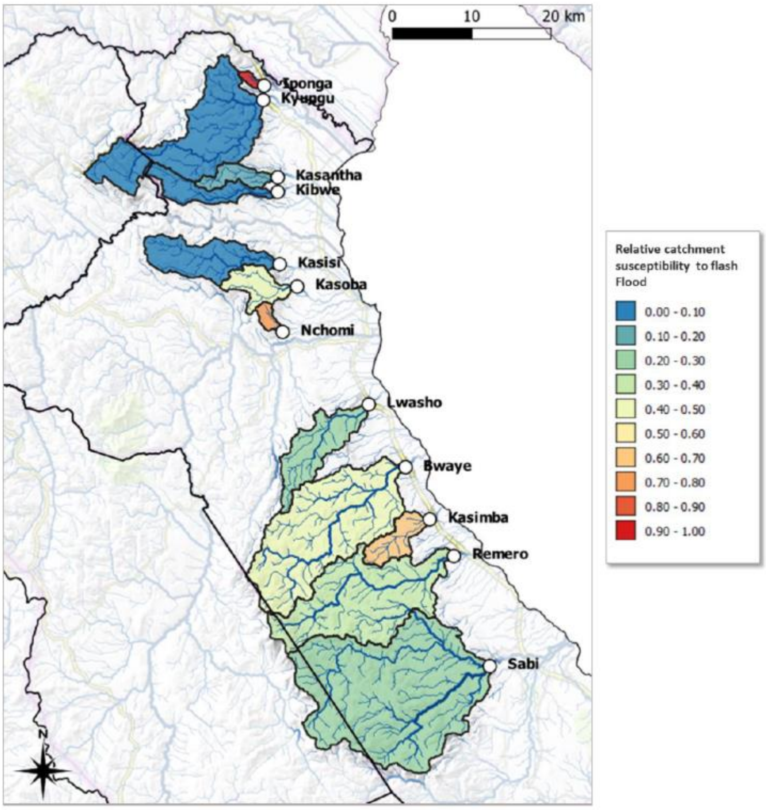


Figure 5:PCA based catchment susceptibility analysis (Bucherie, 2019)

The catchment has been selected later via two main criteria. The first and primary criterium is size. The size of the catchment in the study area is important concerning flood modelling. A bigger catchment means more calculations in openLISEM, which increases the time needed to do an entire simulation run. The second criterium is soil composition. The goal is to select a catchment with as much soil heterogeneity as possible in the context of potential further applicability and validity. Based on these criteria, the Kyungu river catchment has been chosen, as shown in figure 3, south of the border with Tanzania. The study area size is 159.20 km^2. A bigger study area was chosen to add more diversity and heterogeneity, which can aid in representability. Furthermore, the study area is located in an area that has received one of the most amounts of rain in 2020, as can be seen in figure 4. The selection of the study area made it possible to continue to the next step: creating a map database.

2.2 Development of the hydrodynamic flood model

To develop a flood model for a study area, a database of maps needs to be created in order for the model to be able to make its calculations. For the creation of a map database, five primary layers are needed, of which the rest of the maps or parameters will be calculated: a Digital Elevation Map, a soil layer, a land-use layer, a river network layer and a buildings/infrastructure layer. Some of these base layers need to be created or adjusted, for which pre-processing is needed.

2.2.1 openLISEM

The Limburg Soil Erosion Model (LISEM) is a physically-based numerical capable of simulating runoff, sediment dynamics and shallow floods on a catchment scale. The model was initially developed in 1994 to model soil erosion in Limburg, a province in The Netherlands (De Roo et al., 1994). The vision of openLISEM is "To improve the understanding, to help analyse, and to predict the behaviour of land surface processes, our vision is the development of an integrated, holistic and fundamental modelling tool for the numerical simulation of land surface processes." (Jetten, 2018). Included in their vision is to allow for as much societal benefit as possible, which is why the decision was made to make it open-source and freely available. The LISEM model is based on work done by the original developers and scientific literature (Jetten, 2018; Jetten & De Roo et al., 2001).

Multiple hydrodynamic models exist to simulate floods in openLISEM: 1D, 2D and 1D+2D (Teng et al., 2017; van de Bout, 2020). One-dimensional models simulate flow in one dimension, which can be used where higher dimension modelling is unnecessary, for instance, with a pipe or a canal. Two-dimensional modelling is the most used method for flood simulations. It provides the possibility to model a flood plain in another dimension: horizontal velocity of water averaged across a vertical column. Three-dimensional models have been developed to model the potentially dangerous spiral flows that occur at dam breaks or tsunamis. The combination of a 1D and 2D model works in the way that a 1D kinematic wave is used for the runoff and the 2D dynamic wave is used for flooding, where only the water overflowing from the channel is classified as flood.

Further details on the underlying physical principles of openLISEM can be found in Baartman et al. (2012) and Jetten & De Roo (2001).

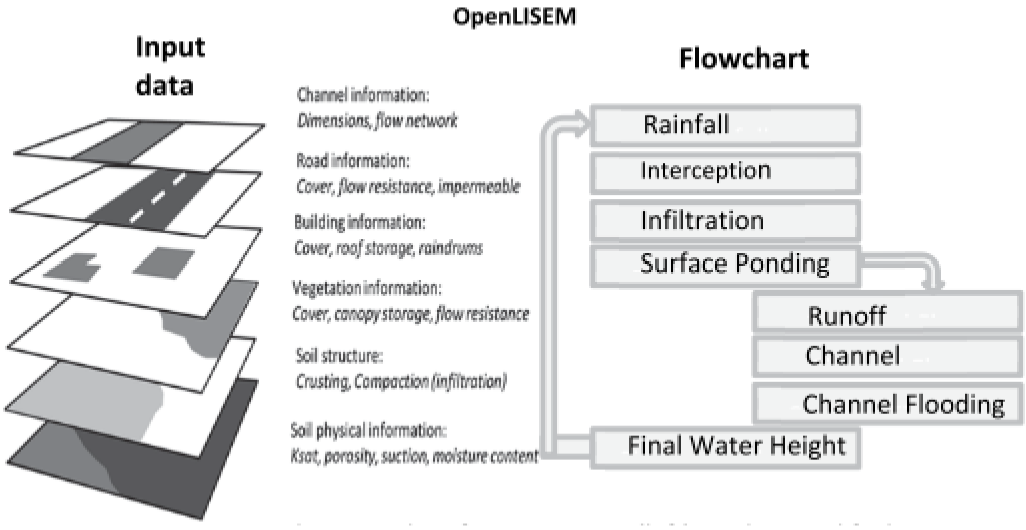


Figure 6: Simplified flowchart of openLISEM process (van den Bout, 2020)

2.2.2 Pre-processing

Digital elevation model

Two sources for the DEM were used for this research: ALOS-PALSAR AW3D30 version 3.1 (12.5-meter resolution) and SRTM (30-meter resolution). Both DEM's were clipped, reprojected and then resampled to a 20-meter resolution. The ALOS-PALSAR DEM proved to be challenging to work with due to irregularities. Water was not flowing correctly over de DEM, which is why the decision was made to use the SRTM DEM. Results were better but still not optimal. Therefore all pixels below 505 meters were processed with a 3x3 averaging filter to smooth the floodplain.

A higher resolution or better quality DEM is usually required for proper use of the 2D dynamic wave and overland flow simulation. This is why the 1D + 2D hydrodynamic model was chosen for this research, as the 1D aspect in this model negates depressions in the DEM. This choice was made under the assumption that pits or depressions in the DEM are caused by errors in the DEM itself.

Mask

Using the DEM, a surface flow network can be created. This was done by first eliminating small depressions in the DEM, after which the local drain direction (ldd) network was used to create the channel network. The outlet point was determined manually at the end of this channel network, after which all pixels that flow into this outlet point were determined and thus, the entire catchment was delineated. The catchment was then used as a mask to clip all input maps, including the ldd. The masked ldd was then used to create the river/channel network of only the catchment

Soil

Soil characteristics contribute to some of the essential flood modelling parameters, such as saturated hydraulic conductivity, porosity, initial moisture and suction. The following soil maps were downloaded from soilgrids.org (De Sousa et al., 2020; Batjes et al., 2019):

- Sand content
- Clay content
- Silt content
- Organic Carbon content
- Gravel content
- Bulk density content

Soilgrids.org is a publicly accessible database that maps soil properties globally with ML methods on a 250-meter resolution using the soil profile observations of the World Soil Information system (WoSIS) to verify their maps (Sousa et al., 2020).

All these soil maps were downloaded for two different depths: top soil (5-15cm) and subsoil (60-100 cm) because openLISEM has the ability to work with the two levels of soil depths.

These maps are in fractions, and for this research, these fractions were converted to g/kg. All these soil maps were used in pedotransferfuctions (Saxton & Rawls, 2006) to calculate the following maps/variables for two different soil depths:

- Porosity
- Saturated hydraulic conductivity
- Initial moisture
- PSI
- Density
- Organic Matter
- Wilting point
- Field capacity
- Plant available water content

Land-use and Coverage

Not only subsurface characteristics are influencing waterheight in the flood model, but also surface elements—for example, land-use. The land-use type determines how much water can be intercepted before it infiltrates into the ground and how high the resistance is when water flows over the surface (manning's N and micro-relief). The assumptions made for the resistance of different land-use types will be highlighted in paragraph 2.3.4.

The land cover map from the European Union's Copernicus program from 2019 was used for this research. This is an already classified land-cover map with a 100m resolution, which was chosen because real-life validation was not possible due to COVID-19 restrictions.

To calculate the amount of intercepted water and the specific amount of vegetation per pixel, an NDVI was also needed. The NDVI used for this research was retrieved from Sentinel 2 images from 10-24-2016, as this was the earliest date with close to cloudless conditions.

Pre-processing for the NDVI included: calculating the Normalized Difference Vegetation Index and removing all the negative values because they are not vegetation.

A bare soil map was not made because the Copernicus land-use map has a bare/sparse vegetation class.

Buildings were added using Openstreetmap (OSM). Because the model's resolution is 20x20 meters, the buildings are added as a fraction of the pixel. This means that the manning's N of that pixel is going to be higher. The roads were added via OSM as well.

Basemap	Resolution	Version/Date	Source
DEM	30 meters	V2.0: 12-11-2020	SRTM Nasa
Landuse	100 meters	2019	Copernicus Global Land Cover (Buchorn et al., 2020)
NDVI	20 meters	24-10-2016	Sentinel 2
Soilmaps	250 meters	V2.0: 05-2020	Soilgrids.org (De Sousa et al., 2020)
Roads and buildings	-	26-01-2021	OpenStreetMap
Rainfall	CHIRPS 0.05°×0.05° GSMAP 0.1°x0.1°	01-01-2001 to 31-12-2020	CHIRPS (Funk et al., 2015) GSMaP (Kubota et al., 2020)

Table 1: An overview of the different characteristics of the base layers

2.3 Assumptions and uncertainty

Flood models are built on assumptions and for event-based models, such as openLISEM, these assumptions influence the output even more. The state at the initialisation of a catchment is important for even-based models, with factors such as soil saturation due to antecedent rainfall, the state of vegetation and crops at the time the flood occurs, the exact size of the channels and the accuracy of the DEM. Many choices need to be made with the creation of the model, which brings uncertainty. Normally, this uncertainty can be decreased by calibrating the model by doing fieldwork. However, not having the ability to do fieldwork due to the COVID-19 pandemic or not having a measured river discharge made calibration difficult.

Nonetheless, it is important to mention that flood models are not meant to represent reality perfectly. Models should always be viewed as a method to assist in water governance decision making. However, an attempt to make the model flood model as realistic as possible should always be made. This section will highlight the parts of the model where the most significant assumptions were made and why.

2.3.1 Reference floods

The problem lies mainly in the fact that real-life data is exceptionally scarce and, while the ECHO III project provided accurate and specific data for various locations in the flood area, it was not possible to connect these damage numbers to specific events. The European Civil Protection and Humanitarian aid Operations (ECHO) III project focuses on the preparedness for response and early action in, among other countries, Malawi. For this, questionnaires were conducted in Malawi for local knowledge on historical flood events.

In order to accurately compare the flood model to real-life, it is imperative to know the amount of precipitation that fell during an event, which in turn requires the exact date to be known. Unfortunately, this is not present in the ECHO III dataset and is, in general, hard to determine if there is not extensive flash flood research and a database present. Attempts have been made to make this model as accurate as possible by looking at average flooding in an area determined as the floodplain in the study area (see figure 7). This area was determined by classifying every pixel below a height of 505 meters in the DEM.

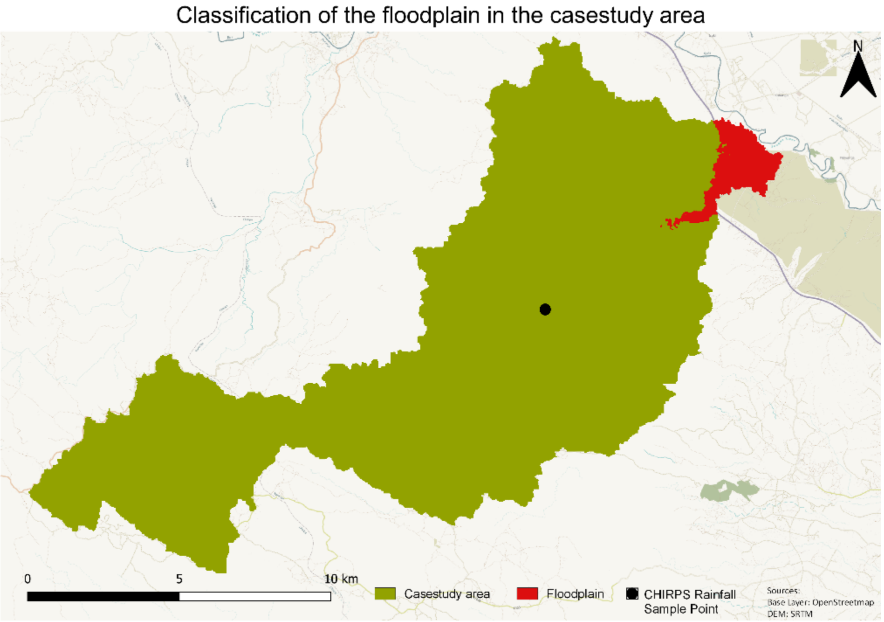


Figure 7: The classified floodplain marked in red, with the sample point for the CHIRPS data in the middle for the study area

By marking the floodplain within the study area with value 1, it became possible to calculate the average waterheight within the floodplain area. This function was then used to calibrate the model as best as possible according to the average waterheight calculated from the ECHO III dataset. However, the ECHO III dataset describes flooding from events from multiple years. In order to have the highest chance of finding the right event, 2018 is the year with the highest chance. When looking for representable rain events, 2018 had significantly fewer extreme rain events than other years. So this, in combination with the fact that in the ECHO III dataset, there was a limited amount of people mentioning 2018 events, provided the highest chance of finding the rain event the interviewees were referring to. However, unfortunately, the waterheight of the flood model using the 2018 rain event did not match the waterheights described by the interviewees. This discrepancy can be attributed to two possible reasons and will be explained in the following paragraphs:

1. Rainfall data uncertainty
2. openLISEM flood model assumptions and uncertainty

2.3.2 Rainfall data

The primary possibility is that the events described in the ECHO III dataset are not the same events as the ones chosen and downloaded from JAXA Global Rainfall Watch. The Global Rainfall Map (GSMaP) by JAXA Global Rainfall Watch' was produced and distributed by the Earth Observation Research Center, Japan Aerospace Exploration Agency (Kubota et al., 2020).
Additionally, when GSMaP data was compared with rain-gauge data, daily summaries for two of the events did not match up. Only for the 2018 event, the rainfall measured with both methods was similar (see table 2), which is why the attempt was made to use this event for calibration. However, only the first event caused flooding in the openLISEM model mainly due to the intensity. While the other two scenarios had comparable amounts of rainfall, the timespan in which this fell was much larger, leading to almost no flooding. This is normal, as rainfall has more time to infiltrate into the

soil instead of becoming runoff. However, the problem is that these events were found based on new articles and red cross documents published describing extensive flooding, while the openLISEM simulations showed no flooding.

Songwe border station	GSMaP	Raingauge
Event 1 (01-04-2017)	160 mm	18.2 mm
Event 2 (16-04-2017 to 21-04 2017)	174.6 mm	383.6 mm
Event 3 (12-04-2018)	132 mm	139 mm

Table 2: Table showing the differences between GSMaP and Rain gauge data at Songwe Border Station

Due to the inconsistent accuracy of the GSMaP satellite data when compared to rain gauge data, the decision was made to run the model with design storms based on return periods of 2, 5, 10, 20, 50, 80 and 100 years, which were created with a GUMBEL analysis. However, rain-gauge stations in Malawi are relatively scarce, but more importantly, inconsistent. With the aim to continue to use publicly available datasets, the Global Historical Climatology Network (GHNC) daily dataset was chosen, with Karonga Airport as rain gauge. This rain gauge was chosen due to its proximity to the study area and temporal coverage. The dataset ranged from 1950-08-01 until 1990-06-30 but only had a data coverage of 67%, leaving significant gaps in the datasets.

Moreover, the days and months missing were primarily in the rain season, which led to 18 years having no rainfall/rainfall data and skewed results of the GUMBEL analysis. So these years had to be omitted from the dataset, which created a dataset that span 22 years where even the usable years had considerable amounts of missing data. The results from the GUMBEL analysis can be seen in Appendix B.

With the unreliability of rain gauge data and satellite data, the decision was made to use data from Climate Hazards Group InfraRed Precipitation with Station data (CHIRPS) rainfall data. CHIRPS is a quasi-global rainfall data set panning 50°S-50°N (and all longitudes) that was created to account for the underestimation of extreme rainfall event intensity by satellite measurements and the often scarce ground-station coverage in rural areas since 1981. CHIRPS blends ground station rain gauge and satellite to produce daily, pentadal, and monthly datasets with a high resolution of 0.05° (Funk et al., 2015; Climate Hazard Centre, n.d.).

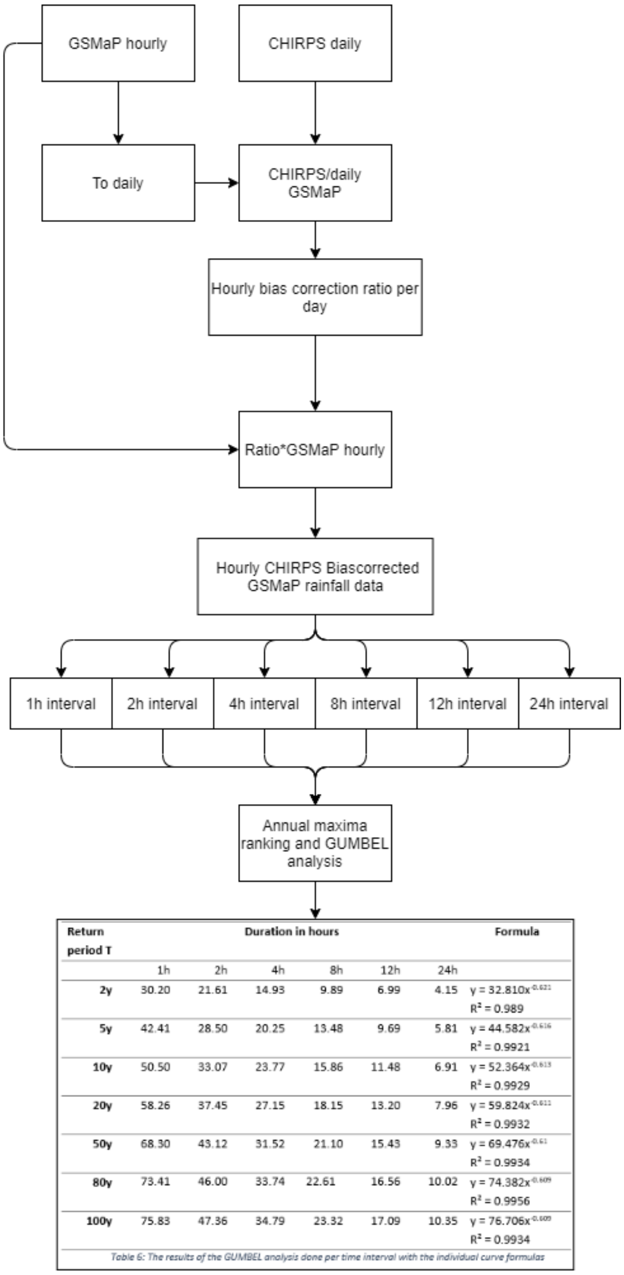
2.3.3 Bias correction

The CHIRPS dataset for the GUMBEL analysis ranged from 01-01-1981 to 31-12-2020. It was obtained via Google Earth Engine, where a CSV file of the daily rainfall was downloaded that is extracted from the raster images for one designated point, see figure 7. The coverage percentage of this dataset is 100%. In order to generate proper IDF curves for the study area, the hourly GSMaP data was multiplicatively bias-corrected using chirps data.

The GSMaP hourly dataset ranged from 01-03-2000 to 01-03-2021, which means that the CHIRPS dataset used for correction has also been cut to the same range. While a period of 21 years is enough to predict return periods of events, multidecadal oscillations occur in extreme precipitation trends (Fadhel et al., 2017). However, disaggregation of the 40 year period CHIRPS dataset was not within the scope of this research.

The bias correction was done in two different ways, with an R package called ‘biascorrection (Malek, 2020) and manually in Excel. The biascorrection R package first aggregates the daily data of both datasets to monthly quantiles, after which the bias is corrected for each month. Then, the monthly data is disaggregated to daily data using a scaling coefficient that bias corrects the raw daily GSMaP input data to match the previously calculated monthly bias-corrected values

For the bias correction that was done in excel, no monthly aggregation was done. Instead, the bias correction was done on a daily scale, where each day of the GSMaP dataset was corrected with the CHIRPS dataset. For every day, the ratio of the difference between the Chirps and the GSMaP dataset was calculated, after which every hour of the GSMaP dataset was multiplied by the ratio that was calculated for the day that hour belonged to. However, this also means that for days where the GSMaP dataset recorded no data, and the CHIRPS did, the GSMaP now has the same multiplication factor for every hour to come to the daily CHIRPS value. This is the case for 290 days out of 7613 days and skews the frequency distribution for 24-hour duration storms.



Return period T	Duration in hours						Formula
	1h	2h	4h	8h	12h	24h	
2y	30.20	21.61	14.93	9.89	6.99	4.15	$y = 32.810x^{-0.621}$ $R^2 = 0.989$
5y	42.41	28.50	20.25	13.48	9.69	5.81	$y = 44.582x^{-0.616}$ $R^2 = 0.9921$
10y	50.50	33.07	23.77	15.86	11.48	6.91	$y = 52.364x^{-0.613}$ $R^2 = 0.9929$
20y	58.26	37.45	27.15	18.15	13.20	7.96	$y = 59.824x^{-0.611}$ $R^2 = 0.9932$
50y	68.30	43.12	31.52	21.10	15.43	9.33	$y = 69.476x^{-0.61}$ $R^2 = 0.9934$
80y	73.41	46.00	33.74	22.61	16.56	10.02	$y = 74.382x^{-0.609}$ $R^2 = 0.9956$
100y	75.83	47.36	34.79	23.32	17.09	10.35	$y = 76.706x^{-0.609}$ $R^2 = 0.9934$

Table 6: The results of the GUMBEL analysis done per time interval with the individual curve formulas

Figure 8: Flowchart of the biascorrection, annual maxima ranking and GUMBEL analysis on which the IDF curves are based

With the hourly bias-corrected GSMaP data, the intensity-duration-frequency curves were calculated based on an annual maxima series, where a GUMBEL analysis was done for every time interval for every return period (See appendix C). These curves combine rainfall event intensity over time (duration) and the return period of said event (Bougadis & Adamowksi, 2006; Rudari et al., 2016). The curve provides a rainfall pattern over time for a specific return period event, of which the equation can be extracted to create synthetic rainfall events. These events were calculated with a duration of 24 hours and designed with an alternating block method to make the events symmetrical for better and clearer discharge (Duka et al., 2018; Chow et al., 1988).

2.3.4 openLISEM Flood model assumptions

As previously mentioned, lack of calibration leads to certain assumptions being made on the following variables.

Channel depth and width

The channel depth is one factor where assumptions were made. It has proven to be challenging to estimate the channel depth precisely. This is due to the fact that no actual concrete data is available. The ongoing COVID-19 crisis makes it difficult to gather data and the resolution of the DEM is not high enough to assess channel depth accurately. So channel depth has been based on the channel width, where the channel depth is 0.21*times the channel width. The channel width has been estimated using google earth as a reference, mainly in the downstream floodplain area. However, it has not been possible to model the inversion of the width of the channel as it transitions into the floodplain.

Another problem was that the catchment in the study area ends in a floodplain, as previously explained. Therefore the depth of the channel at a certain point starts to decrease. However, it is not known when this exactly starts and how shallow the channel becomes.

Manning's N and micro-roughness

Manning's N and the micro-roughness of the surface were partly determined from literature and partly estimated. The manning's N is also known as a descriptor of the density of vegetation. The higher the density and manning's N, the slower water can move through that vegetation. The Copernicus classified land-use map shows nine different land-use types in the study area:

Land-use type	**Manning's N values**
Cropland	0.04
Herbaceous vegetation	0.035
Built-up	0.02
Shrub land	0.05
Deciduous broadleaved open	0.15
Deciduous broadleaved closed	0.15
Unknow type closed forest	0.15
Unkown type open forest	0.15
Herbaceous wetland	0.1825

Table 3: The different manning's N values for the different land-use types. The Manning's N values were taken from literature (Hessel et al., 2003; Chow, 1959, Liu et al., 2019).

