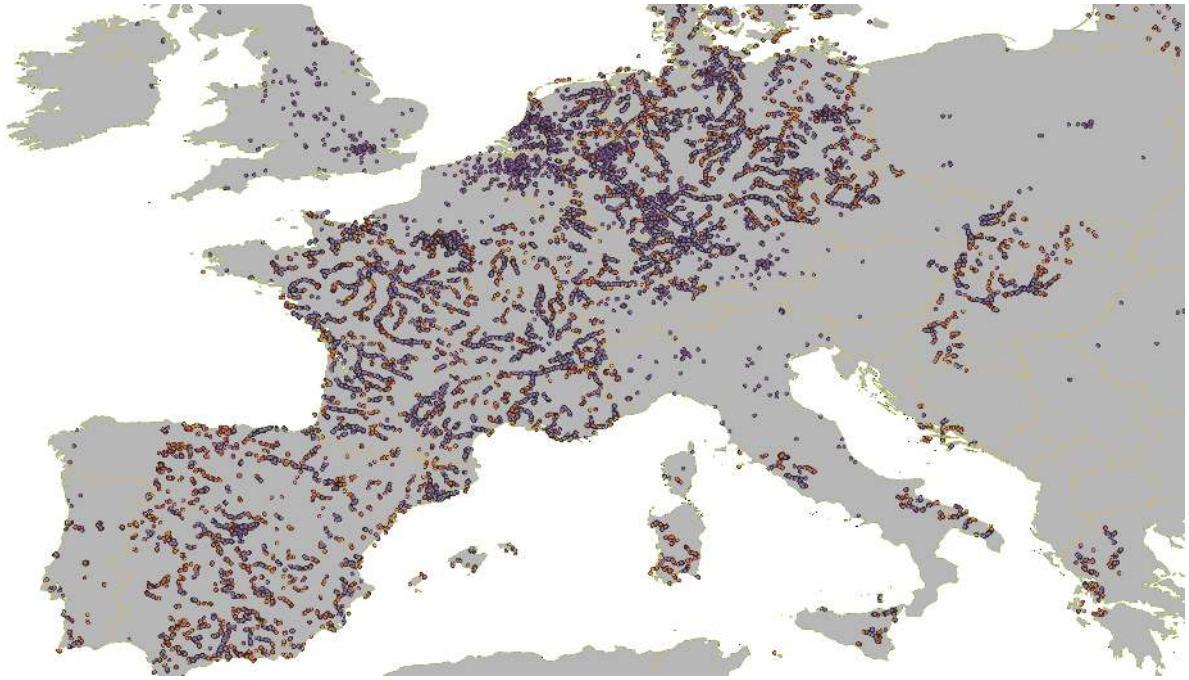
Predicting Flooding from Local Features



Ian C. Ronk

Layout: typeset by the author using LaTeX.
Cover illustration: QGis

Predicting Flooding from Local Features

Predicting Flooding Risk for Pan-European REIT Assets using Local Features

Ian C. Ronk
12664804

Bachelor thesis
Credits: 18 EC

Bachelor *Kunstmatige Intelligentie*

University of Amsterdam
Faculty of Science
Science Park 904
1098 XH Amsterdam

Supervisor
Sander Van Splunter
Arjan Knibbe

KR&A
Vijzelstraat 68
1017 HL Amsterdam

Semester 2, 2022

Abstract

Fluvial flooding is an increasingly occurring phenomenon that bears enormous financial consequences for the regions affected. Several hydro-logical flood simulations have been designed to simulate the flood return periods, including a pan-European simulation by [Dottori et al., 2016] to be able to map out flood risks. In this project, the aim is to continue on the research conducted by [Dottori et al., 2016] and investigate whether it is possible to predict the flood return period of a location on its local features to increase explainability and adaptability of the consequently created model. Examples of such local features are the distance to the nearest river or the artificial imperviousness percentage of the surrounding area. This project consists of two parts: first the local features are identified by a literary research and consequently these features are extracted from open source data sets and a model is trained on these features. The results show that local features are able to predict whether a location will flood within a 20 year flood return period with an accuracy of 97.19% using a Random Forest Classifier. Classifying whether a location will not flood or has a flood return period of 10, 50 or 200 years was done with an accuracy of 90.17% by a Random Forest Classifier. The highest contributing features to these results are the surrounding artificial imperviousness *(Micro 18-21)* and relative height to surrounding area *(Meso 2-4)*.

Contents

1 Introduction

The total amount of damages globally by flooding in 2010 were estimated to be 78 billion dollars and projections estimate that these damages will ramp up to 395 billion dollars by 2050 [Ward and Winsemius, 2018] and that an increased number of people worldwide will be exposed to increased flooding [Tellman et al., 2021]. Having insights in the flooding risk on a certain location is of the uttermost importance as flooding brings enormous damages and can even have disastrous consequences in the case of nuclear reactors if this phenomenon is not taken into account [Bankoff and Lee, 1983]. There are five different types of flooding. The three most important are briefly discussed below[1].

1. **Coastal flooding:** coastal flooding is any flooding that results from the sea. This type of flooding is very sudden, because of tide changes.
2. **Pluvial/Flash flooding:** Flash flooding is the result of exceptionally heavy rain that overwhelms the sewage system. These floods are very unpredictable and sudden.
3. **Fluvial flooding:** fluvial flooding is any type of flooding that results from the overflow of rivers and waterways. They are usually the result of heavy rain that exceeds the river capacity and causes them to overflow.

There have been various models and simulations created, that encapsulate the hydrodynamics of the water. These models may be roughly classified as surface wave stability theories, static equilibrium theories, envelope or limiting operation theories and slug formation theories [Bankoff and Lee, 1983]. Usually both the pluvial and fluvial flooding play a role in urban flooding [Chen et al., 2005], but they are mostly simulated separately as they have a different cause The European Commission, in collaboration with the Joint Research Centre (JRC), has released a fluvial flooding simulation, which uses hydrodynamics of the river basins in Europe to create flooding return period maps for the entirety of Europe [Dottori et al., 