Soil Erosion

The openLISEM software was initially developed in 1989 to provide detailed data to help mitigate runoff and soil erosion in The Netherlands. Soil erosion encompasses the process of soil detaching, transportation and deposition downstream by wind or water (Jetten, 2016).

For this research, erosion was not considered when the specific events selected were modelled, as the erosion process understanding in the study area is not the goal of this research. However, in the study area, erosion does occur and has affected the flood extent greatly, as sedimentation was occurring in channels and culverts (Bucherie, 2019). However, the influence of erosion for the developed hydrological model and thus different return period events is outside of the scope of this research.

Interception

Depending on the study area, most of the rain falling during a rain-driven event will be intercepted by other surfaces than the soil (Linsley, 1982; Jetten, 2018). This interception can be calculated for different types of vegetation, buildings and rain drums in LISEM.

The maximum canopy storage is different for different vegetation types and is primarily determined by the Leaf Area Index (LAI). Within LISEM, this canopy storage is modelled as fixed storage and thus takes away from the rainfall. Because the Copernicus land-use map showed mainly broadleaved vegetation, the broadleaved forest setting was selected.

Infiltration

Infiltration is one of the most essential processes of openLISEM. It describes the transport of water downwards into and through the soil, from the surface into the subsurface (Jetten, 2018). The infiltration is determined by the hydraulic conductivity and soil moisture content.

Within openLISEM, there is the option to choose between three different infiltration models: Green and Ampt (1911), Smith and Parlange (1978) and SWATRE (Bastiaansen et al., 1996).

For the openLISEM model created for this research, the Green and Ampt model was selected, as it provides the option of 2 layer infiltration. Additionally, an impermeable lower soil boundary was implemented. Flow and advanced settings were left at their default settings.

Baseflow

Because openLISEM does not model groundwater, thus baseflow needs to be assumed because baseflow is coming from groundwater. The baseflow was assumed to be at half the channel width.

2.4 Machine learning materials and methods

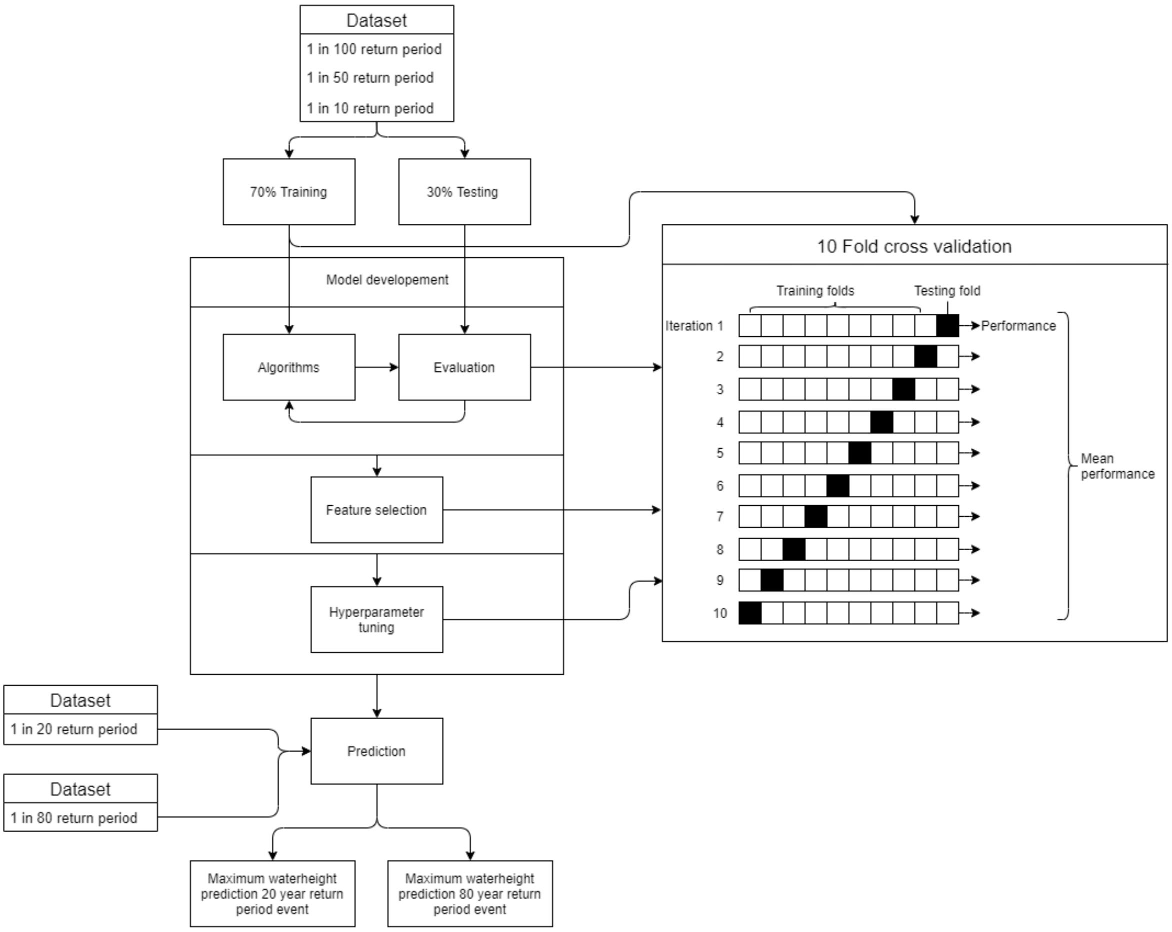


Figure 9: Simplified flowchart of the ML model development process for this research. Where the training data was used for the cross validation, which was used for baseline model evaluation, feature selection and post hyperparameter tuning evaluation.

Language	**Python**
Platform	Google Colab
Main packages	Sklearn Numpy Matplotlib

Table 4: The main tools for the ML model development

2.4.1 Features

The features have been selected and created based on their importance in the hydrological model in openLISEM and their importance seen from a hydrological modelling perspective. It is important to mention that the majority of the features chosen for the ML model did not need any further pre-processing than what was done for their creation for the openLISEM flood model.

The goal is to select features that do not need an extensive GIS analysis for their creation because not having features that need a significant amount of time to create is essential when considering the ML model in the context of using it as a potential method for early warning.

1. Relative DEM (DEM)

The DEM has been included as it lies at the core of a hydrodynamic model, as can be read in paragraph 2.2.2**.** It has been made relative to account for possible future uses in other study areas.

2. Mannings' N (n)

Mannings' N is a descriptor of the density of vegetation. The higher the density and manning's N, the slower water can move through that vegetation and over the surface (Arcement & Schneider, 1989).

3. Leaf area index (lai)

The maximum canopy storage is different for different vegetation types and is primarily determined by the Leaf Area Index (lai). Because canopy storage Is fixed in openLISEM, it decreases the amount of rainfall that hits the ground and thus infiltrates.

4. Saturated hydraulic conductivity (ksat1)

Ksat or the saturated hydraulic conductivity of the soil is one of the most essential variables in hydrodynamic modelling, as it explains how fast water can move through the soil when saturated. The higher the Ksat, the less runoff is generated because water can infiltrate quicker. The saturated hydraulic conductivity can be calculated using the Saxton and Rawls (2006) pedotransfer functions (Jetten, 2018).

5. Roads (roads)

Roads were included as they can influence infiltration rates greatly. More road means more area with poor infiltration, which could have a more significant influence in urban areas.

6. Porosity (thetas1)

The porosity of the soil is the amount of voids or empty spaces present in a material, or in this case, soil type.

7. Initial moisture (Thetai1)

Theta is a measure for how wet the soil is and is calculated for openLISEM by multiplying the initial moisture content with the porosity of the soil. A higher theta means a wetter soil, thus less infiltration and thus more runoff (Jetten, 2018).

8. Wetting front suction head (PSI)

PSI defines the average suction wetting front of a soil or the capillary suction of a soil. The capillary suction is the transportation of water or a liquid through porous solids such as a soil by the surface tension of the liquid (Jetten, 2018; Uzoegbo, 2020). A higher PSI means a drier soil with more suction and thus less run-off.

9. Landuse (landunit)
Land-use can significantly impact runoff and infiltration rate. In this regard, it can be considered as redundant, as the manning's N, ksat, PSI and theta are also included. Nevertheless, this redundancy remains to be analysed.

10. Gradient (grad)
Gradient is the local slope of the cells, which impacts overland flow. The higher the gradient, the steeper the slopes

11. Flow accumulation (ups)
The ups is a map that is a cumulative map that adds all the upstream cells to take the upstream area into account. This map is derived from the ldd and has higher values in the channel the further you go downstream because of its cumulative nature.

12. Saturated hydraulic conductivity of dirt roads (ksatcomp)
Ksat comp describes the saturated hydraulic conductivity on dirt roads. The study area has more dirt roads than tarmac, potentially influencing the general infiltration rate and thus the amount of runoff generated. Ksat for dirt roads is often lower because the dirt is compacted by vehicles driving over it. This was calculated by multiplying the ksat values on dirt roads by 0.1.

13. Rainfall (rainfall)
Rainfall is one of the most important variables in hydrodynamic flood modelling, and to distinguish between different events, rainfall is also included as a feature. This is the only feature that changes between different events, which is crucial in predicting the difference in waterheight per return period event. For this research, every pixel has the same amount of rainfall.

14. Averaged slope length (slopelength)
While ups is a good feature for showing the entire upstream area, it can create bias due to the high values in the channel, specifically downstream in the floodplain.
To consider the quantity of upstream area for the pixels in the river channel (more upstream area means more runoff that flows down the slope), slope length was used. Using the ldd, it accumulates friction through each cell until the outflow cell is reached with the following formula in PCRaster:

distance x ((friction(sourcecell)+friction(destinationcell))/2

However, this means that the channel cells will accumulate as well, which will still create bias. To correct this, all channel pixels were replaced with an average of six surrounding pixels.

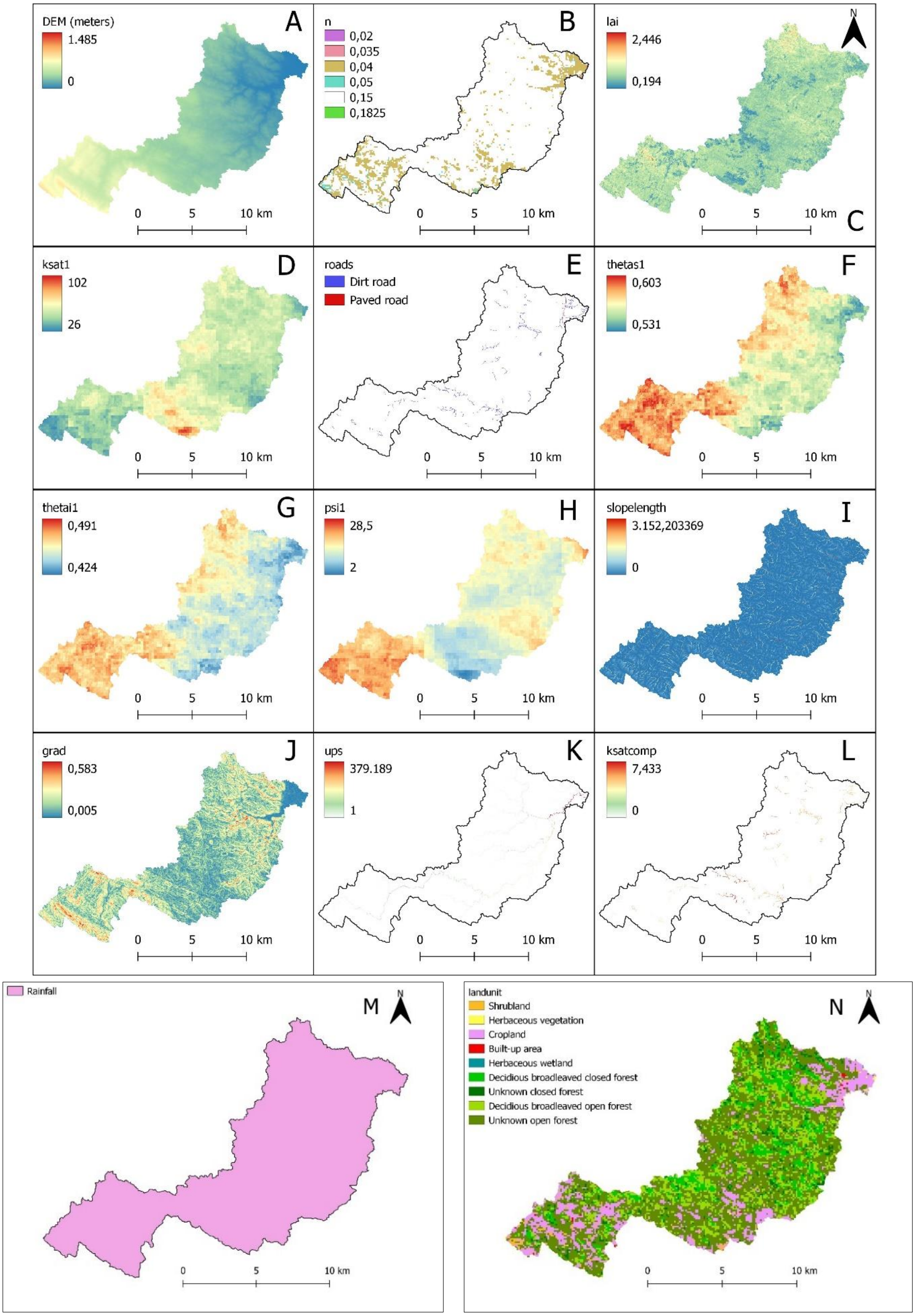


*Figure 10: Maps of all predictors selected for the ML model with a) DEM, b) manning's N, c) LAI, d) Saturated hydraulic Conductivity, e) Roads, f) initial moisture, g) thetai, h) wetting front suction head, i) Slopelength, j) Gradient, k) upstream cumulative, l) ksat of dirt roads, m) *Rainfall and n) land-use*

**Rainfall amount is the same for all pixels*

Feature	Mean	Std	Min	Max
1. DEM (m)	216.69	164.55	0.00	707.30
2. n	0.12	0.05	0.02	0.18
3. lai	0.76	0.20	0.13	2.45
4. ksat	52.44	9.34	27.28	96.14
5. roads*	0.06	0.24	0.00	2.00
6. thetas	0.56	0.01	0.53	0.60
7. thetai	0.45	0.01	0.43	0.48
8. PSI	14.51	3.94	2.73	26.62
9. landuse*	102.80	36.36	20.00	126.00
10. grad	0.12	0.09	0.01	0.62
11. ups	5419.13	33688.23	1.00	397750.00
12. ksatcomp	0.28	1.21	0.00	8.74
13. rainfall (mm)	228.28	36.32	179.13	265.76
14. slopelength	8.67E+27	2.94E+29	0	1.00E+31
15. whmax (10, 50, 100) (m)	0.17	0.56	0	16.79

Table 3: The descriptive statistics for all features and the target variable of the buffer (figure 11).
**The features roads and land-use are both classified, so descriptive statistics for these two features can be disregarded*

2.4.2 Algorithms

For this research, tree-based ensemble algorithms have been chosen. As in literature, the random forest or similar ensemble decision tree algorithms have often proven to generate the most accurate results for flash flood and landslide susceptibility mapping (Pourghasemi, 2020; Pham et al., 2017; Shafizadeh-Moghadam, 2018; Khosravi et al., 2018). Support vector machines were also considered, but they have been excluded because of their increased computational time, especially with larger datasets like the one used for this research (Pourghasemi, 2020).

Random Forest

The core of the random forest algorithm (RF) starts at individual decision trees, as described in Breiman (2001). A decision tree is grown by choosing between features or predictors to get the most significant reduction in error or impurity (node impurity) at each split. This continues until the end of the branch is reached. The end of the branch of a tree, also known as the terminal node or leaf, is reached when another split cannot happen without going under the user-specified minimum amount of samples per node. RF is different compared to a standard single decision tree method as it generates a user-specified number of trees (n_tree), after which the algorithm selects random features of a user-specified amount per node split (m_try). Each tree then provides a predicted value, and the values of all the trees are averaged in one single value. This random sampling of features and data points gives the RF algorithm its strength, as it does not rely on a single tree that only has one combination of samples of features, but many trees all grown with random samples and features. The averaging of all tree predictions allows for the growth of decision trees with a high variance while not having to take as big of a bias penalty. This means that overfitting is less of a problem with RF algorithms. Additionally, because decisions in each node can be tracked, it is also possible to find feature importances. Here, the most important feature is the feature that reduces the amount of error the most in each split, which provides further valuable insight and possibilities for feature selection.

Lastly, RF has shown flexibility regarding input data, often needing limited pre-processing and being able to train and predict with big datasets in a relatively short timeframe.

There are, however, disadvantages as well. For instance, depending on the use case, RF might not be useful as the algorithm cannot extrapolate. This means that a trained random forest model is not able to predict beyond the minimum and maximum target variable value it has been trained on (Zurita-Milla, 2020, Muñoz et al., 2018; Pourghasemi, 2020). For the application of the random forest algorithms in this research, this is not an issue. As return period events and the maximum and minimum amount of rainfall coupled to said events can be calculated beforehand. With this data available and the events that will be predicted, extrapolation is not needed.

Extra Randomized Trees

The extra randomized trees (EXRF) algorithm is nearly identical to the standard random forest algorithm. There is, however, one main difference. EXRF is similar to RF in sampling random features for splitting the parent node into two children nodes. However, where RF picks a feature to split the node based on the greatest reduction in error, EXRF splits a node randomly without looking at which feature has a more significant reduction. This increases the randomness and, therefore, the variance but slightly increases bias. EXRF has been shown to deal better with noise in datasets (Geurts et al., 2006; Pedregosa et al., 2011; The Kerneltrip, 2018; Lawson et al., 2017).

Extreme Gradient Boost

Gradient boosting algorithms are also similar to RF in the way that both methods generate a forest of decision trees. However, the main differences can be found in how the trees are built and how the prediction results are combined. With gradient boosting, the forest is built each tree one at a time, compared to the more random and independent tree building in RF. By building one tree at a time, it is possible to evaluate and identify weak learners or observations that are difficult to predict. Like RF, the individual trees are grown by minimizing the loss or error at each split, the main difference with the RF algorithms is that gradient boosting reduces error by reducing bias and RF reduces error by reducing variance.

The goal is then that the next tree that is built improves upon the predictions of the first tree, which means that the predictions of the final gradient boosting model is the weighted sum of the predictions of the previously grown trees (Brownlee, 2020; Harshdeep, 2018; Morde, 2019; XGBoost, n.d.).

Extreme Gradient Boost (XGB) is an optimized gradient boosting algorithm that was created for maximum speed and performance by optimizing software and hardware usage (Morde, 2019; Brownlee, 2016; Chen & Guestrin, 2016).

2.4.3 ML model performance

The performance of each ML algorithm will be monitored and validated by splitting the training dataset set into training and validating sets, of which 70% of the data is used for training and the remaining 30% is used for validating. The performance will be tested for the baseline, feature selection and post hyperparameter tuning with 10 fold cross-validation, shown in figure 10. When validating the trained algorithm, three metrics will be considered: the R-squared (R^2), the mean absolute error (MAE) and the root-mean-square-error (RMSE). Generally, to have a high prediction accuracy, the R^2 should be as close as possible to 1, and both the MAE and RMSE are as low as possible.

2.4.4 Feature selection and hyperparameter tuning

The feature selection has been made using sci-kit's Recursive feature elimination with built-in cross-validation (RFECV) to select the best feature set automatically. The function automatically ranks the features according to their importance, after which it starts eliminating the least important feature from the feature set sequentially. RFECV then evaluates the cross-validated scores and highlights the best performing feature set.

Machine learning algorithms and their performance can be improved by tuning hyperparameters. These parameters define the thresholds and boundaries within which the algorithms operate. Because each ML algorithm is different in the way they operate, they all have different hyperparameters. For example, previously mentioned hyperparameters for ensemble methods such as RF or EXRF are the m_try and n_tree. XGBoost, because of the different way of operating, also has essential hyperparameters such as learning_rate and gamma to tune the minimum amount of loss to split a node.

This hyperparameter tuning was done by running the model and looking at the ML performance metrics and using the gridsearchCV function of sci-kit learns python package.

A buffer was made on which the algorithms were trained to cut down on processing time, shown in figure 11 below. The buffer was made using slope gradient as a constriction and referenced to the maximum waterheight of the highest return period event to make sure all or most of the flooding for every event was inside of the buffer.

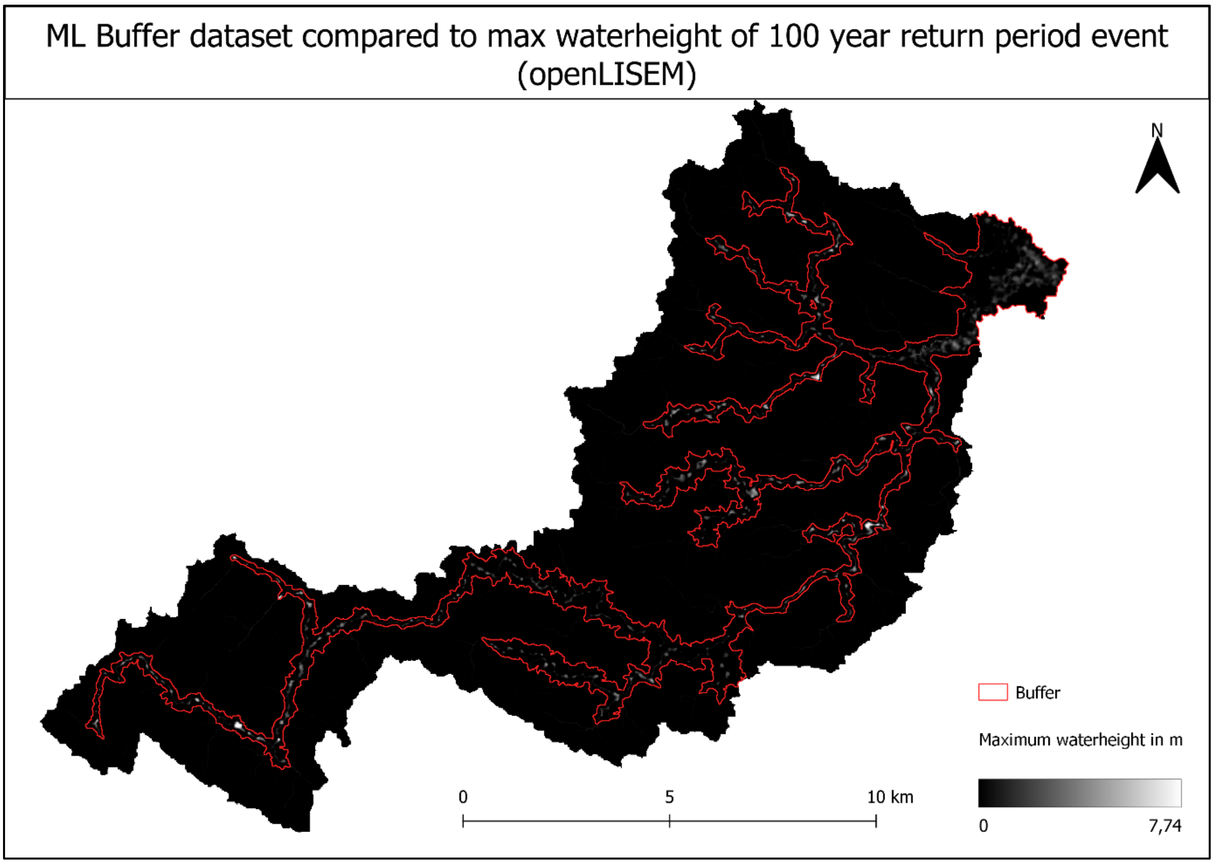


Figure 11: The created buffer of the channel in the study area to include maximum return period event flood extent

2.5 ML model output and openLISEM output comparison

This research aims to evaluate and compare the maximum flood height maps from both the openLISEM and the ML model. First, the prediction performance for both a 20 and an 80 year return period event was calculated. Consecutively, the difference between the openLISEM flood model and ML prediction was quantified and visualized by calculating the difference between the two models, which shows where the different ML algorithms are underestimating and overestimating compared to the openLISEM flood model. Furthermore, the maximum flooded area, percentage of flooded pixels, and maximum flood volume were calculated for the openLISEM and ML models and compared.

3. Gumbel analysis and synthetic event creation results

The differences between the bias correction methods to correct the hourly data for the CHIRPS rainfall dataset can be seen in the cumulative distribution function curves (CDF) in figure 12, with the R package and manual excel bias correction named GSMaP_monthly_daily_corrected and GSMaP_daily_corrected, respectively. The closer to chirps, the better, as CHIRPS already contains ground station corrected satellite data, leading to the choice of the manual excel bias_correction method to generate hourly bias-corrected GSMaP data as it matches the Chirps dataset exactly.

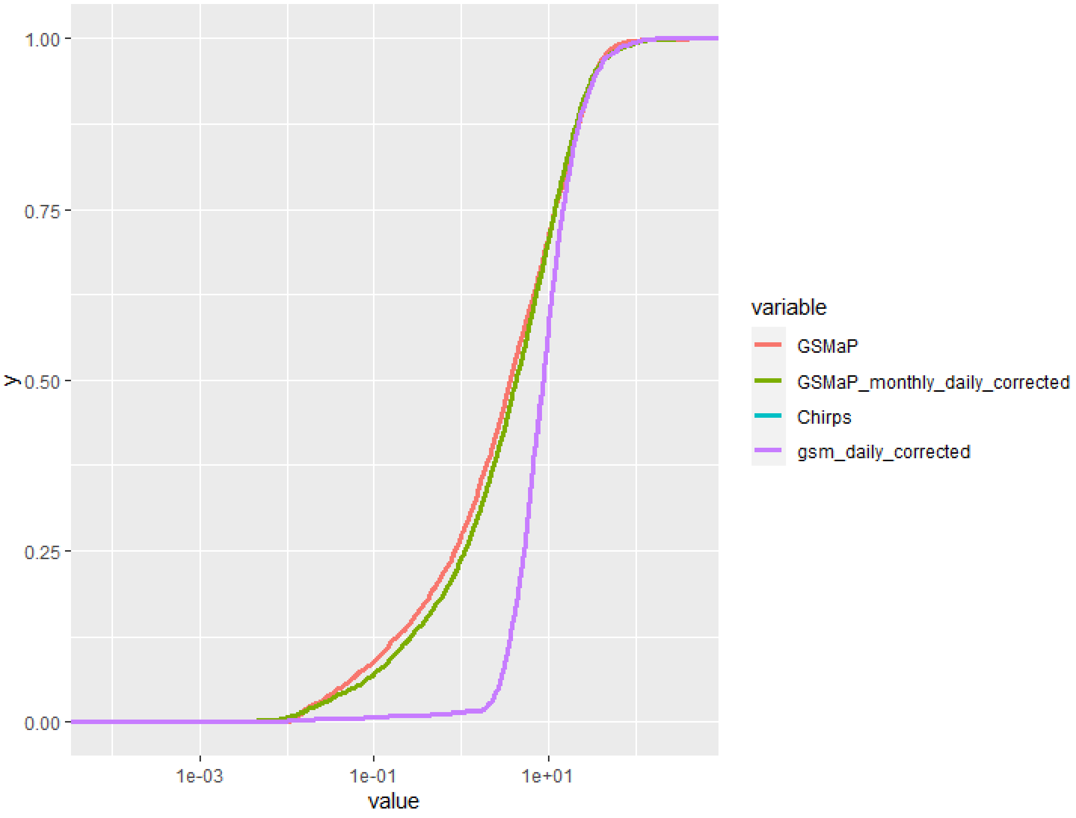


Figure 12: CDF curves showing the differences between the two different bias correction methods

The bias-corrected data was then used to calculate the IDF curves (see figure 12), which was done, as explained previously in paragraph 2.3.3, with a GUMBEL analysis for every time interval and every return period. The results of this GUMBEL analysis can be seen in table 6.
IDF curves were created with the bias-corrected data for 2, 5, 10, 20, 50, 80 and 100 year return periods, with intervals of 2, 4, 8, 12 and 24 hours.

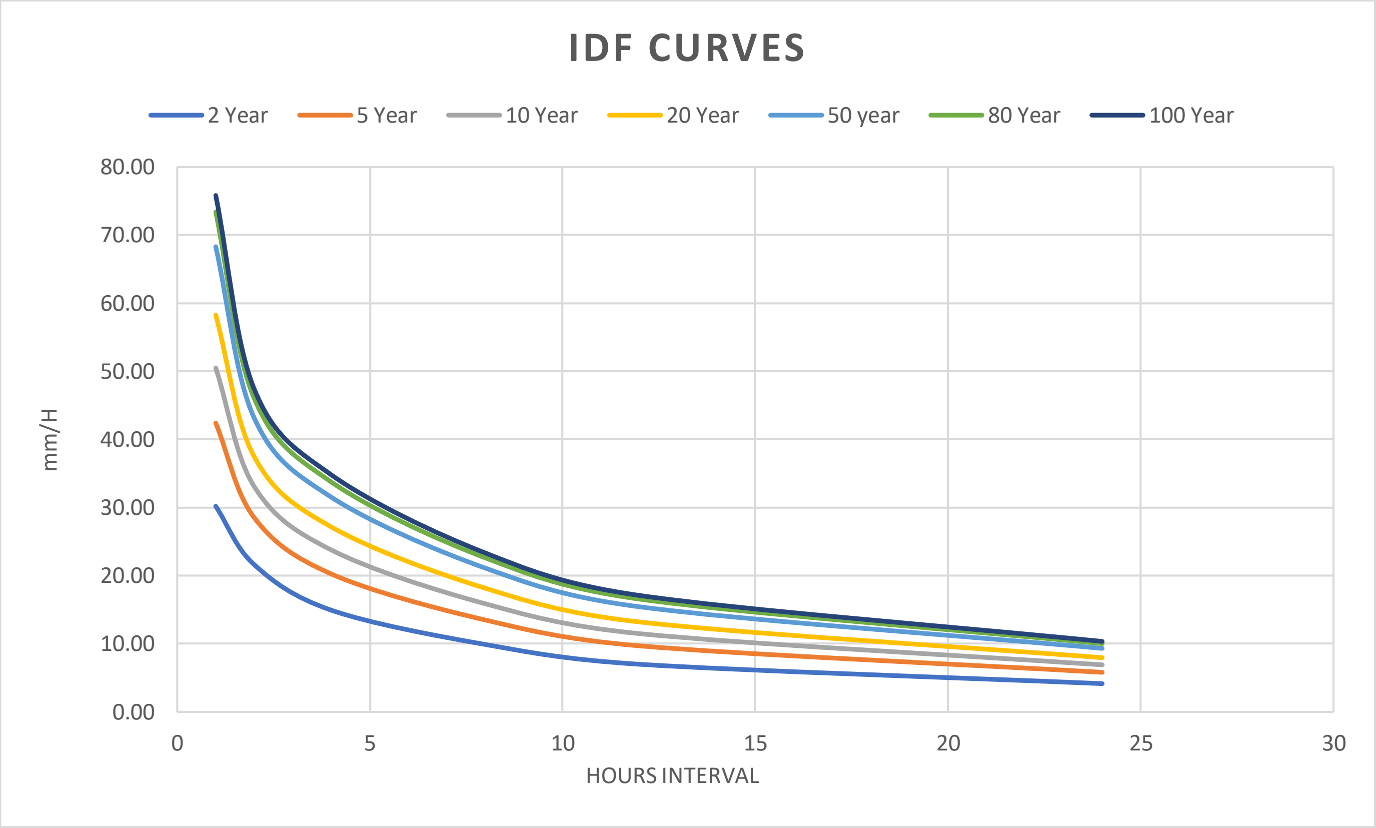


Figure 13: IDF curves for the 7 different return periods for a maximum of 24 hours.

Return period T	Duration in hours						Formula
	1h	2h	4h	8h	12h	24h	
2y	30.20	21.61	14.93	9.89	6.99	4.15	$y = 32.810x^{-0.621}$ $R^2 = 0.989$
5y	42.41	28.50	20.25	13.48	9.69	5.81	$y = 44.582x^{-0.616}$ $R^2 = 0.9921$
10y	50.50	33.07	23.77	15.86	11.48	6.91	$y = 52.364x^{-0.613}$ $R^2 = 0.9929$
20y	58.26	37.45	27.15	18.15	13.20	7.96	$y = 59.824x^{-0.611}$ $R^2 = 0.9932$
50y	68.30	43.12	31.52	21.10	15.43	9.33	$y = 69.476x^{-0.61}$ $R^2 = 0.9934$
80y	73.41	46.00	33.74	22.61	16.56	10.02	$y = 74.382x^{-0.609}$ $R^2 = 0.9956$
100y	75.83	47.36	34.79	23.32	17.09	10.35	$y = 76.706x^{-0.609}$ $R^2 = 0.9934$

Table 4: The results of the GUMBEL analysis done per time interval with the individual curve formulas

The curve formulas displayed in the last column of table 6 made it possible to create the 24 hour long synthetic events. Using the previously described (see paragraph 2.3.2) alternating block method, they become as symmetrical as possible, as shown in figure 14. The total rainfall per return period event can be found in table 7.

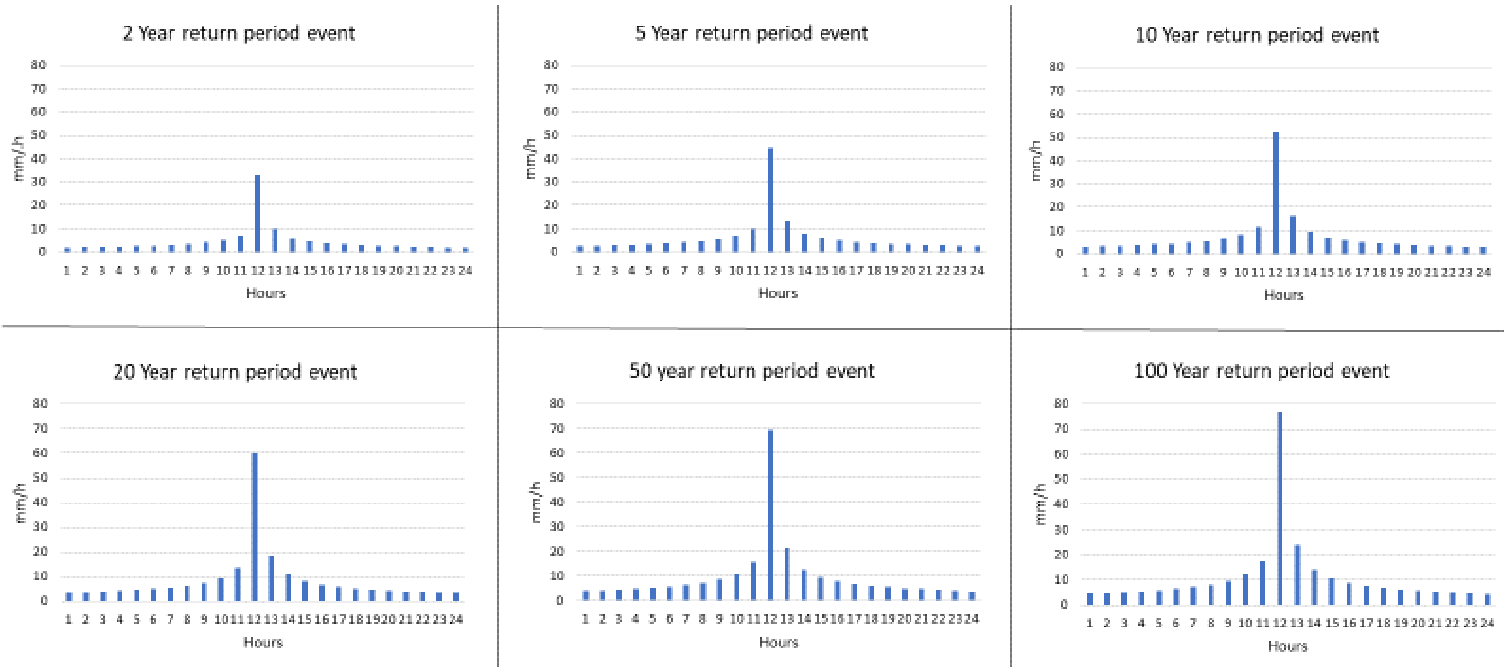


Figure 14: The synthetically created return period events

Return period (Years)	Total (mm)
2	109.4
5	151.0
10	179.1
20	205.9
50	239.9
80	257.7
100	265.8

Table 5: Total rainfall per synthetic return period event in millimetres

4. openLISEM results

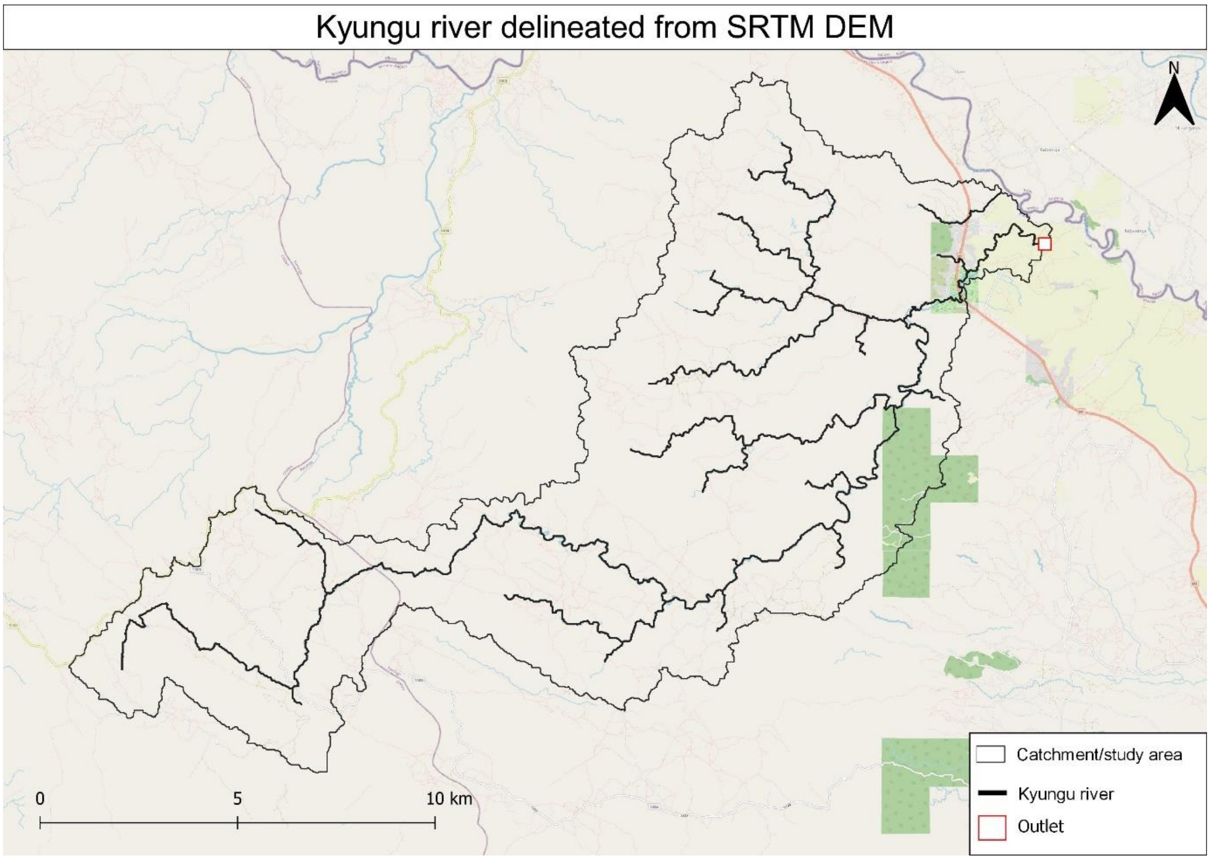


Figure 15: The delineated Kyungu river network calculated from the SRTM based ldd flow network

The Kyungu river catchment has many branches, as shown in figure 15, with an approximate total length of 95 km. This length can be seen in the different hydrographs, where, for all events, the discharge at the outlet point increases and decreases after the initial peak rainfall has fallen and water from the immediate vicinity of the outlet point starts to collect and flow towards it. Then after 23.6 hours for the 10 year, 20.5 hours for the 50 year and 20 hours for the 100 year return period events, peak discharge occurs, which can be seen in figure 16a, 17a and 18a, respectively. With the peak rainfall happening between the 10th and 11th hour for every event, it can be seen that the lag time between the peak rainfall and peak discharge decreases as the return period increases with a lag time of 13.5, 10.5 and 10 hours for 10, 50 and 100 year return period events respectively. After calibration in openLISEM, the simulations for all events were run for 2200 minutes or about 37 hours (simulated time) as that was when the discharge returned to normal.

Peak discharge, max flood volume and max flood area all increased when the return period increased, as seen in table 8 below:

Return period	**Peak discharge l/s**	**Max flood volume m^3**	**Max flooded area m^2**
2	33771.64	24	400
5	45816.23	131.864	900.800
10	95578.92	922.304	2.537.600
50	166339.17	6.108.668	7.576.000
100	190057.76	8.019.072	9.264.800

Table 6: The flood characteristics of the openLISEM simulations of different return period events

Maximum waterheight and the flood pattern of the 50 and 100 year return period events are similar. In comparison, a relatively low amount of flood volume and flood area can be observed in table 8 for the 10 year return period event. The results for the 2 and 5 year return period were not used, as near-zero flooding occurred. Hydrographs and flood maps for the 2 and 5 year return period events can be found in Appendix D.

In the maximum waterheight maps, clusters of pixels with extremely high waterheight values can be observed, which can be a result of pits in the DEM. These pits are partly a consequence of the 20x20 meter resolution but are primarily caused by inaccuracies in the DEM, which prevents the water from flowing out of the pit resulting in pools.

4.1 10 Year return period event

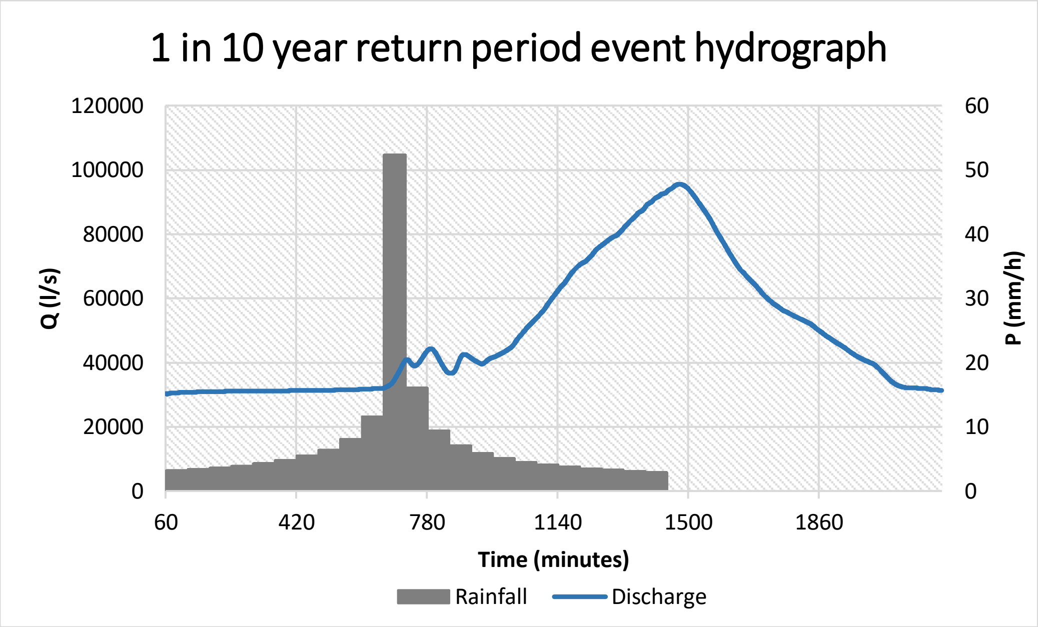


Figure 16a: Hydrograph for a 10 year return period event

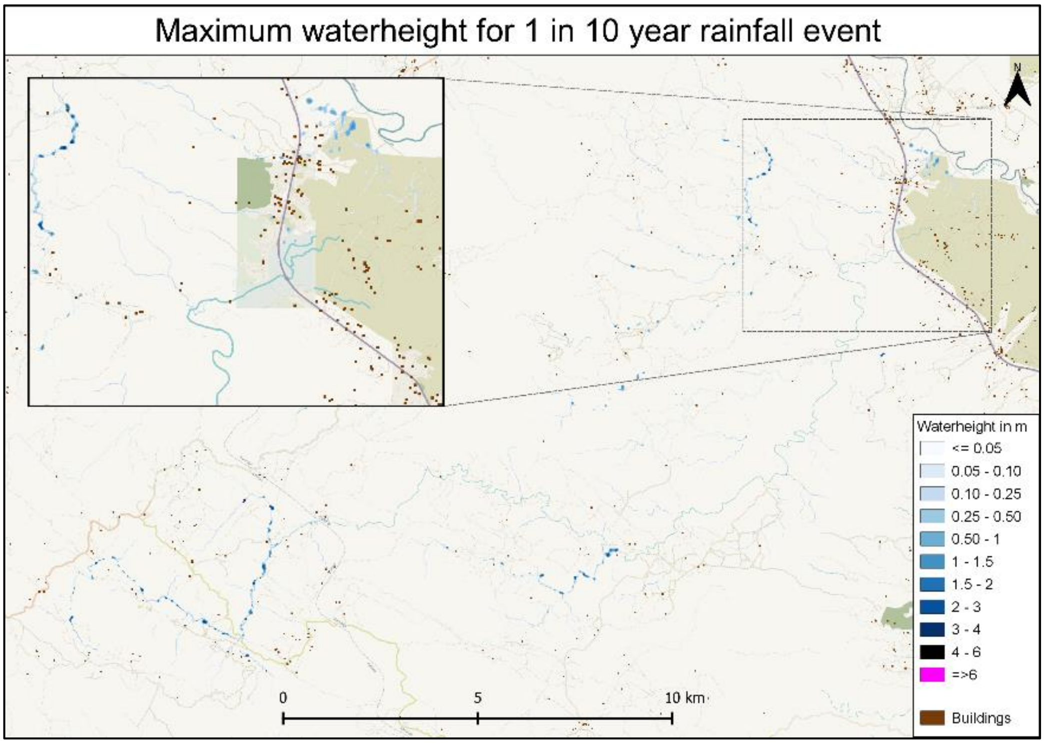


Figure 16b: Maximum waterheight map generated for a 10 year return period event using openLISEM

4.2 50 year return period event

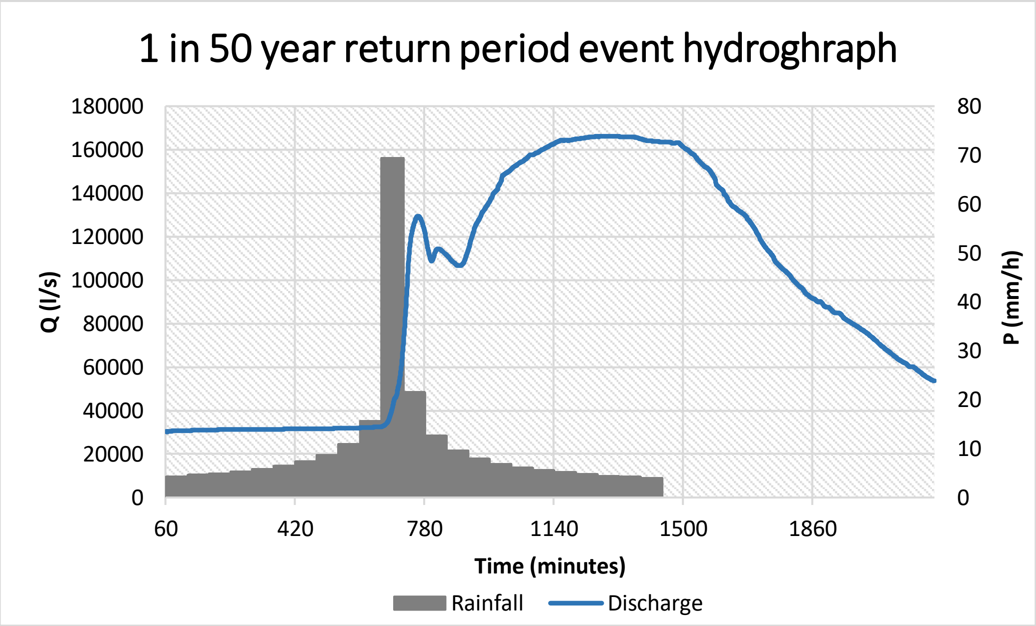


Figure 17a: Hydrograph for a 50 year return period event

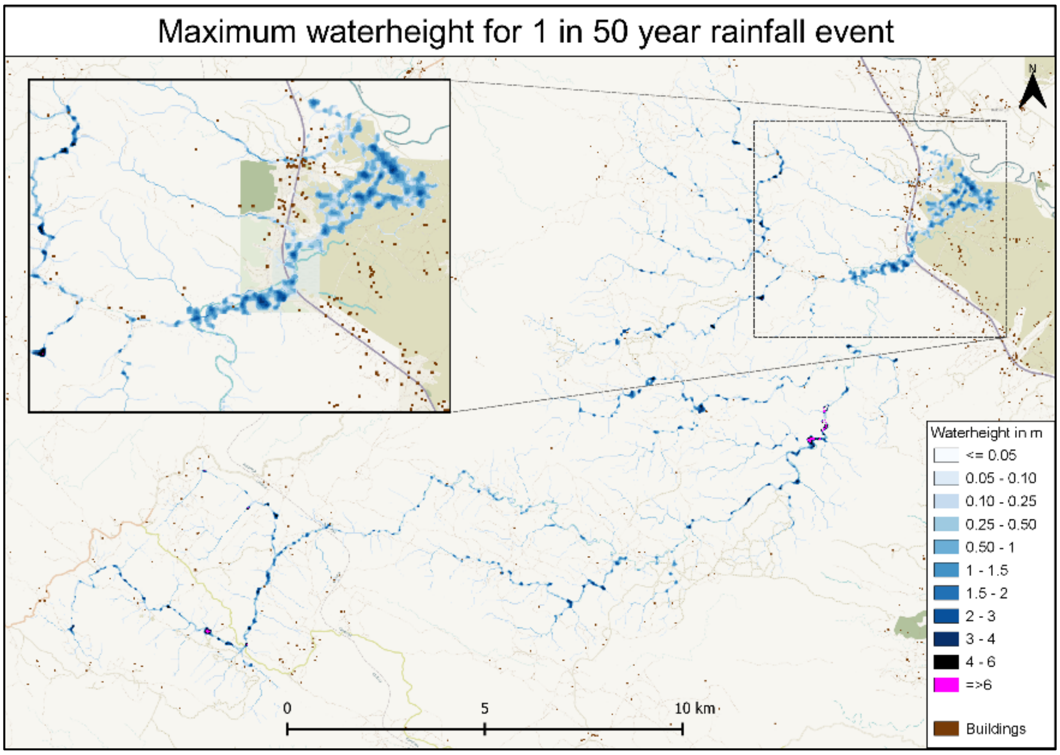


Figure 17b: Maximum waterheight map generated for a 50 year return period event using openLISEM

4.3 100 year return period event

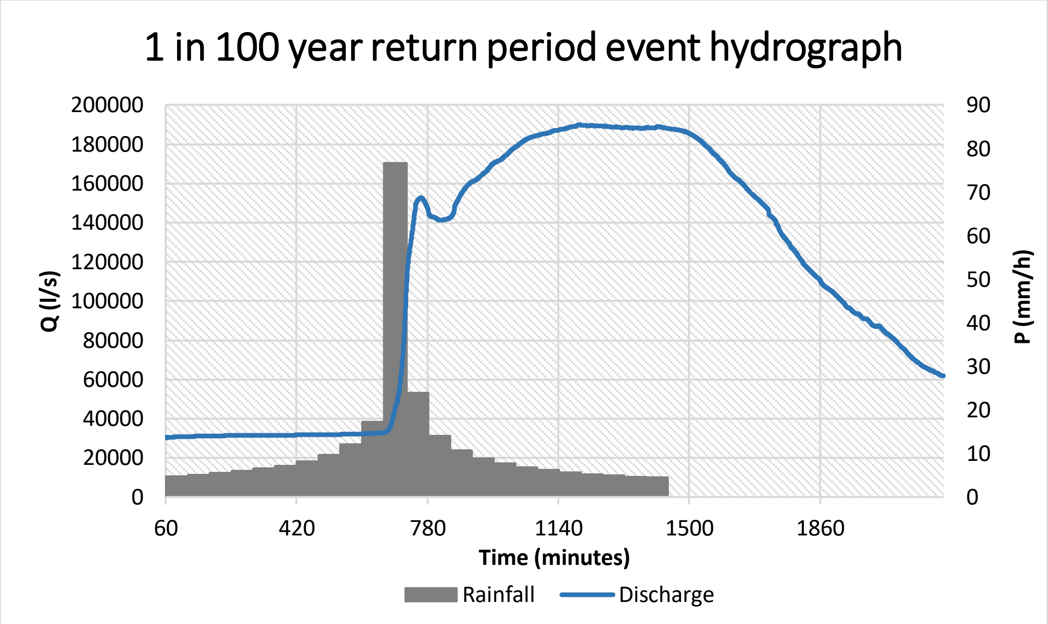


Figure 18a: Hydrograph for a 100 year return period event

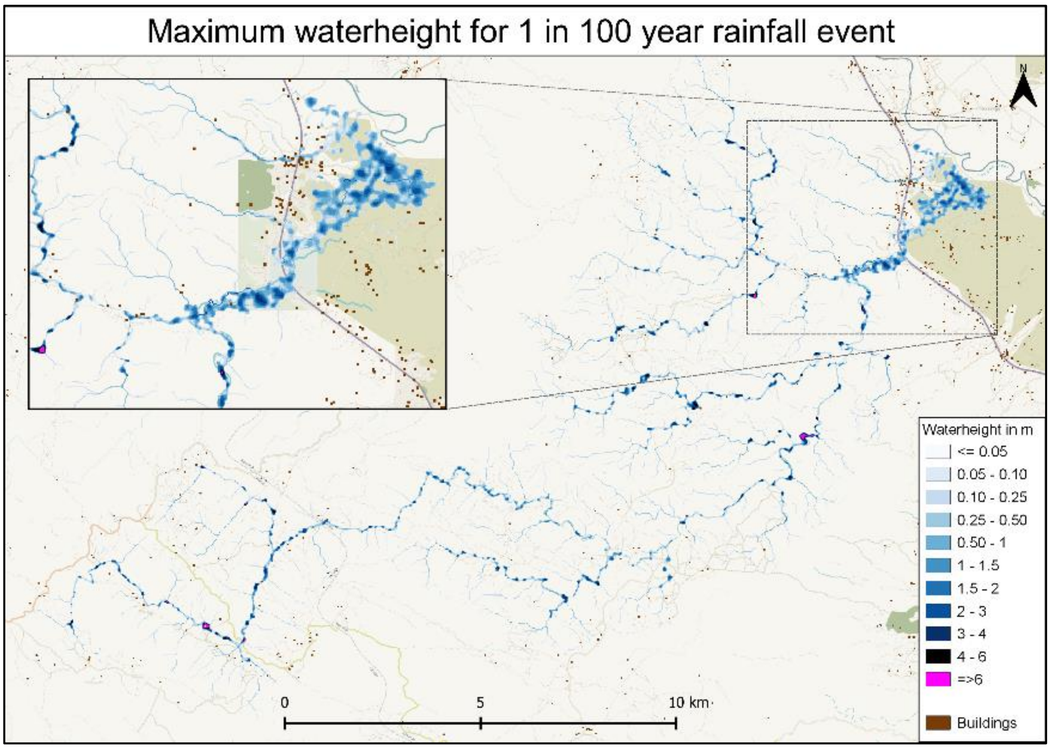


Figure 18b: Maximum waterheight map generated for a 100 year return period event using openLISEM

5. ML model training and validation

5.1. Baseline performance

As shown in figure 9, the first step in tuning and generating the results is establishing a baseline of the performance of the different algorithms only using default settings, which can be seen in table 9 below:

10 fold cross-validated Baseline performance

Algorithm	R^2 mean	R^2 std	MAE mean	MAE std	RMSE mean	RMSE std
LM	0.001	0.001	0.248	0.004	0.304	0.018
DT	0.436	0.061	0.096	0.004	0.177	0.023
RF	0.681	0.025	0.094	0.002	0.099	0.014
EXRF	0.684	0.016	0.104	0.003	0.099	0.013
XGB	0.255	0.017	0.188	0.003	0.222	0.018

Table 7: The 10 fold cross-validated baseline performance results for LM, DT, RF, EXRF and XGB algorithms

Baseline

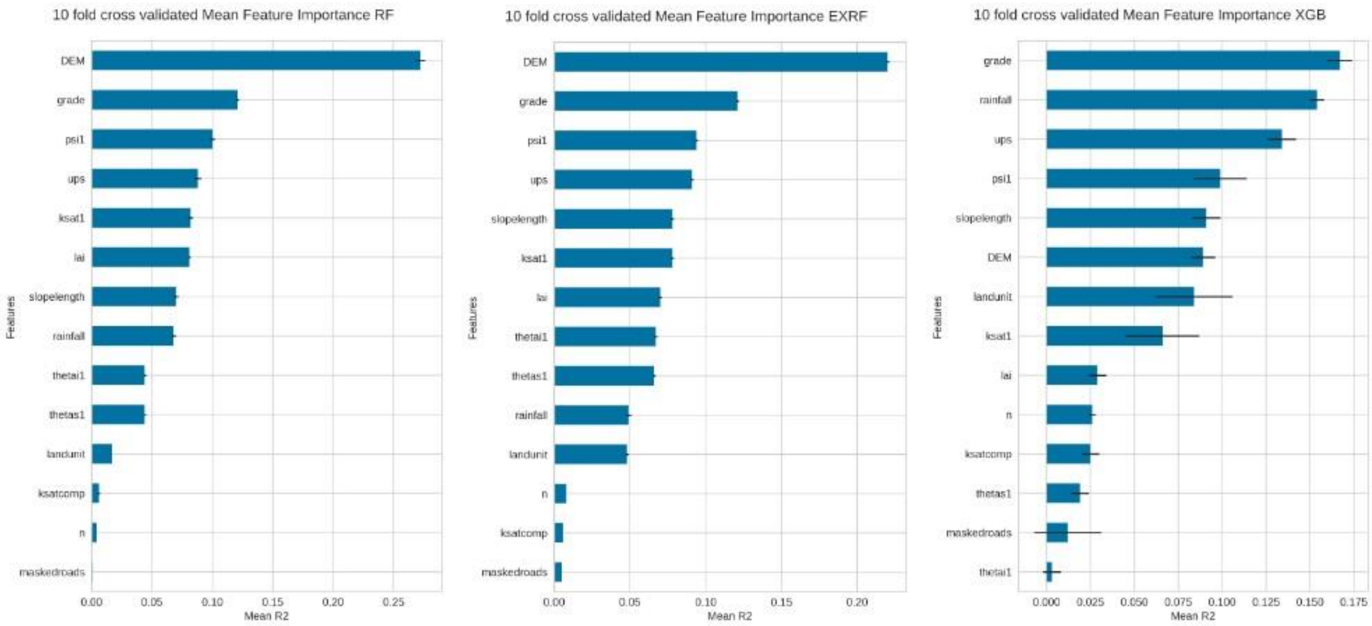


Figure 19: Baseline feature importance ranking from most important to least important for the RF, EXRF and XGB algorithms

The ML models have been trained on a dataset containing all features mentioned in paragraph 2.4.1 and with the target variable (maximum waterheight) of a 10 year, 50 year and 100 year return period event. Table 9 above shows what the performance is for the raw algorithms.
RF and EXRF show similar feature importance ranking, with DEM in both being the most important feature, as well as the least important feature, roads, being the same.
The XGB algorithm shows a different initial ranking than the other two ensemble algorithms, with gradient, rainfall, land-use, ups and DEM as the top 5 most important features. Additionally, compared to the RF and EXRF algorithms, XGB shows a more significant standard deviation for every feature.

5.2. Feature selection

The feature selection was based on the best performing algorithm, which was the EXRF algorithm. However, the RF algorithm is also quite similar in performance, with the RF and EXRF often switching positions of the best performing algorithm by small margins. Figure 20 below shows the sequential cross-validated recursive feature selection results, where the feature with the lowest importance is eliminated from the training set each iteration. The first run has all fourteen features selected and the last run only has zero features.
As shown in figures 20 and 19, after the fourth feature, land-use, has been eliminated from the dataset, the model's accuracy decreases significantly, indicating that land-use and all features ranked above are important in predicting waterheight.

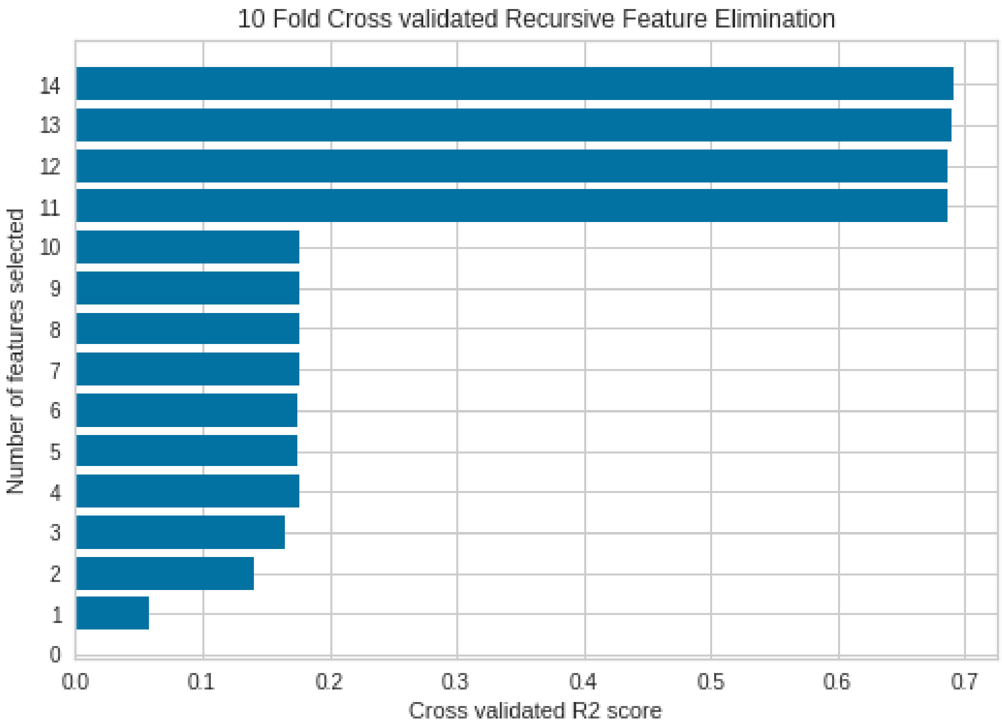


Figure 20: The 10 fold cross-validated Recursive Feature Selection results, showing the 10 fold average R^2 score per run and with the first run including all features (14) and the last run none.

5.3 Post hyperparameter tuning performance

Hyperparameter tuning has not significantly increased the performance for EXRF and RF. Instead, it has increased the R^2 standard deviation, as seen in table 10 below, while hyperparameter tuning showed more significant model performance improvement in non-cross validated models. This is not reflected here. Hyperparameter tuning mainly focused on m_try and n_tree, where for all algorithms, an increase in n_tree provided an increase in performance, especially for the XGB algorithm.

Table 8: Results of the 10 fold cross-validated performance post hyperparameters tuning for the RF, EXRF and XGB algorithms
**RF was cross-validated with 8 folds due to hardware limitations*

10 fold cross-validated Post hyperparameter tuning performance							
Algorithm	**R^2 mean**	**R^2 std**	**MAE mean**	**MAE std**	**RMSE mean**	**RMSE std**	**Fit time (sec)**
RF*	0.682	0.03	0.095	0.002	0.097	0.012	335
EXRF	0.685	0.03	0.103	0.002	0.097	0.013	107
XGB	0.722	0.024	0.103	0.003	0.087	0.014	51

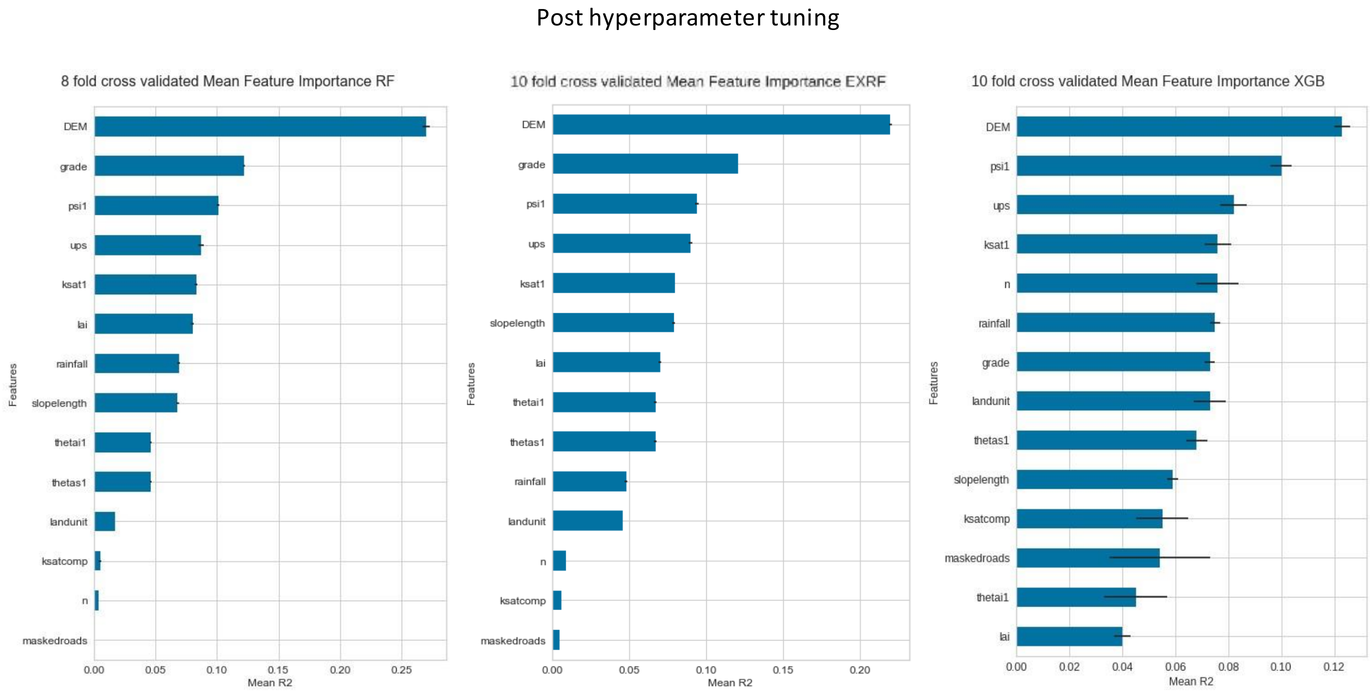


Figure 21: The post hyperparameter tuning feature importance ranking from most important to least important for the RF, EXRF and XGB algorithms

Besides a significant increase in accuracy, the XGB algorithm also is the fastest, which can be attributed to the focus on decreasing memory usage and calculation time when the algorithm was developed.

The feature importance ranking after hyperparameter tuning generally remains the same, with the main changes in ranking happening for the XGB algorithm. However, similarly to the baseline cross-validated feature importances, the features have a high standard deviation which could lead to different rankings for different runs.

6. ML maximum waterheight prediction results

The ML algorithms have been trained on 10, 50 and 100 year return period events of which the water height maps can be seen in chapter 4, and the performance of the trained models in chapter 5. To predict the maximum waterheight, two different return period events, a 20 and 80 year return period event, were created and used. Because these events have not been used in training, data leakage has been prevented as much as possible. Because these events were also run in openLISEM, it is possible to see the prediction performance for these two events.

6.1 20 year return period event prediction

20 Year return period prediction performance			
Algorithm	**R^2**	**MAE**	**RMSE**
RF	0.353	0.07	0.27
EXRF	0.599	0.056	0.213
XGB	0.482	0.079	0.244

Table 9: 20 year return period prediction performance for the RF, EXRF and XGB algorithms

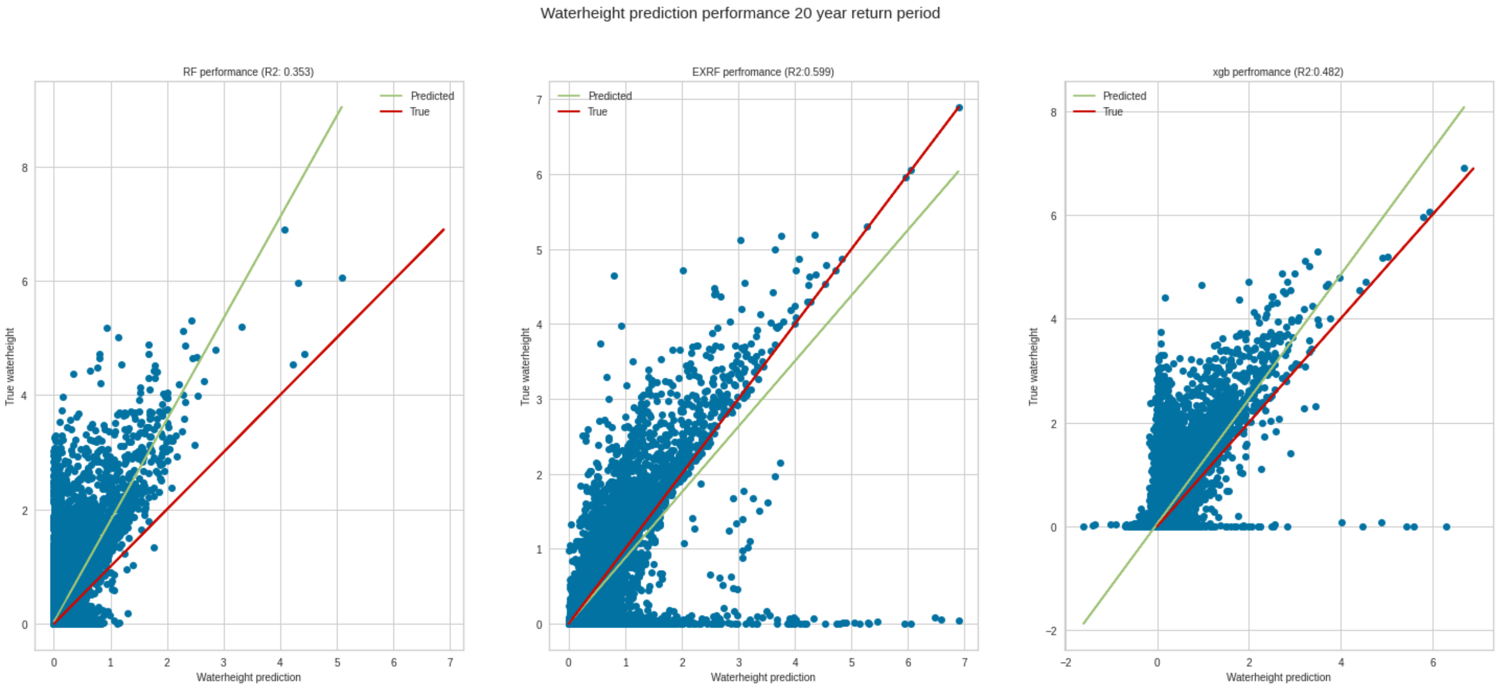


Figure 22:The maximum waterheight prediction performance for the 20 year return period event, showing the difference between the true/observed values (red) and the predicted values (green) for the RF, EXRF and XGB algorithms

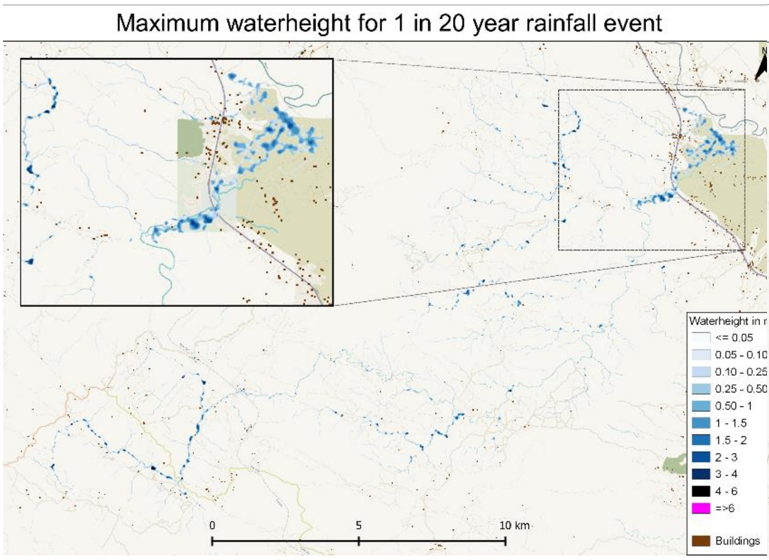


Figure 23a: 20 year return period maximum waterheight map simulated by openLISEM

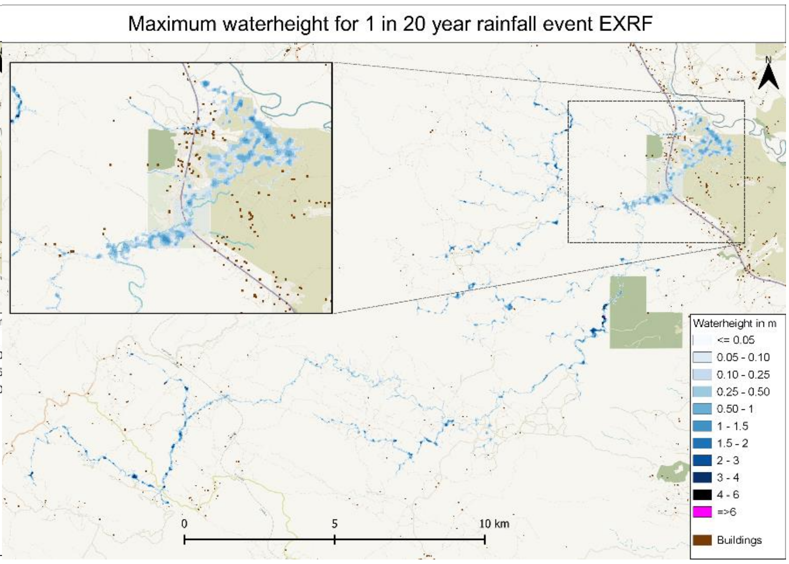


Figure 23b: 20 year return period maximum waterheight map predicted by the ML model using the EXRF algorithm

Comparison of flood characteristics for 20 year return period event

	Max flood area m^2 (Buffer)	% of flooded (Buffer)	Max flood volume m^3 (Buffer)
openLISEM	4.206.000	14.25%	2.459.363
ML EXRF	7.056.000	23.90%	2.705.309
Difference	2.850.000	9.65%	245.946

Table 10: Quantitative comparison of the openLISEM simulation and the ML model EXRF predicted maximum waterheight maps for the 20 year return period event. Flooded pixels are pixels with a maximum waterheight of >=0.05m

For the 20 year return period prediction, the EXRF algorithm performs the best and the resulting maximum waterheight map can be seen in figure 23b. When compared with the openLISEM maximum waterheight map, the flood patterns are relatively similar with water collecting in areas where, according to the true openLISEM values, water is supposed to collect. However, when visually compared, there are more flooded pixels, and more flooding seems to occur midstream and upstream for the ML prediction. This is supported by the calculated area that is flooded, where the flooded amount of square meters is significantly higher. Overall the ML prediction is underestimating, as shown in figure 22.

As displayed in figure 24, most of the pixels in the study area fall within the MAE. However, downstream in the floodplain, the ML model is generally underestimating. Further upstream, there is an area that is severely overestimated with a maximum overestimation (positive) of 6.90 meters and a maximum underestimation (negative) of 3.90 meters.

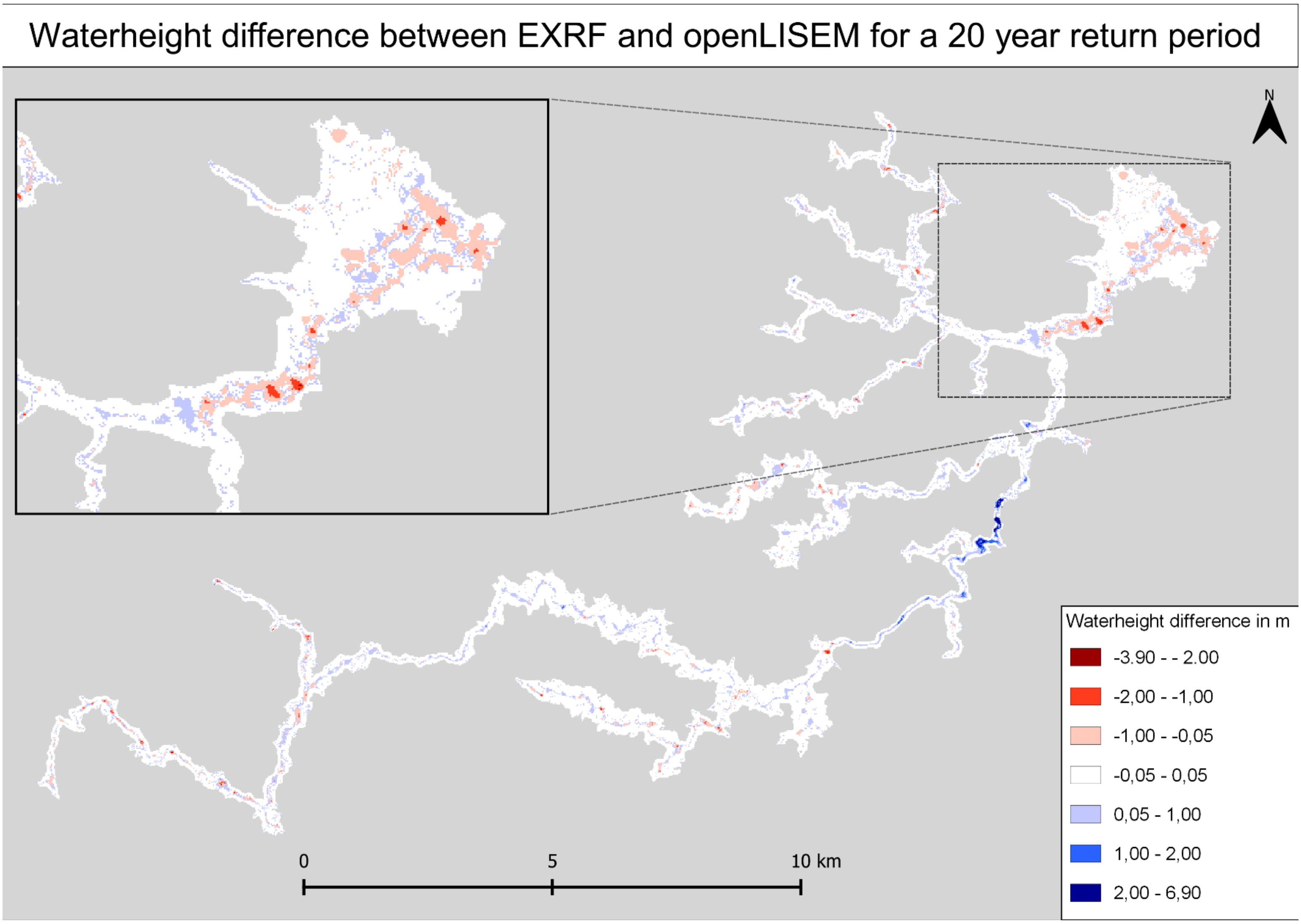


Figure 24: Difference in waterheight between the ML EXRF prediction and the openLISEM simulation for a 20 year return period event. With the ML model overestimating compared to openLISEM for the positive numbers (blue) and underestimating for the negative numbers (red)

6.2 80 year return period event prediction

80 Year return period prediction performance			
Algorithm	**R^2**	**MAE**	**RMSE**
RF	0.892	0.070	0.214
EXRF	0.914	0.050	0.190
XGB	0.913	0.075	0.192

Table 11: 80 year return period prediction performance for the RF, EXRF and XGB algorithms

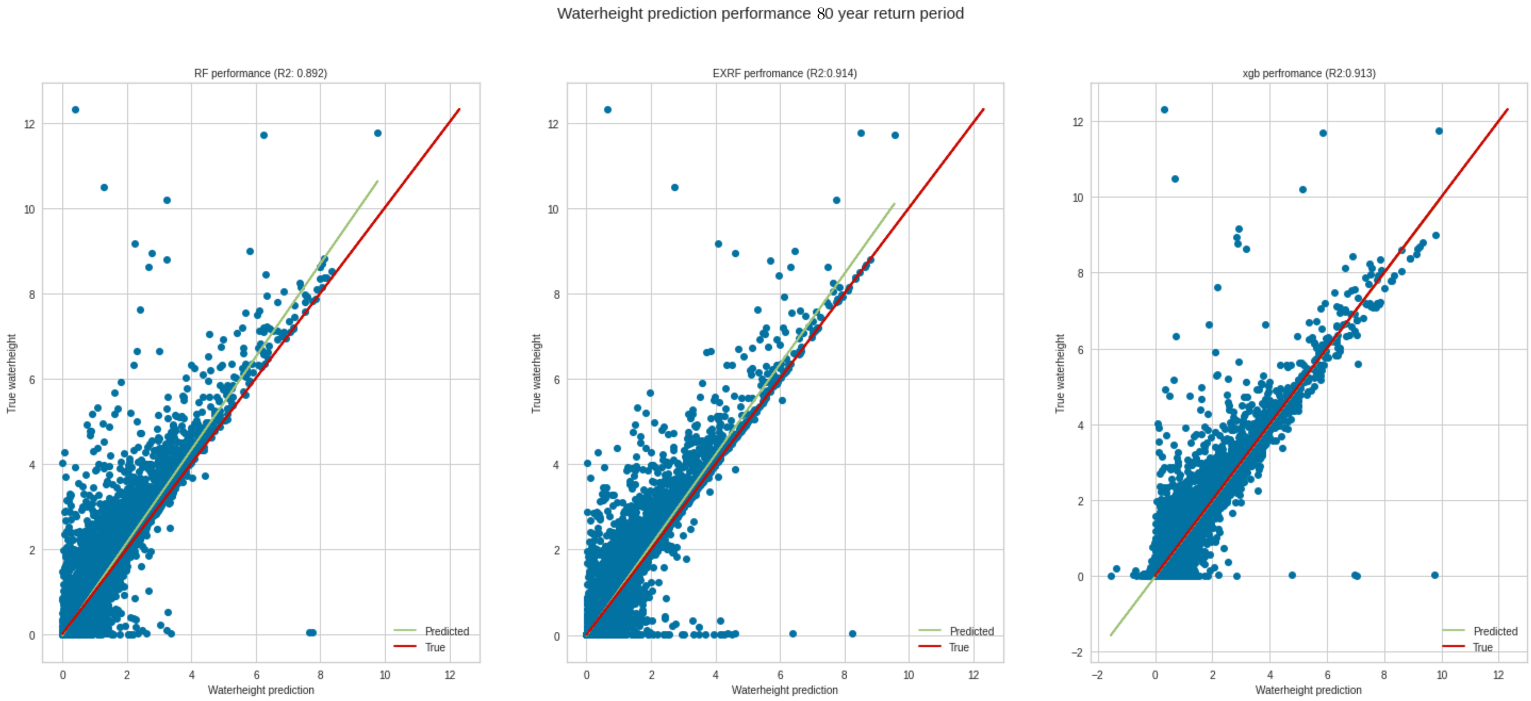


Figure 25: The maximum waterheight prediction performance for the 80 year return period event, showing the difference between the true/observed values (red) and the predicted values (green) for the RF, EXRF and XGB algorithms

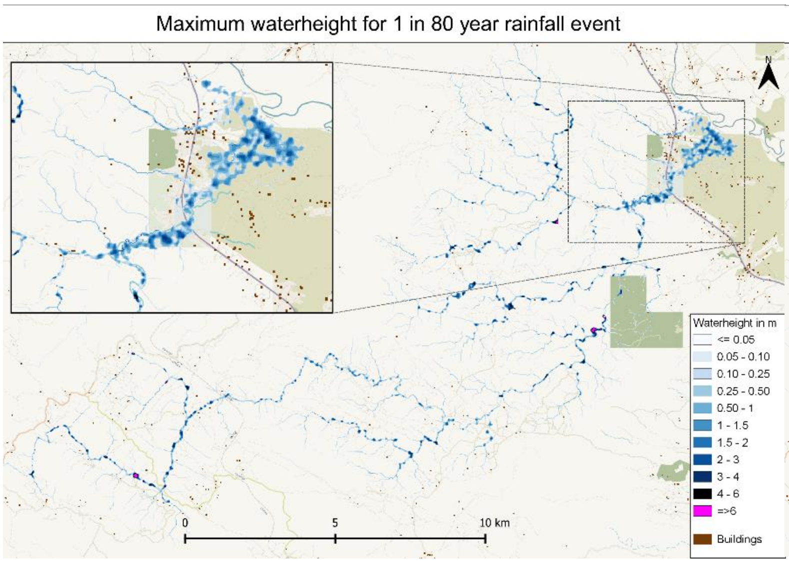


Figure 26a: 80 year return period maximum waterheight map simulated by openLISEM

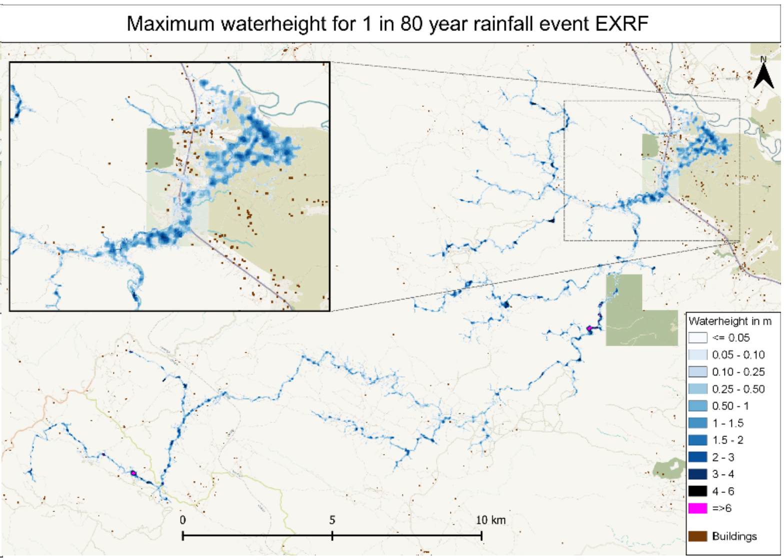


Figure 26b: 80 year return period maximum waterheight map predicted by the ML model using the EXRF algorithm

Comparison of flood characteristics for 80 year return period event

	Max flood area m^2 (Buffer)	% of pixels flooded (Buffer)	Max flood volume m^3 (Buffer)
openLISEM	7.687.600	26.04%	6.554.384
ML EXRF	10.358.400	35.09%	6.577.111
Difference	2.670.800	9.05%	22.727

Table 12: Quantitative comparison of the openLISEM simulation and the ML model EXRF predicted maximum waterheight maps for the 80 year return period event. Flooded pixels are pixels with a maximum waterheight of >=0.05m

For the 80 year return period prediction, the EXRF algorithm performs the best and the resulting maximum waterheight map can be seen in figure 26b. The accuracy for the 80 year return period event prediction is high, with an R^2 0.91, an MAE of 0.05 and RMSE of 0.19, as shown in table 13. The high accuracy results in a prediction that is decidedly similar to the true openLISEM simulation, especially in the floodplain downstream.

Similarly to the 20 year return period prediction, the amount of flooded square meters for the 80 year return period prediction is significantly higher compared to openLISEM simulation for the same event, with an almost equal overestimation of flooded pixels for both return period predictions.

The 80 year prediction is overestimating slightly, as can be seen in figure 25.

As shown in figure 27, the maximum overestimation and underestimation maximum waterheight values are 8.20 and 11.7 meters, respectively.

Similarly to the 20 year prediction, the same upstream area proves challenging to predict, with near maximum overestimation and underestimation values occurring within the same area.

Contrary to the 20 year prediction, the floodplain has been more accurately predicted for the 80 year prediction, with overestimated and underestimated pixels being mixed and differences rarely exceeding the 1-meter category.

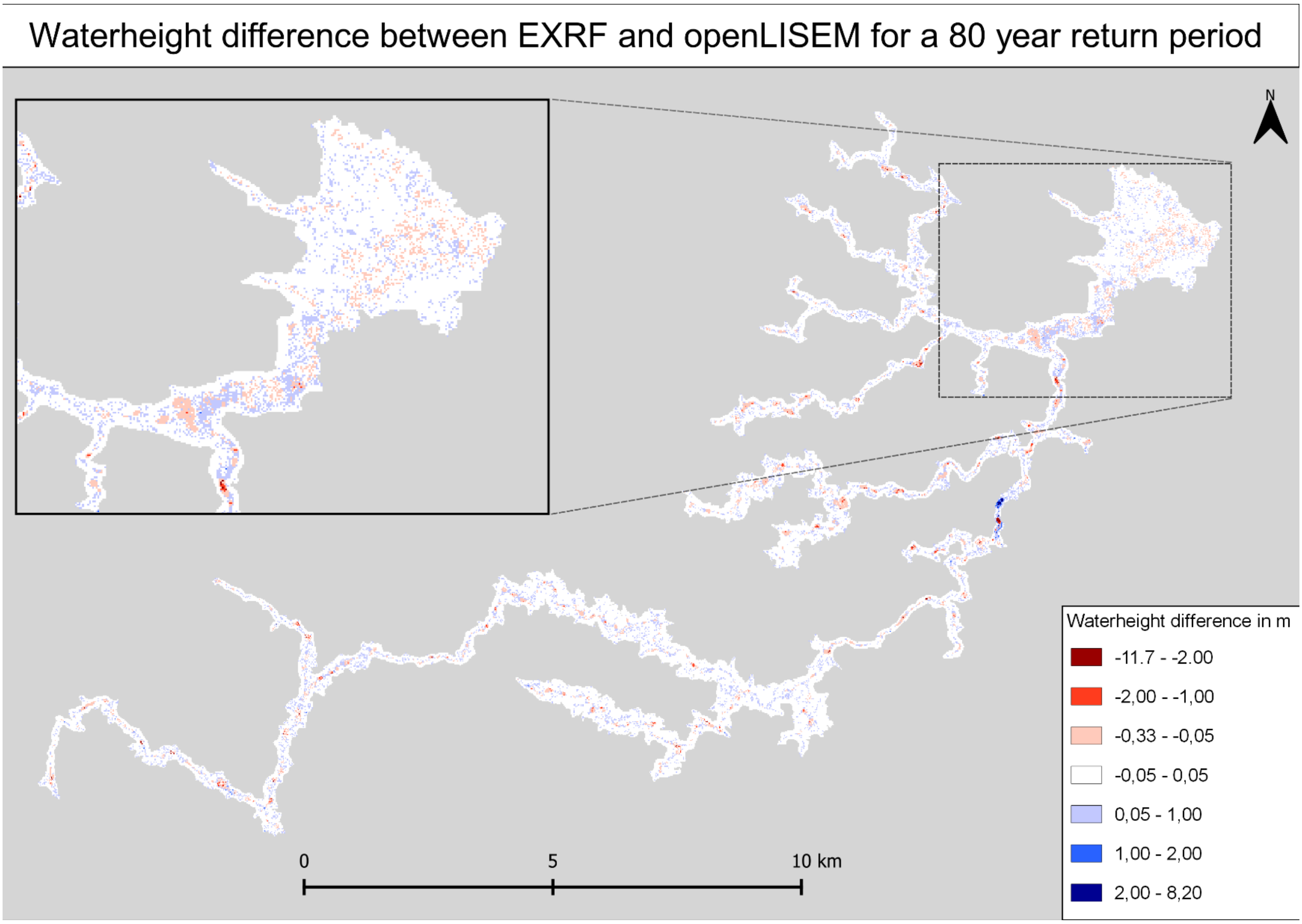


Figure 27: Difference in waterheight between the ML EXRF prediction and the openLISEM simulation for an 80 year return period event. With the ML model overestimating compared to openLISEM for the positive numbers (blue) and underestimating for the negative numbers (red)

7. Discussion

This research aimed to analyze the feasibility of using machine learning to predict the waterheight of flash floods in the context of early warning and early action in Northern Malawi. Additionally, the utilized tools and data needed to be easily accessible and usable.
This research proposed the conceptual method of using a hydrodynamic flood model to train an ML model to estimate maximum waterheight for flash floods and to do this by selecting predictors that are important in hydrodynamic flood modelling itself.

7.1 Sub-objective 1

To develop a realistic flash flood model of the study area in openLISEM to establish a ground truth on which the Machine Learning algorithms can be trained.

7.1.1 Gumbel analysis and synthetic rainfall event creation

The purpose of this analysis was to create synthetic rainfall events that could be used for the hydrodynamic modelling in openLISEM as the next step in the methodology outlined in this research. This purpose was achieved by creating seven different synthetic rainfall events based on 20 years of hourly bias-corrected CHIRPS data for the study area, therefore answering the sub-research question: **What are the different amounts of rainfall for different return period events for the study area?** The first approach of using historical events that caused significant flooding that could be found via news articles or red-cross disaster response trigger documents proved to be unachievable. The underestimation of events caused close to no flooding, which could prove problematic when training the ML model. There was too much uncertainty with the underestimating GSMaP data and the information available on the flooding the selected events caused, which will be highlighted in the next paragraph. This is why the decision was made to 1. Use CHIRPS data, and 2. Create synthetic rainfall events. Using synthetic events also allowed for more flexibility in the creation of additional events.

7.1.2 Hydrodynamic flood model

The purpose of the openLISEM hydrodynamic flood model was to serve as the first step in the methodology pipeline proposed in this research, to generate maximum waterheight maps for different return period events. The openLISEM flood model showed that the time between peak rainfall and peak discharge decreased for this study area with higher return period events. The maximum waterheight maps generated by openLISEM were then used to predict maximum waterheight using different machine learning algorithms.

While the sub-research question: '**what is the maximum waterheight and flood extent in the study area for different return period events?**' has been answered, this cannot be confidently said about the sub-objective. The output generated by openLISEM was what had been expected and provided what was needed for the next step. However, validating the maximum waterheight maps and the overall accuracy of the hydrodynamic flood model was not possible due to several factors.
First of all, hydrodynamic flood models are usually calibrated with fieldwork data. For example, river

widths can be measured, river bed types can be analyzed and a general view of the study area can be obtained. This was, however, not possible due to the COVID-19 epidemic. Moreover, local knowledge has been proven to provide precious insights into the operation of river networks and the history of such a river network (Bucherie, 2019). It has remained an untapped resource for this research.

Second of all, the amount of historical flood data availability is minimal and often not detailed enough to verify the accuracy of the flood model.

Consequently, there is the possibility that the maximum waterheight maps generated by the openLISEM hydrodynamic flood model are not entirely accurate. However, this does not completely invalidate the flood model, as it has been calibrated as best as possible and is developed with publicly available and reasonably accurate data.

7.2 Sub-objective 2

To develop a Machine Learning model for the study area using predictors created from a hydrodynamic modelling perspective to predict waterheight.

7.2.1 Machine learning model development

This objective aimed to create a working ML model and analyze its performance for different ML algorithms. Following this, the goal was to analyze the importance of the selection of predictors, which was made to represent the foundation on which hydrodynamic flood models operate.

Which combination of predictors is selected and what is their importance in predicting maximum waterheight?

The development of the model was done using a combination of 13 different predictors that were chosen based on their importance in the openLISEM hydrological model and created based on important aspects of a physically based and driven flood model that cannot be simulated with machine learning. Examples of the latter type of predictors are the ‘slopelength’ and ‘ups’ features. The eventual combination of features that were selected through recursive feature elimination was all of them. When the features: ‘roads’, ‘ksatcomp’ and ‘n’ are taken out of the dataset, the cross-validated score does not decrease significantly. So regarding optimal run time, these predictors could all be eliminated from the dataset. However, these eliminated features could have a significant effect on flooding in other study areas. The study area does not have any major dense vegetation, such as dense agricultural fields or forests. This can be seen in the land-use map (figure 10n) and the Leaf area index map (figure 10c). Study areas in other biomes could see the mannings’ N ranking higher in feature importance. The same goes for more urban areas with more hard roads, which, for instance, have significantly lower ksat values, affecting local infiltration speed and capacity (Schmidt et al., 2020). So considering the applicability to other study areas, these features were not removed from the dataset.

When looking at the feature importance between the different algorithms (figures 19 and 21), the RF and EXRF algorithms show a similar ranking with a low standard deviation for both pre and post hyperparameter tuning. In contrast, the XGB algorithm shows a different ranking, with most features having large standard deviations. This can be attributed to the correlation between features, as the

XGB algorithm will not (try to) focus on a feature once a link to the target variable has been established. This means that if there are two correlated features, all the importance will be on one of the features and not both. This leads to a less diluted importance scoring and possibly cause some features to have high standard deviations (Chen et al., 2018). It is important to mention that, while having a correlation between features is generally considered flawed in machine learning, decision trees and decision tree ensemble methods such as RF, EXRF and XGBoost are largely unaffected due to their random feature sampling at each split. However, as previously mentioned, it can dilute the importance of features when they are correlated, as their specific importance in predicting the target variable is similar. Thus both can be chosen when looking for the feature with the least amount of loss in a split.

When analyzing the feature importance in figure 21, all algorithms determined the DEM as the most important feature. This can likely be explained by the fact that, when looking at the openLISEM maximum waterheight maps, most flooding is generally and logically occurring in areas with a lower elevation. Additionally, it is evident that two of the most crucial parameters for hydrological modelling are represented in the top five most important features for all three algorithms: saturated hydraulic conductivity (Ksat) and the suction of the wetting front (PSI) of a soil. Both features are part of the foundation on which infiltration speed and capacity are based, which are essential in determining how much runoff is generated and thus how much water flows into the channel. Consequently, these features ranking high in importance might explain the relatively high accuracy in predicting the maximum flood volume for both events (see tables 12 and 14).
Seen from a hydrological modelling perspective and the way XGB deals with correlated features, the feature importance ranking of the XGB algorithm can be considered the most accurate one. With features such as PSI, Ksat, rainfall, mannings' N, gradient and land-use all having a similar importance and being able to switch positions due to their standard deviations, it signifies that the ML model is close to matching what the openLISEM model is doing when seen from the hydrological modelling parameter perspective.
Considering these features ranking high for the XGB algorithm, it is interesting that the only non-static feature, rainfall, does not rank very high in importance. Especially for the RF and EXRF algorithms. This can be explained, however. Because rainfall, just like in hydrological modelling, rainfall generally is not important for influencing where flood is going to occur, but rather how much.

What is the performance during the training of the different machine learning algorithms?
When developing the ML model showed that it can predict maximum waterheight with reasonable accuracy for all algorithms, with a mean validation R^2 of approximately 0.7 for RF, EXRF and XGB and MAE of approximately 10 cm depending on the algorithm. Overall, the XGB algorithm performed the best with the highest R2 of 0.72, an MAE of 0.103 and an RMSE of 0.078. When looking at the performance metrics before and after hyperparameter tuning, the XGB algorithm is displaying interesting behaviour once more, with a significant increase in prediction accuracy, which can be attributed to the sequential way the XGB algorithm builds its decision trees. This suggests that the default n_tree value is inadequate to predict maximum waterheight for flash floods accurately. Regarding the training and validation data, the XGB algorithm can be regarded as the most

favourable algorithm, specifically because of the possibly less diluted feature importance and because the mean calculation time is significantly lower than the other algorithms.
This shorter calculation time also allowed for more extensive hyperparameter tuning. As for the RF and EXRF algorithms, the cross-validated grid search was limited to n_tree and m_try tuning due to runtimes of over 12 hours to complete the grid search.

Subobjective 7.3

To generate a predicted flash flood maximum waterheight map for different return period events in the study area

7.3.1 Machine learning maximum waterheight prediction

What is the prediction accuracy?
When predicting a 20 and an 80 year return period event, the R^2 decreased to 0.59 and increased to 0.91 outside of the validation standard deviation, respectively, with an MAE of approximately 5 cm for both events. The RMSE varied slightly with 0.21 for the 20 year and 0.19 for the 80 year return period events. The ML model consistently overestimates the amount of flooded square meters by the same relative amount compared to the openLISEM simulations but seems to predict the total flood volume relatively accurately. For both events, the overall flood pattern is accurate compared to the openLISEM simulations, with major flooding occurring in the same locations.
One of the main differences, when looking at the over and underestimation patterns, is that for 80 year return period event, the model is significantly less inaccurate in predicting waterheight for the floodplain than for the 20 year return period event. While the maximum and minimum waterheights

The difference in R^2 between the 20 year return period and the 80 year return period is interesting, mainly because the differences lie far out of the standard deviation observed in the 10 fold cross-validation. Moreover, this difference in accuracy is not reflected in the MAE and RMSE, as they are quite similar for both events.
One of the possible causes could be attributed to the training dataset. When observing the maximum waterheight of the 10 year return period event in table 8 and figure 16b, relatively minor flooding occurs compared to the 50 year and 100 year maximum waterheight maps. Moreover, the 50 and 100 year events are similar in their flooding patterns. They have a significantly smaller relative difference in total flood volume between each other than between the 10 and 50 year events. The same applies to the total flood area. Considering this, there is likely considerable overlap in the flood extent patterns and flood heights of the 80 year predicted event and the 50 and 100 year trained events. This would make it easier to predict higher flood heights with higher return period events because of a higher representation of similar data in the dataset.
Additionally, the 80 year prediction is close between the lower 50 year and a higher 100 year return period events, which means it is rather clear what lower and upper boundaries there are in flood height prediction. This could explain why the model is primarily underestimating the 20 year event, as seen in figures 22 and 24, because the closest event and thus closest matching data is of an event with a lower return period than the event being predicted.

How does the ML prediction compare in flood extent?

The EXRF predictions for both events are accurate in predicting the flood patterns. With water pooling in areas where, according to the openLISEM flood models, it is supposed to. The ML model seems to overestimate the maximum flooded area by nearly the same amount for both events, with 9.65% for the 20 year and 9.05% for the 80 year (see tables 12 and 14). The main difference between the two events is in maximum flood volume. This suggests that most of the difference in the accuracy of the 20 year return period event, compared to the 80 year return period event, expresses itself in the flood volume. This is confirmed by the residual plots, where for the 20 year event predictions are moderately underestimated (See figure 22), and the 80 year event only slightly overestimated (See figure 25).

This is shown in the difference maps as well. In figure 24, there is significant underestimation occurring in the floodplain. This is likely due to, the in the previous section mentioned reason, that the 10 year return period event the closest event is to the 20 year. When looking at figure 16b and 23a, there is a significant difference between flood extent. This also suggests that for the openLISEM flood model, a threshold is exceeded between the 10 and 20 year return period events, causing much more flooding. Evidently, this underestimation of the floodplain is not happening (figure 27) for the 50, 100 and 80 year return period events because there is no significant difference. However, despite this apparent threshold, the ML model still managed to predict the 20 year return period flood extent with reasonable accuracy.

Therefore, the overestimation of the area upstream for the 20 year event seems curious, as the 10 year return period event shows no flooding in that region (figure 24). A combination of training data and feature importances could explain this, as this area is signified by high gradients (figure 10j). More importantly, both openLISEM simulations that did have flooding in this area (50 and 100 year) show very high levels of flooding, leading the ML model to overestimate.

A solution to this accuracy difference problem could be to train, validate and test only on a single event. This would mean that the ML model would be split into, for example, 80% training, 10% validating and 10% testing to prevent data leakage. While this could improve accuracy and, more importantly, stability of performance, this was not possible for the objectives of this research. When only 10% of randomly sampled pixels can be used for testing/predicting, it is impossible to accurately compare the prediction to the openLISEM flood model.

Additionally, because openLISEM simulations are needed to train the algorithm, the benefit, in an early warning application context, of the faster ML model compared to the openLISEM flood model is nullified when a simulation needs to be run for every different forecasted return period event to get an accurate result.

Considering this, a different combination of events could be used to train the algorithm. However, training on events similar to the event you are trying to predict could impact the flexibility of application when a forecasted rainfall event falls outside of the minimum or maximum return period events it has been trained on. While the XGB algorithm is capable of extrapolation, the RF and EXRF algorithms are not. This extrapolation ability has not been tested for this research.

Additional measures that could improve the accuracy is data scaling, a method used to improve ML model accuracy by decreasing the absolute difference of values between different variables. For this research, there are significant absolute differences in values between features (see table 5). However, tree-based algorithms, such as those used for this research, are unaffected by significant data scale differences due to how they operate (Brownlee, 2019; Praveen, 2020; Bhushan; 2020). Additionally, most data scaling methods normalize the data between 0 and 1, which would produce useless maximum waterheight maps.

Lastly, different training dataset creation methods could be used. For this research, the decision was made to create a buffer that included the full flood extent for the maximum (100 year) return period event to include the same pixels for every event in the training dataset. There are other options, however. For instance, only selecting pixels with a waterheight equal to or bigger than 0.05m. This would reduce the row count of the dataset and removes most of the as non-flooded classified pixels. This would also improve the class imbalance of the target variable, as most pixels within the buffer are non-flooded pixels (<0.05 m) and could therefore improve accuracy. However, it has not been proven that non-flooded pixels are useless in accurately predicting flood patterns. They might present an overrepresentation of high return period event pixel count and thus maximum waterheight values compared to low return period events. Limited experimentation with primarily buffer size was done early on in the ML development, which can be found in appendix F. These results have not been cross-validated.

7.4 Proof-of-concept

The concept of using a hydrodynamic flood model to train an ML model to predict flash flood susceptibility is in its infancy. As described in the introduction, most flood susceptibility analysis literature uses a historical database, which means no return period events, predicting by classification and possibilities in variation in accuracy with limited validation options available. This limits comparable literature severely.

A recent paper by Mahato et al. (2021) dives into the use of the HEC-RAS hydrological flood model for validation. However, they use this to predict flood susceptibility and not waterheight. Additionally, the hydrodynamic flood model was only being used for validation, not training. The training was done on a database of historical flood events and predictors were selected based on existing literature. Using the calibrated 2D HEC-RAS flood model for validation, they confirmed that the correlation between the flood model and the ML susceptibility model for flooded area varied between 0.62 and 0.69 for different algorithms. In this regard, it is comparable to the R^2 of 0.68 for the training dataset. However, it is hard to compare with fundamental differences such as classification versus regression and maximum waterheight versus susceptibility.

Hosseiny et al.'s (2020) research did use a hydrodynamic flood model to create the training set for the ML model. The ML approach was more complex, however, using RF classification to identify flooded and non-flooded pixels and then an artificial neural network (ANN) regression to predict river depth together with elevation data. This was done for different discharges, suggesting a return period event-based approach. The flooded and non-flooded pixels were predicted with an accuracy of 0,985. After this, the ANN predicted flood depth with an accuracy of 0.88 and an RMSE of 0.02m. Additionally, the error in predicted depth increased with higher discharges. It was trained on the following discharges: 10, 50, 95, 120, 150, 300 and 400 m^3/s. It predicted on the following discharges 20, 30, 45, 225, and 350 m^3/s. While Hosseiny et al. (2020) do not clarify the higher inaccuracy for higher discharges, the cause could be the same as the one proposed in the chapter above. It could be easier to predict with more similar data for lower discharges (for this research, the higher return period events).

7.5 Strengths and limitations

Most importantly, being able to predict water height using ML can provide a significant increase of information that can be used for impact analyses in a short time frame. This, in comparison with previous ML flash flood susceptibility analyses, only providing susceptibility of an area to flooding. Additionally, using a hydrological flood model provides a significant amount of precise data that the ML model can be trained, which can be helpful in areas with poor or no flood extent data availability. With this detailed training data, the ML model has proven that it is capable of predicting waterheight with reasonable accuracy for different return period events and doing this fast. With a minimum training time of approximately one minute and a prediction time of approximately 10 seconds, it has proven to be significantly faster than a hydrological flood model.

Furthermore, it is flexible. With the tree-based ensemble methods being able to handle all the selected predictors, it shows it is easy to adapt, work with and does not need extensive computing power.

However, with the current state of the method proposed and the knowledge gained in this research, it has not been proven yet to function without a hydrological flood model providing training data. This means that to utilize the fast calculation times of the ML model in an early warning scenario, flood model simulations should have been run in advance with different return period events for a specific study area. In this case, the ML model is there to give an accurate flood extent map of a forecasted event that has not been simulated with a hydrological flood model. Moreover, developing a hydrological flood model for a study area can be a complex task that can cost a significant amount of time. Additionally, depending on the size of the study area and the resolution of the data, hydrological modelling simulations can take a long time to complete and require significant computational power.

8. Conclusion

This research has looked at the feasibility of using machine learning to predict the maximum waterheight of flash floods in the context of early warning and early action in Northern Malawi.

The important first step was to develop a hydrodynamic flood model for the Kyungu river catchment and this objective has been fulfilled. The reason for using a hydrodynamic flood model for the proposed method as a foundation stems from the common problem of not having a sufficient flood event database in place. The importance of having a good database has been proven for this research, as it was impossible to calibrate the model to real-world data. Having calibration data was not crucial for this research, however, as the primary goal of this research was to compare ML maximum waterheight predictions to that of a hydrodynamic flood model.

To make the flood model work, an annual maxima (GUMBEL analysis) on bias-corrected rainfall data has been done, resulting in the creation of seven IDF curves and seven synthetic return period events.

The results of training the ML model showed that the different ensemble algorithms perform similarly. Additionally, the feature importance of the ML model has shown that fundamental hydrological modelling parameters that have been used as predictors are important for predicting maximum waterheight.

The waterheight predictions on a 20 year return period and an 80 year return period have shown different performance compared to the training, with lower accuracy for the 20 year return period and higher accuracy for the 80 year return period events. Both event predictions show similar trends in maximum flooded area overestimation, with the primary accuracy difference between the two events being in maximum flood volume. This suggests dependability on the similarity of training events to the predicted event to get a high accuracy prediction. However, both ML models can sufficiently predict flood patterns similar to those simulated by the openLISEM flood model for both events.

To answer the main research question:

What is the feasibility of using machine learning to predict the maximum waterheight of flash floods in the context of early warning and early action in Northern Malawi?

This research has proven that with the selection of predictors, it is feasible to predict maximum waterheight with ML. However, the proposed method of using hydrodynamic flood modelling output to train the ML model and the possible dependability on the similarity of training events to get a very accurate prediction influences the method's applicability in an early warning context. There are many things on the method proposed in this research, that can still be researched and improved upon in the future. This could result in higher accuracy and/or a more stable performance, which can make it more feasible for use in an early warning context.

9. Recommendation

Most importantly, future research should focus on the applicability of this method to other study areas. Whether the ML model with the, in this research proposed combination of predictors, can reproduce similar results in other study areas with different geological and climatological features. Additionally, it can be valuable to gain more knowledge on and, primarily, confirm the link between the variance in the ML maximum water height prediction accuracy and the specific return period events the ML model has been trained on.

Furthermore, future research should consider whether this method needs to be trained on hydrological modelling simulations run on the specific study area or not and if different hydrological modelling software packages affect the ML maximum waterheight prediction accuracy.

It is also relevant to research the impact of the buffer size used to generate the training dataset. Possible improvements could be made here in regards to calculation time and prediction accuracy.

Lastly, while waterheight is important in estimating the flood extent and potential impact, the flow velocity is equally important, especially for flash floods. With often high flow velocities being one of the more dangerous characteristics of flash floods, it would be useful to see whether this can be predicted with this method as well.

10. References

510 Global. (2018). *Data Responsibility Policy*. Retrieved from https://www.510.global/wp-content/uploads/2018/12/510-Data-Responsibility-policy-V2.2-20181211-PUBLIC-USE.pdf

510 GLOBAL. (n.d.). *Humanitarian Aid Through Data & Digital Products*. Retrieved October 14, 2020, from

Aleotti, P., & Chowdhury, R. (1999). *Landslide hazard assessment: summary review and new perspectives.* Bulletin of Engineering Geology and the Environment 1999 58:1, 58(1), 21–44. https://doi.org/10.1007/S100640050066

Arcement, G. J., & Schneider, V. R. (n.d.). *Guide for Selecting Manning's Roughness Coefficients for Natural Channels and Flood Plains*.

Baartman, J., Jetten, V. G., Ritsema, C. J., & De Vente, J. (2012). *Exploring effects of rainfall intensity and duration on soil erosion at the catchment scale using OpenLISEM.* EGUGA, 14, 651. Retrieved from https://ui.adsabs.harvard.edu/abs/2012EGUGA..14..651B/abstract

Bastiaanssen, W. G. M., Huygen, J., Schakel, J. K., & Broek, B. J. van den. (1996). *Modelling the soil-water-crop-atmosphere system to improve agricultural water management in arid zones (SWATRE)* (pp. 13–30). Retrieved from https://research.wur.nl/en/publications/modelling-the-soil-water-crop-atmosphere-system-to-improve-agricu

Batjes, N. H., Ribeiro, E., & Van Oostrum, A. (2020). *Standardised soil profile data to support global mapping and modelling (WoSIS snapshot 2019).* Earth System Science Data, 12(1), 299–320. https://doi.org/10.5194/ESSD-12-299-2020

Borga, M., Gaume, E., Creutin, J. D., & Marchi, L. (2008). *Surveying flash floods: gauging the ungauged extremes Re-Analysis of Radar-Based Precipitation Estimates of the Flash-Flood Producing Storm.* Process. https://doi.org/10.1002/hyp.7111

Borga, M., Stoffel, M., Marchi, L., Marra, F., & Jakob, M. (2014). *Hydrogeomorphic response to extreme rainfall in headwater systems: Flash floods and debris flows*. Journal of Hydrology, 518(PB), 194–205. https://doi.org/10.1016/j.jhydrol.2014.05.022

Bougadis, J., & Adamowski, K. (2006). *Scaling model of a rainfall intensity-duration-frequency relationship*. Process, 20, 3747–3757. https://doi.org/10.1002/hyp.6386

Bout, B. van den. (2020). *INTEGRATED PHYSICALLY-BASED MULTI-HAZARD MODELLING*.

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

Broeken, M. (2018). *Exploring vulnerability and impact of floods in Malawi*. Retrieved from https://run.unl.pt/handle/10362/48099

Brownlee, J. (2016). *A Gentle Introduction to XGBoost for Applied Machine Learning*. Retrieved June 10, 2021, from https://machinelearningmastery.com/gentle-introduction-xgboost-applied-machine-learning/

Brownlee, J. (2019). *How to use Data Scaling Improve Deep Learning Model Stability and Performance*. Retrieved July 30, 2021, from https://machinelearningmastery.com/how-to-improve-neural-network-stability-and-modeling-performance-with-data-scaling/

Brownlee, J. (2020). *A Gentle Introduction to the Gradient Boosting Algorithm for Machine Learning*. Retrieved June 12, 2021, from https://machinelearningmastery.com/gentle-introduction-gradient-boosting-algorithm-machine-learning/

Bucherie, A. (2019). *On the predictability of flash floods and their impacts in North Malawi*. Delft : IHE Delft Institute for Water Education.

Bucherie, A., Werner, M., van den Homberg, M., & Tembo, S. (2021). NHESSD - *Flash Flood warning in context: combining local knowledge and large-scale hydro-meteorological patterns*. Retrieved July 13, 2021, from https://nhess.copernicus.org/preprints/nhess-2021-107/

Buchhorn, M. ; Smets, B. ; Bertels, L. ; De Roo, B. ; Lesiv, M. ; Tsendbazar, N. - E. ; Herold, M. ; Fritz, S. *Copernicus Global Land Service: Land Cover 100m: collection 3: epoch 2019*: Globe 2020.

Central Intelligence Agency. (2020). *The World Factbook Africa: Malawi.* Retrieved October 6, 2020, from https://www.cia.gov/library/publications/the-world-factbook/geos/mi.html

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

Chen, T., & Guestrin, C. (2016). XGBoost: *A Scalable Tree Boosting System*. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). New York, NY, USA: ACM. https://doi.org/10.1145/2939672.2939785

Chen, T., Benesty, M., & He, T. (2018). *Understand your dataset with XGBoost — xgboost 1.5.0-dev documentation*. Retrieved July 29, 2021, from https://xgboost.readthedocs.io/en/latest/R-package/discoverYourData.html

Chow, V. T. (1959*). Manning's n for Channels*. Retrieved March 2, 2021, from http://www.fsl.orst.edu/geowater/FX3/help/8_Hydraulic_Reference/Mannings_n_Tables.htm

Chow, V.T., Maidment, D.R. and Mays, L.W. (1988). *Applied Hydrology*. McGraw-Hill, New York, 572

Climate Hazards Center. (n.d.). CHIRPS: *Rainfall Estimates from Rain Gauge and Satellite Observations*. Climate Hazards Center - UC Santa Barbara. Retrieved March 29, 2021, from https://www.chc.ucsb.edu/data/chirps

Cloke, H. L., & Pappenberger, F. (2009, September 15). *Ensemble flood forecasting: A review*. Journal of Hydrology, Vol. 375, pp. 613–626. https://doi.org/10.1016/j.jhydrol.2009.06.005

Corominas, J., Copons, R., Vilaplana, J. M., Altimir, J., & Amigó, J. (2003). *Integrated Landslide Susceptibility Analysis and Hazard Assessment in the Principality of Andorra*. Natural Hazards 2003 30:3, 30(3), 421–435. https://doi.org/10.1023/B:NHAZ.0000007094.74878.D3

De Roo, A. P. J., Wesseling, C. G., Cremers, N. H. D. T., Offermans, R. J. E., Ritsema, C. J., & Van Oostindie, K. (1994). *LISEM: a new physical-based hydrological and soil erosion model in a GIS-*

environment, theory and implementation DESIRE: Desertification mitigation and remediation of land View project Horizon 2020: Interactive Soil Quality Assessment in Europe and China f. Retrieved from https://www.researchgate.net/publication/279555181

De Sousa, L. M., Poggio, L., Dawes, G., Kempen, B., & Van Den Bosch, R. (2020). Computational Infrastructure of SoilGrids 2.0. In *International Symposium on Environmental Software Systems (ISESS 2020): Environmental Software Systems. Data Science in Action* (pp. 24-31). (IFIP Advances in Information and Communication Technology ; Vol. 554). Springer. https://doi.org/10.1007/978-3-030-39815-6_3

Dodma. (2019). *Malawi 2019 floods post disaster needs assessment report*. 1–111.

Douben, K.-J. (2006). *Characteristics of river floods and flooding: a global overview, 1985–2003.* Irrigation and Drainage, *55*(S1), 9–21. https://doi.org/10.1002/ird.239

Duka, M. A., Lasco, J. D. D., Veyra, C. D., & Aralar, A. B. (2018). *Comparative assessment of different methods in generating design storm hyetographs for the Philippines.* Journal of Environmental Science and Management, 21(1), 82–89.

EM-DAT. (n.d.). *Classification*. Retrieved August 30, 2020, from https://www.emdat.be/classification

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. & Duchesnay, É. (2011). *Scikit-learn: Machine Learning in Python*. In Journal of Machine Learning Research (Vol. 12). Retrieved from http://scikit-learn.sourceforge.net.

Fadhel, S., Rico-Ramirez, M. A., & Han, D. (2017). *Uncertainty of Intensity–Duration–Frequency (IDF) curves due to varied climate baseline periods.* Journal of Hydrology, 547, 600–612. https://doi.org/10.1016/j.jhydrol.2017.02.013

Floodlist news. (2015). *2015 Floods Leave Malawi Facing Worst Food Crisis in 10 Years.* Retrieved July 31, 2021, from https://floodlist.com/africa/floods-malawi-facing-worst-food-crisis-10-years

Funk, C., Peterson, P., Landsfeld, M., Pedreros, D., Verdin, J., Shukla, S., Husak, G., Rowland, J., Harrison, L., Hoell, A. and Michaelsen, J. (2015*). The climate hazards infrared precipitation with stations—a new environmental record for monitoring extremes*. Scientific Data 2015 2:1, *2*(1), 1–21. https://doi.org/10.1038/sdata.2015.66

Gaume, E., Bain, V., Bernardara, P., Newinger, O., Barbuc, M., Bateman, A., Blaškovičová, L., Blöschl, G., Borga, M., Dumitrescu, A., Daliakopoulos, I., Garcia, J., Irimescu, A., Kohnova, S., Koutroulis, A., Marchi, L., Matreata, S., Medina, V., Viglione, A., Preciso, E., Sempere-Torres, D., Stancalie, G., Szolgay, Jan., Tsanis, I., Velasco, D. & Viglione, A.(2009). *A compilation of data on European flash floods.* Journal of Hydrology, 367(1–2), 70–78. https://doi.org/10.1016/j.jhydrol.2008.12.028

Geurts, P., Ernst, D., & Wehenkel, L. (2006). *Extremely randomized trees*. Mach Learn (). https://doi.org/10.1007/s10994-006-6226-1

Guha-Sapir, D., Vos, F., Below, R., & Ponserre, S. (2011). *Annual Disaster Statistical Review 2011 The numbers and trends*.

Heber Green, W., & Ampt, G. A. (1911). *Studies on soil physics*. The Journal of Agricultural Science, 4(1), 1–24. https://doi.org/10.1017/S0021859600001441

Hessel, R., Jetten, V., & Guanghui, Z. (2003). *FLOW VELOCITY*

Hoeppe, P. (2016). *Trends in weather related disasters - Consequences for insurers and society*. Weather and Climate Extremes, 11, 70–79. https://doi.org/10.1016/j.wace.2015.10.002\

Hosseiny, H., Nazari, F., Smith, V., & Nataraj, C. (2020). *A Framework for Modeling Flood Depth Using a Hybrid of Hydraulics and Machine Learning.* Scientific Reports 2020 10:1, 10(1), 1–14. https://doi.org/10.1038/s41598-020-65232-5

Hu, W., She, D., Shao, M., Chun, K. P., & Si, B. (2015). *Effects of initial soil water content and saturated hydraulic conductivity variability on small watershed runoff simulation using LISEM. Https://Doi.Org/10.1080/02626667.2014.903332*, *60*(6), 1137–1154.

IFRC. (2018). *Leaving Millions Behind*. Retrieved from https://media.ifrc.org/ifrc/world-disaster-report-2018/

IFRC. (n.d.). *Types of disasters.* Retrieved October 18, 2020, from https://www.ifrc.org/en/what-we-do/disaster-management/about-disasters/definition-of-hazard/

International Displacement Monitoring Centre. (2018). *Internal Displacement 2018*.

International Displacement Monitoring Centre. (2020). *Internal Displacement 2020*.

ITC faculty of geo-information. (n.d.). *About – openlisem.* Retrieved October 19, 2020, from https://blog.utwente.nl/lisem/about/

ITC faculty of geo-information. (n.d.*). Running lisem – openlisem*. Retrieved october 19, 2020, from https://blog.utwente.nl/lisem/running-lisem/

Jetten, V. (2016). *LISEM INTEGRATED WATERSHED MODEL THEORY AND PRINCIPLES*.

Jetten, V. G. (2018). *OpenLISEM Multi-Hazard Land Surface Process Model Documentation & User Manual*.

Jetten, V. G., & Roo, A. P. J. de. (2001). *Spatial Analysis of Erosion Conservation Measures with LISEM.* Landscape Erosion and Evolution Modeling, 429–445. https://doi.org/10.1007/978-1-4615-0575-4_14

Jonkman, S. N. (2005). *Loss of life due to floods.* Journal of Flood Risk Management, 1(1), 43–56. https://doi.org/10.1111/j.1753-318x.2008.00006.x

Kapkar, B. (2020). *Which Machine Learning requires Feature Scaling(Standardization and Normalization)? And Which not?* Retrieved July 30, 2021, from https://www.kaggle.com/getting-started/159643

Khosravi, K., Pham, B. T., Chapi, K., Shirzadi, A., Shahabi, H., Revhaug, I., Prakash, I. & Tien Bui, D. (2018). A *comparative assessment of decision trees algorithms for flash flood susceptibility modelling* *at Haraz watershed, northern Iran*. Science of the Total Environment, 627, 744–755. https://doi.org/10.1016/j.scitotenv.2018.01.266

Khosravi, K., Shahabi, H., Pham, B. T., Adamowski, J., Shirzadi, A., Pradhan, B., Dou, J., Ly., H. B., Gróf, G., Ho, H. L., Hong, H., Chapi, K. & rakash, I. (2019). *A comparative assessment of flood susceptibility modelling using Multi-Criteria Decision-Making Analysis and Machine Learning Methods*. Journal of Hydrology, 573, 311–323. https://doi.org/10.1016/j.jhydrol.2019.03.073

Kubota, T., Aonashi, K., Ushio, T., Shige, S., Takayabu, Y. N., Kachi, M., Arai, Y., Tashima, T., Masaki, T., Kawamoto, N., Mega, T., Yamamoto, M. K., Hamada, A., Yamaji, M., Liu, G., & Oki, R. (2020). Global Satellite Mapping of Precipitation (GSMaP) Products in the GPM Era. *Advances in Global Change Research*, *67*, 355–373. https://doi.org/10.1007/978-3-030-24568-9_20

Lawson, E., Smith, D., Sofge, D., Elmore, P., & Petry, F. (2017). Decision forests for machine learning classification of large, noisy seafloor feature sets. *Computers and Geosciences*, *99*, 116–124. https://doi.org/10.1016/j.cageo.2016.10.013

Lin, L., Wu, Z., & Liang, Q. (2019). Urban flood susceptibility analysis using a GIS-based multi-criteria analysis framework. *Natural Hazards 2019 97:2*, *97*(2), 455–475. https://doi.org/10.1007/S11069-019-03615-2

Linsley Jr, R. K., Kohler, M. A., & Paulhus, J. LH., 1982. *Hydrology for Engineers*. McGraw--Hill Series in Water Resources and Environmental Engineering (pp. 508): McGraw--Hill, Inc

Liu, Z., Merwade, V., & Jafarzadegan, K. (2019). Investigating the role of model structure and surface roughness in generating flood inundation extents using one- and two-dimensional hydraulic models. *Journal of Flood Risk Management*, *12*(1). https://doi.org/10.1111/jfr3.12347

Madruga De Brito, M., & Evers, M. (2016). Multi-criteria decision-making for flood risk management: a survey of the current state of the art. *Hazards Earth Syst. Sci*, *16*, 1019–1033. https://doi.org/10.5194/nhess-16-1019-2016

Mahato, S., Pal, S., Talukdar, S., Saha, T. K., & Mandal, P. (2021). Field based index of flood vulnerability (IFV): A new validation technique for flood susceptible models. *Geoscience Frontiers*, *12*(5), 101175. https://doi.org/10.1016/j.gsf.2021.101175

Malek, K. (2020)*. biascorrection: Compares model simulated against observations and bias corrects model simulations to remove systematic biases*. R package version 0.1.0.

Manda, M., Kamlomo, D., Mphande, C., Wanda, E., Msiska, O., Kaunda, J., Kushe, J. (2016). Karonga town: Growth and Risk Profile. Retrieved October 6, 2020, from Urban Africa Risk Knowledge

Marchi, L., Borga, M., Preciso, E., & Gaume, E. (2010). Characterisation of selected extreme flash floods in Europe and implications for flood risk management. *Journal of Hydrology*, *394*(1–2), 118–133. https://doi.org/10.1016/j.jhydrol.2010.07.017

Morde, V. (2019). XGBoost *Algorithm: Long May She Reign!* Retrieved June 10, 2021, from https://towardsdatascience.com/https-medium-com-vishalmorde-xgboost-algorithm-long-she-may-rein-edd9f99be63d

Muñoz, P., Orellana-Alvear, J., Willems, P., & Célleri, R. (2018). Flash-Flood Forecasting in an Andean Mountain Catchment—Development of a Step-Wise Methodology Based on the Random Forest Algorithm. *Water*, 10(11), 1519. https://doi.org/10.3390/w10111519

Murray-Hudson, M., Wolski, P., & Ringrose, S. (2006). Scenarios of the impact of local and upstream changes in climate and water use on hydro-ecology in the Okavango Delta, Botswana. *Journal of Hydrology*, *331*(1–2), 73–84. https://doi.org/10.1016/J.JHYDROL.2006.04.041

Murray-Hudson, M., Wolski, P., & Ringrose, S. (2006). Scenarios of the impact of local and upstream changes in climate and water use on hydro-ecology in the Okavango Delta, Botswana. *Journal of Hydrology*, *331*(1–2), 73–84. https://doi.org/10.1016/J.JHYDROL.2006.04.041

National Resilience Taskforce. (2018). *National disaster risk reduction framework consultation draft*. (October), 1–24. Retrieved from https://www.homeaffairs.gov.au/emergency/files/national-disaster-risk-reduction-framework.pdf

National Statistical Office. (2019). *Malawi in Figures 2019 Edition*. Retrieved from http://www.nsomalawi.mw/images/stories/data_on_line/general/malawi_in_figures/2019 Malawi in Figures.pdf

Pappenberger, F., Matgen, P., Beven, K. J., Henry, J. B., Pfister, L., & Fraipont, P. (2006). Influence of uncertain boundary conditions and model structure on flood inundation predictions. *Advances in Water Resources*, *29*(10), 1430–1449. https://doi.org/10.1016/J.ADVWATRES.2005.11.012

Pham, B. T., Khosravi, K., & Prakash, I. (2017). Application and Comparison of Decision Tree-Based Machine Learning Methods in Landside Susceptibility Assessment at Pauri Garhwal Area, Uttarakhand, India. *Environmental Processes*, *4*(3), 711–730. https://doi.org/10.1007/s40710-017-0248-5

Pham, B. T., Tien Bui, D., Dholakia, M. B., Prakash, I., & Pham, H. V. (2016). A Comparative Study of Least Square Support Vector Machines and Multiclass Alternating Decision Trees for Spatial Prediction of Rainfall-Induced Landslides in a Tropical Cyclones Area. *Geotechnical and Geological Engineering 2016 34:6*, *34*(6), 1807–1824. https://doi.org/10.1007/S10706-016-9990-0

Pham, B. T., Tien Bui, D., Prakash, I., & Dholakia, M. B. (2017). Hybrid integration of Multilayer Perceptron Neural Networks and machine learning ensembles for landslide susceptibility assessment at Himalayan area (India) using GIS. *CATENA*, *149*, 52–63. https://doi.org/10.1016/J.CATENA.2016.09.007

Pourghasemi, H. R., Kariminejad, N., Amiri, M., Edalat, M., Zarafshar, M., Blaschke, T., & Cerda, A. (2020). Assessing and mapping multi-hazard risk susceptibility using a machine learning technique. *Scientific Reports*, *10*(1), 1–11. https://doi.org/10.1038/s41598-020-60191-3

Rudari, R., Beckers, J., De Angeli, S., Rossi, L., & Trasforini, E. (2016). Impact of modelling scale on probabilistic flood risk assessment: The Malawi case. *E3S Web of Conferences*, *7*(January). https://doi.org/10.1051/e3sconf/20160704015

Ruiter, M. C., Couasnon, A., Homberg, M. J. C., Daniell, J. E., Gill, J. C., & Ward, P. J. (2020). Why We Can No Longer Ignore Consecutive Disasters. *Earth's Future*, *8*(3). https://doi.org/10.1029/2019EF001425

Samanta, S., Pal, D. K., & Palsamanta, B. (2018). Flood susceptibility analysis through remote sensing, GIS and frequency ratio model. *Applied Water Science*, *8*(2), 66. https://doi.org/10.1007/s13201-018-0710-1

Saxton, K. E., & Rawls, W. J. (2006). Soil water characteristic estimates by texture and organic matter for hydrologic solutions. *Soil Science Society of America Journal*, 70(5), 1569-1578.

Saxton, K. E., & Rawls, W. J. (2006). Soil Water Characteristic Estimates by Texture and Organic Matter for Hydrologic Solutions. *Soil Science Society of America Journal*, *70*(5), 1569–1578. https://doi.org/10.2136/SSSAJ2005.0117

Schmidt, L., Heße, F., Attinger, S., & Kumar, R. (2020). Challenges in Applying Machine Learning Models for Hydrological Inference: A Case Study for Flooding Events Across Germany. *Water Resources Research*, *56*(5), e2019WR025924. https://doi.org/10.1029/2019WR025924

Shafizadeh-Moghadam, H., Valavi, R., Shahabi, H., Chapi, K., & Shirzadi, A. (2018). Novel forecasting approaches using combination of machine learning and statistical models for flood susceptibility mapping. *Journal of Environmental Management*, *217*, 1–11. https://doi.org/10.1016/j.jenvman.2018.03.089

Shrestha, M. S., & Takara, K. (2008). Impacts of floods in South Asia. *Journal of South Asia Disaster Study*, *1*(1), 85-106.

Simpson, A., Phillips, E., Balog, S., Fraser, S., Jongman, B., Ledden, M. van, Murnane, R and Himmelfarb, A. (2019). *Disaster Risk Profile Malawi*.

Singh, H. (2018). *Understanding Gradient Boosting Machines*. Retrieved June 12, 2021, from https://towardsdatascience.com/understanding-gradient-boosting-machines-9be756fe76ab

Smith, R. E., & Parlange, J. Y. (1978). A parameter-efficient hydrologic infiltration model. *Water Resources Research*, 14(3), 533-538.

Tehrany, M. S., Pradhan, B., & Jebur, M. N. (2015). Flood susceptibility analysis and its verification using a novel ensemble support vector machine and frequency ratio method. *Stochastic Environmental Research and Risk Assessment 2015 29:4*, *29*(4), 1149–1165. https://doi.org/10.1007/S00477-015-1021-9

Teule, T. (2019). *Assessing Two Methods to Potentially Improve the Flood Early Warning System in the Lower Shire Valley in Malawi*.

The International Federation of Red Cross and Red Crescent Societies. (2008). *Early warning Early action*. Retrieved from https://media.ifrc.org/ifrc/wp-content/uploads/sites/5/2017/11/Early-warning-early-action.pdf

The Kerneltrip. (2018). Random forest vs extra trees. Retrieved June 10, 2021, from https://www.thekerneltrip.com/statistics/random-forest-vs-extra-tree/

Thenraj, P. (2020). Do Decision Trees need Feature Scaling? Retrieved July 30, 2021, from https://towardsdatascience.com/do-decision-trees-need-feature-scaling-97809eaa60c6

UNDRR. (n.d.). Hazard. Retrieved October 18, 2020, from https://www.undrr.org/terminology/hazard

United Nations Office for the Coordination of Humanitarian Affairs (2013). Disaster response in Asia and the Pacific: A guide to international tools and services. Bangkok. UN OCHA 2013b, "Report: The after Action Review/Lessons Learned Workshops Typhoon Bopha Response", Manila

Uzoegbo, H. C. (2019). Dry-stack and compressed stabilized earth-block construction. In *Nonconventional and Vernacular Construction Materials: Characterisation, Properties and Applications* (pp. 305–350). https://doi.org/10.1016/B978-0-08-102704-2.00012-3

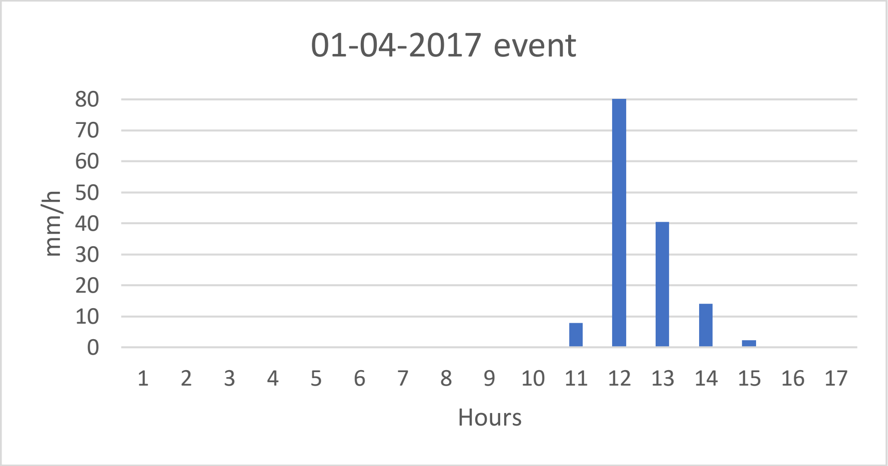
van Milligen de Wit, M. (2021). *Urban surface water flooding in Mzuzu*.

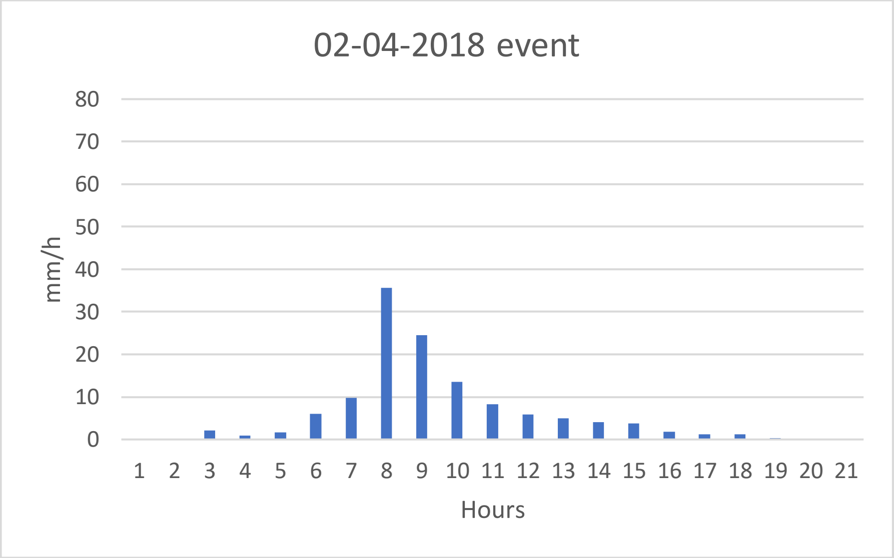
Varnes, D. J.: 1984, Landslide hazard zonation: a review of principles and practice, Natural Hazards, 3. UNESCO. Paris. 63 pp

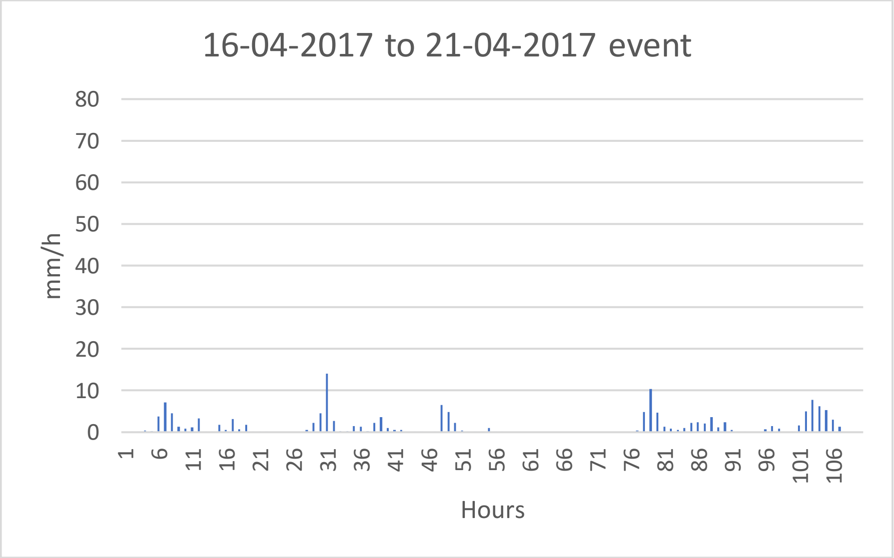
Zurita-Milla, R. (2020). *Classification and regression via decision trees and random forests*.

10. Appendices

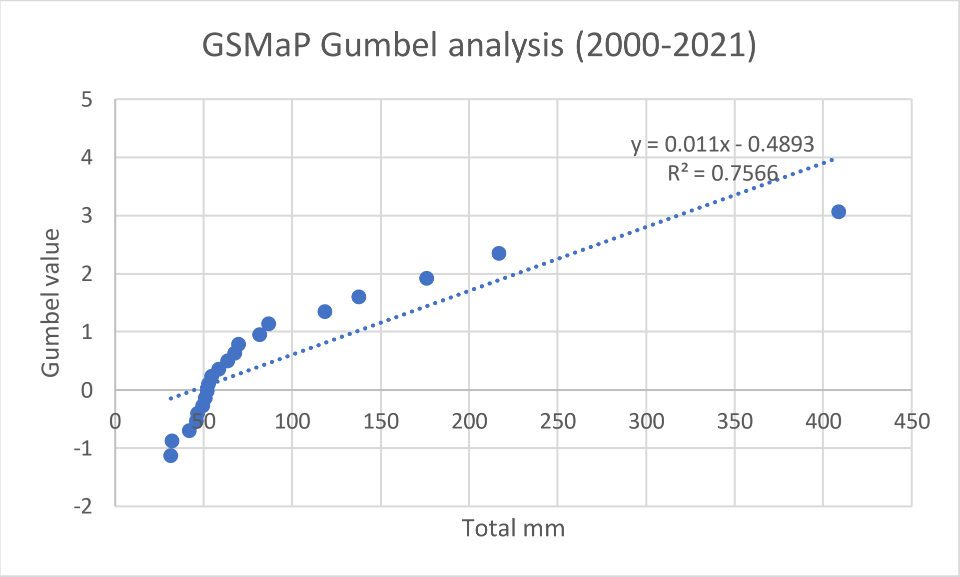
Appendix A: initial event rainfall event selection candidates





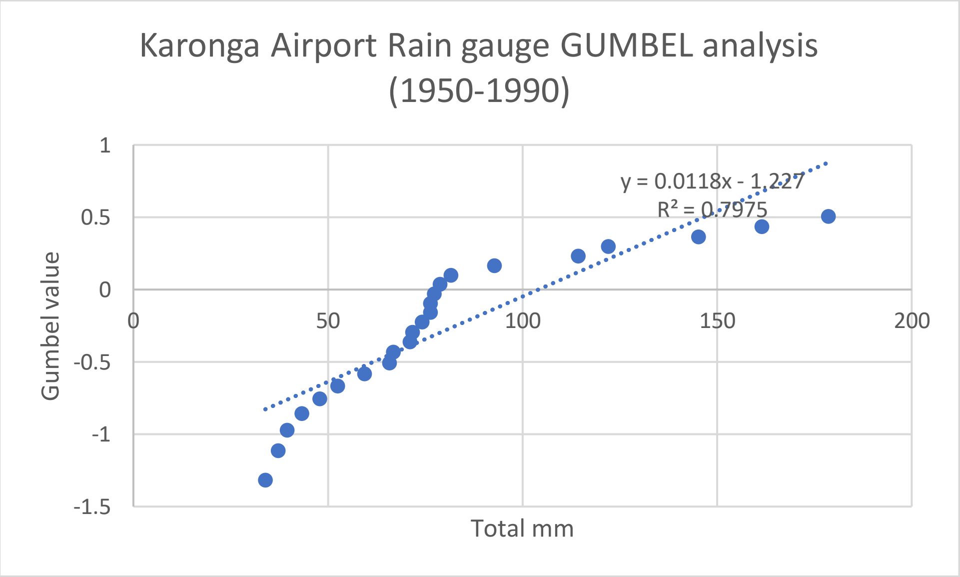


Appendix B: GUMBEL analysis results for GSMaP and Rain gauge datasets



GSMaP Gumbel Results

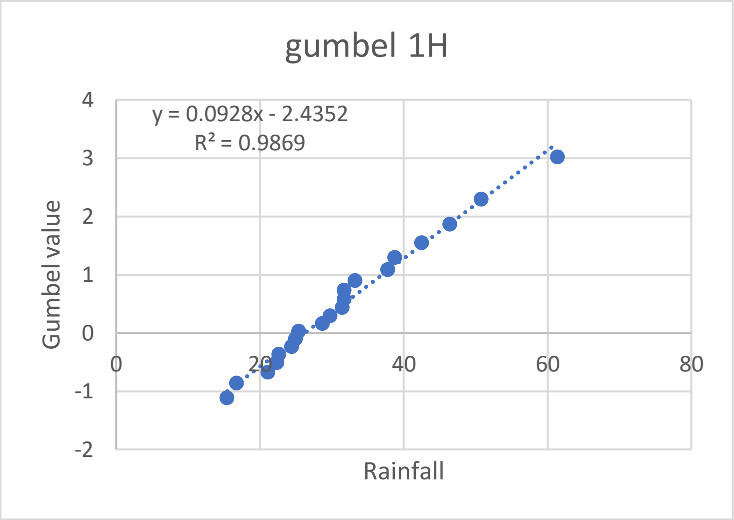
T (year)	Total (mm)
5	180.84
10	249.0607
20	314.4996
50	399.2035
100	462.6772

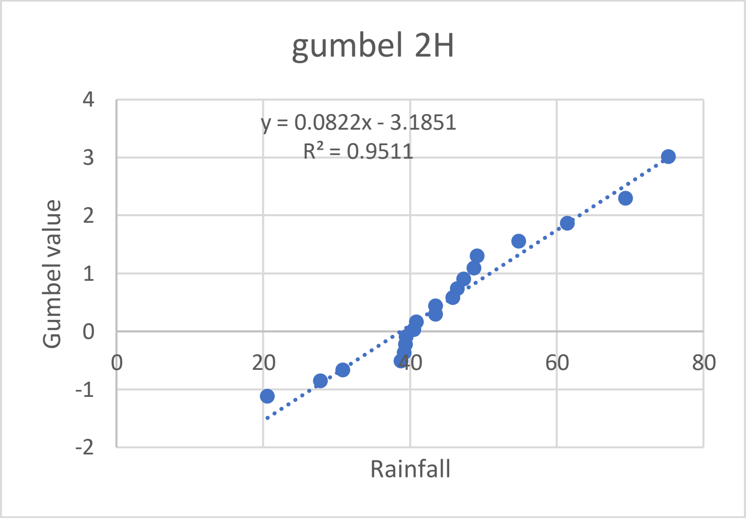


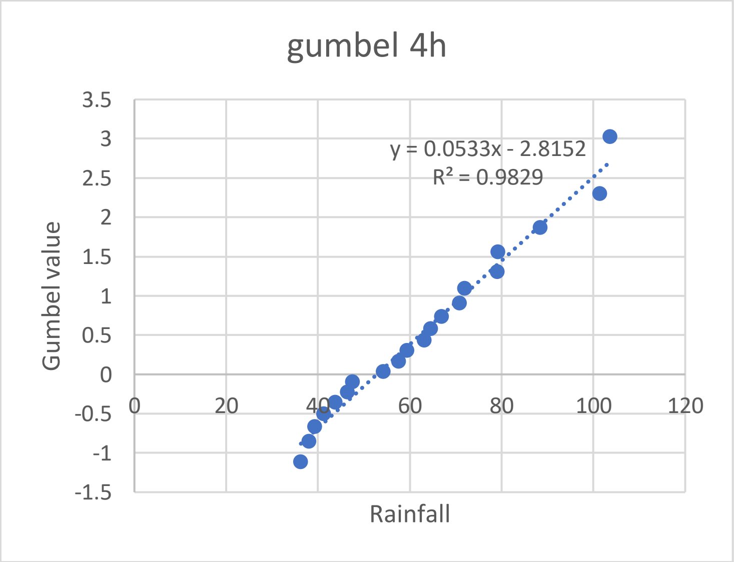
Rain gauge GUMBEL results

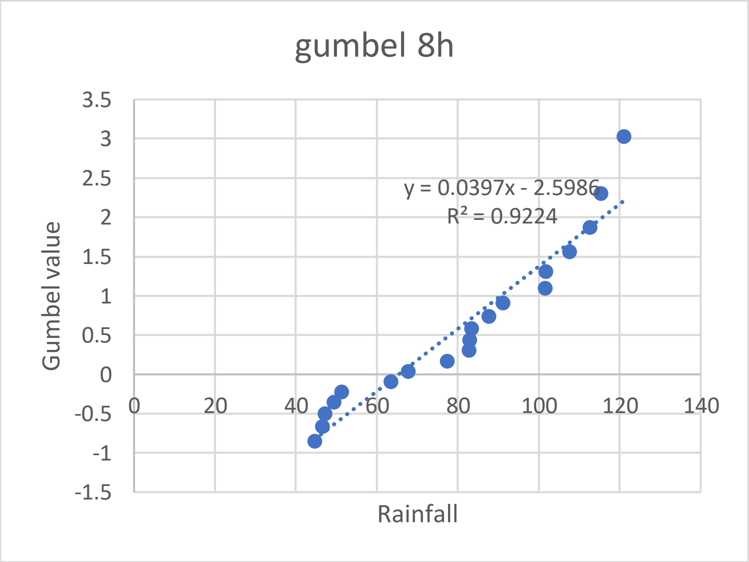
T(year)	Total (mm)
5	231.0966
10	294.6921
20	355.6945
50	434.6558
100	493.8262

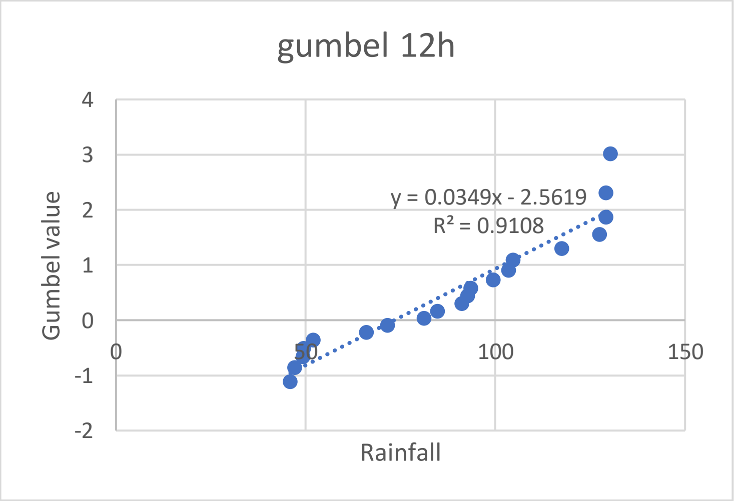
Appendix C: Biascorrected CHIRPS GUMBEL analysis

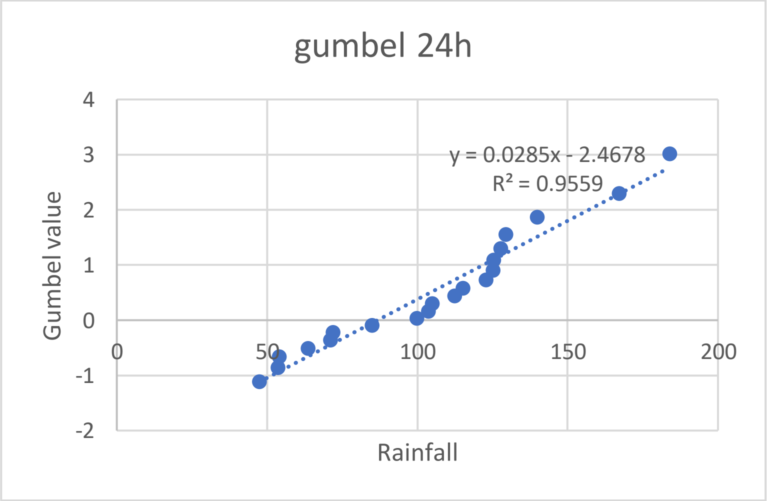




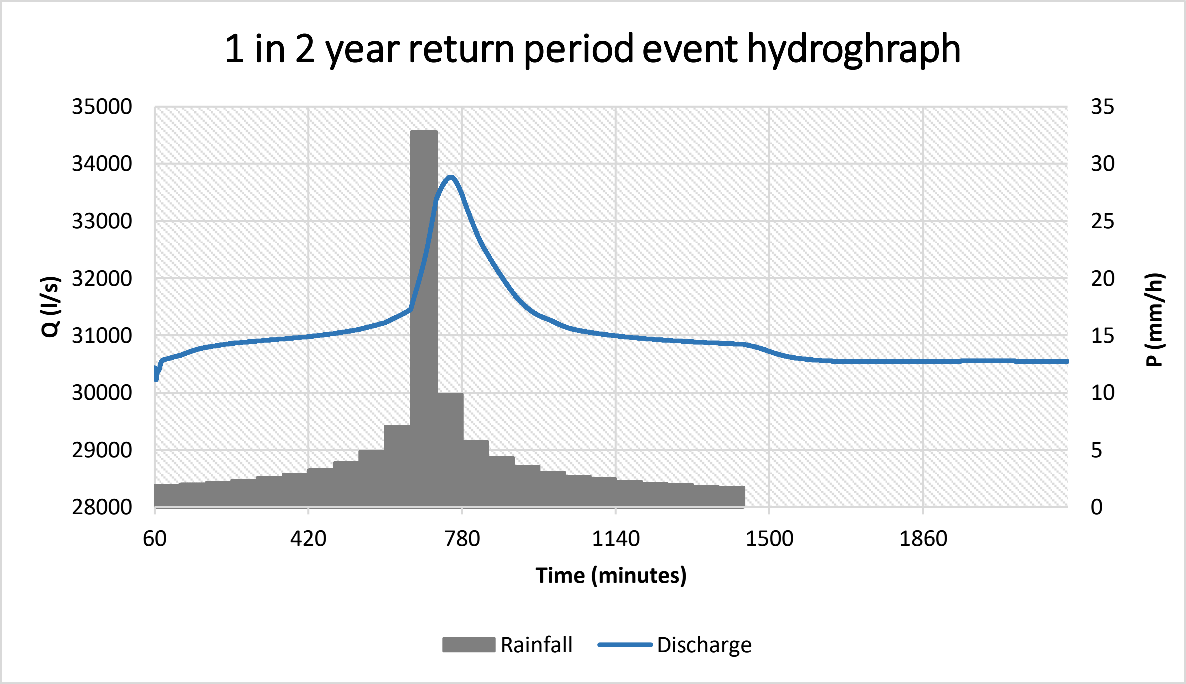


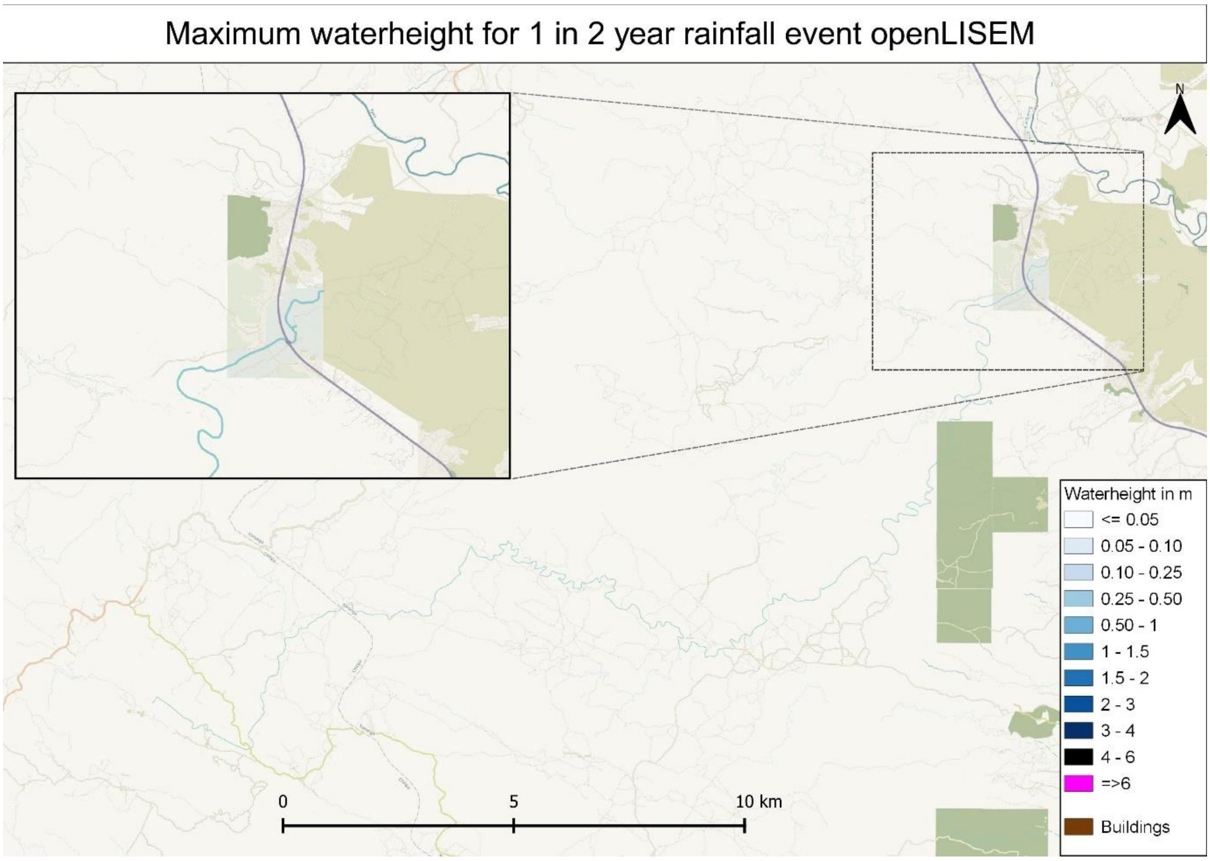


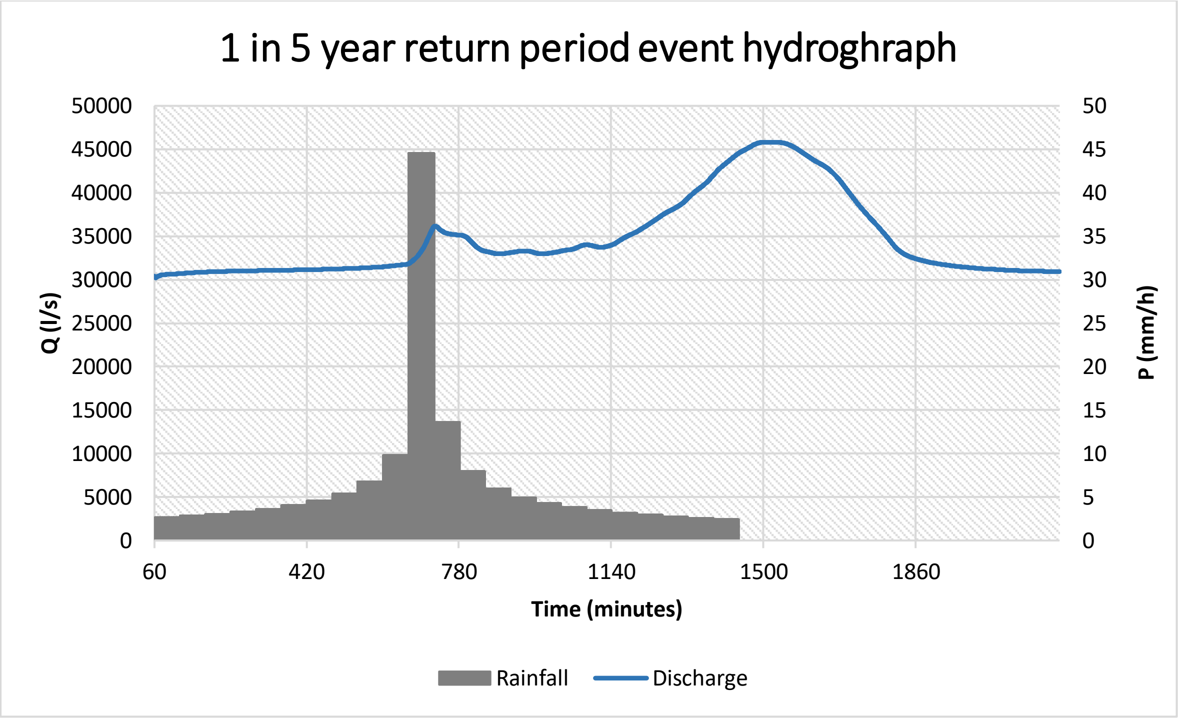


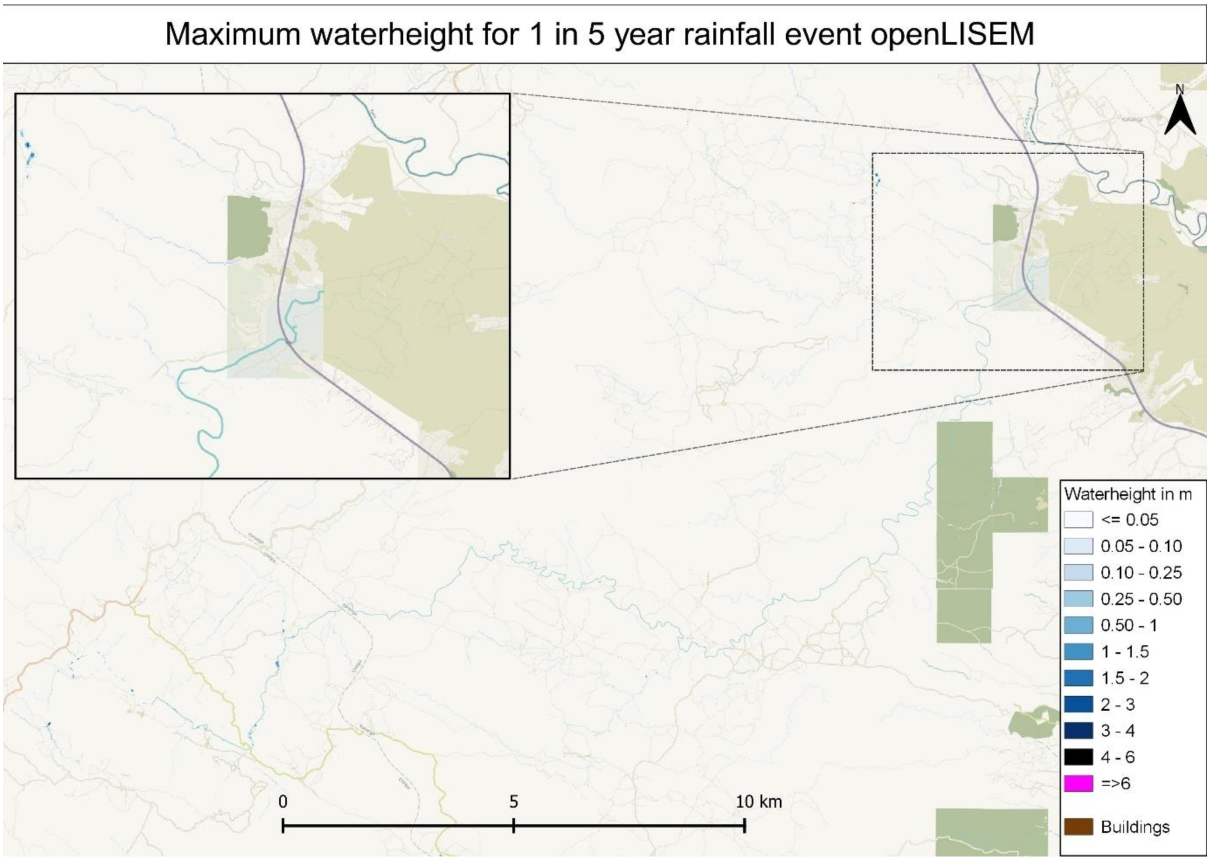


Appendix D: Hydrographs and maximum waterheight maps for 2 and 5 year return period events

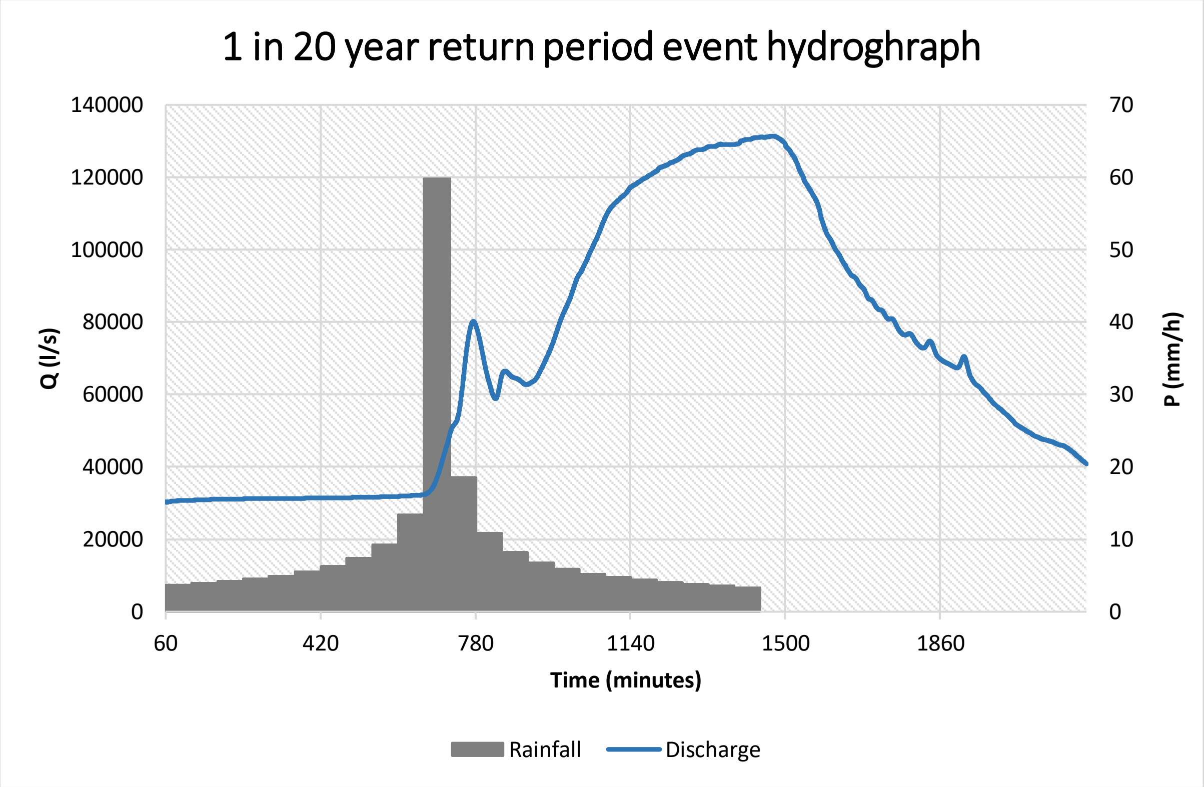

1 in 2 year return period event hydroghraph

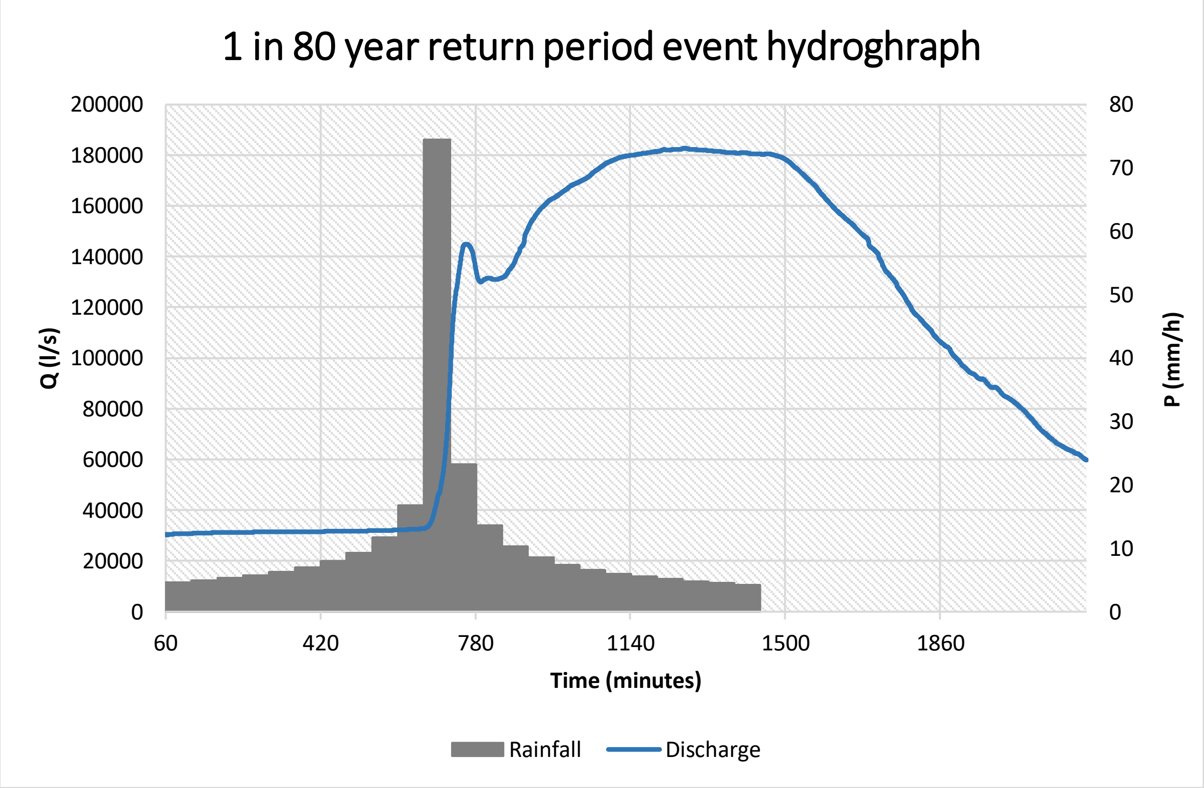

Maximum waterheight for 1 in 2 year rainfall event openLISEM


1 in 5 year return period event hydroghraph
Q (l/s)
50000
45000
40000
35000
30000
25000
20000
15000
10000
5000
0
P (mm/h)
50
45
40
35
30
25
20
15
10
5
0
60
420
780
1140
1500
1860
Time (minutes)
Rainfall
Discharge


Maximum waterheight for 1 in 5 year rainfall event openLISEM
N
Waterheight in m
<= 0.05
0.05 - 0.10
0.10 - 0.25
0.25 - 0.50
0.50 - 1
1 - 1.5
1.5 - 2
2 - 3
3 - 4
4 - 6
=>6
Buildings
0
5
10 km

Appendix E: Hydrographs for the 20 and 80 year return period events





Appendix F: initial input data experimentation

Results waterheight threshold y>0.05=0

lm: 0.021

rmse_lm: 0.3

DT: 0.207

rmse_dt: 0.27

RFR2: 0.586

rmse_rf: 0.195

EXRFR2: 0.631

rmse_exrf: 0.184

ml results river buffer

lm: 0.017

rmse_lm: 0.668

DT: 0.355

rmse_dt: 0.541

RFR2: 0.653

rmse_rf: 0.397

EXRFR2: 0.639

rmse_exrf: 0.405

ml results extra features + buffer

lm: 0.098

rmse_lm: 0.633

DT: 0.173

rmse_dt: 0.606

RFR2: 0.533

rmse_rf: 0.456

EXRFR2: 0.527

rmse_exrf: 0.459

ml results all features + buffer+ coordinates

lm: 0.104

rmse_lm: 0.641

DT: 0.308

rmse_dt: 0.564

RFR2: 0.64

rmse_rf: 0.407

EXRFR2: 0.617

rmse_exrf: 0.42

ml results all features, flooded pixels (=>0.05 only), no buffer

lm: 0.101

rmse_lm: 0.879

DT: 0.362

rmse_dt: 0.74

RFR2: 0.676

rmse_rf: 0.527

EXRFR2: 0.66

rmse_exrf: 0.541

ml results all features, flooded pixels only + rainfall and 100y + 50y + 10y events

lm: 0.112

rmse_lm: 0.797

DT: 0.366

rmse_dt: 0.674

RFR2: 0.683

rmse_rf: 0.476

EXRFR2: 0.675

rmse_exrf: 0.482

ml results all features, buffer + rainfall and 100y + 50 + 10y events

lm: 0.086

rmse_lm: 0.536

DT: 0.622

rmse_dt: 0.345

RFR2: 0.753

rmse_rf: 0.279

EXRFR2: 0.693

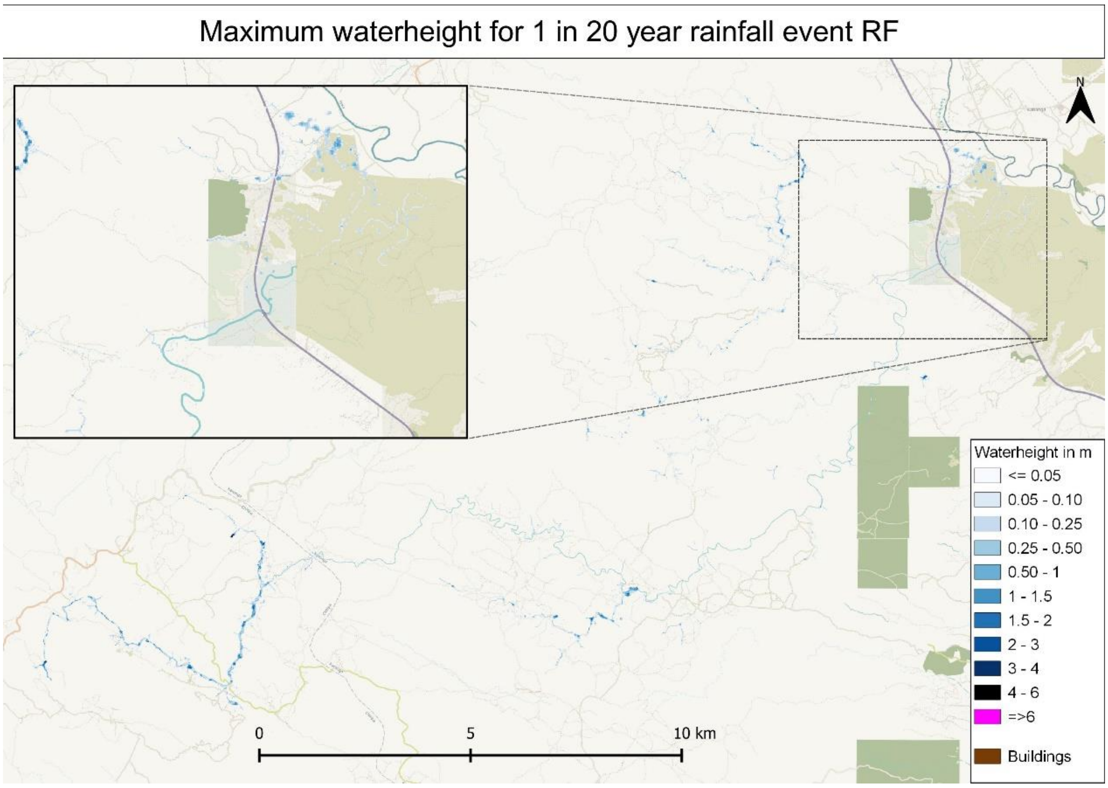
rmse_exrf: 0.311

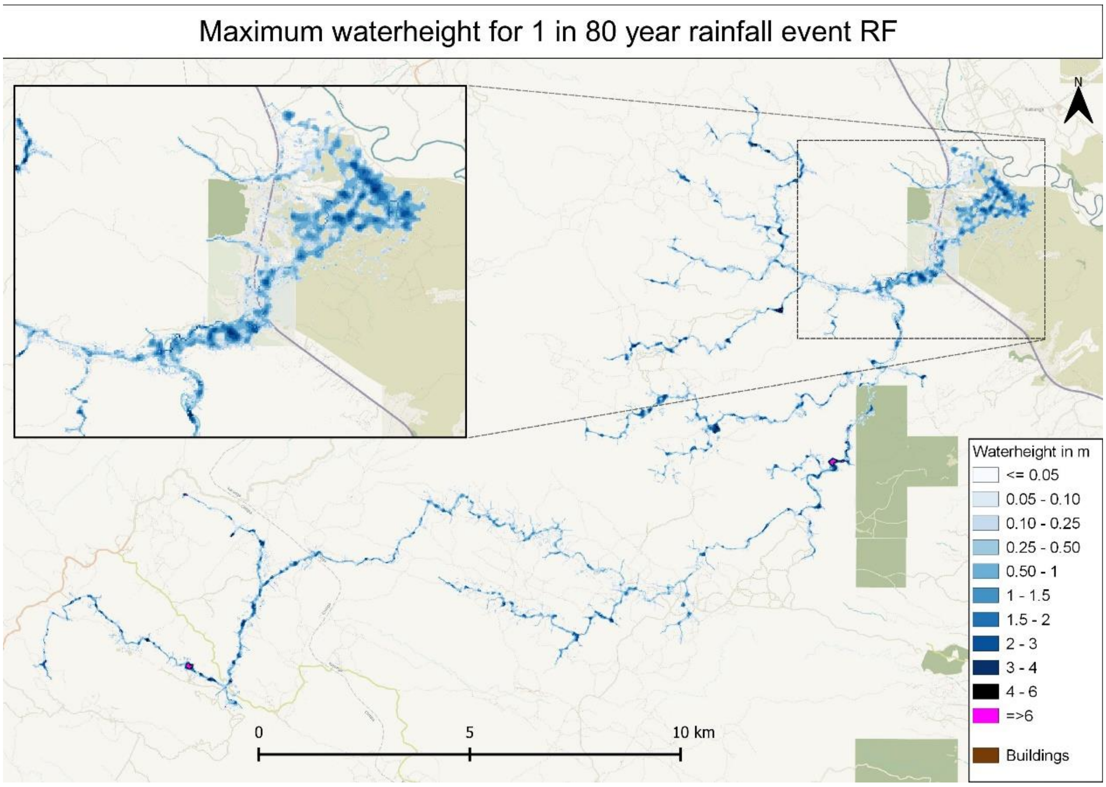
Appendix G: ML algorithms Hyperparameter settings

	RF	EXRF
n_tree	200	200
m_try	‘auto’	‘auto’

	XGB
Objective	‘Squarederror’
n_tree	1000
max_depth	8
tree_method	‘hist’
learning_rate	0.2
m_try	1

Appendix H: Maximum waterheight predictions 20 and 80 year return period events for RF and XGB algorithms


Maximum waterheight for 1 in 20 year rainfall event RF


Maximum waterheight for 1 in 80 year rainfall event RF

Maximum waterheight for 1 in 20 year rainfall event XGB

Maximum waterheight for 1 in 80 year rainfall event XGB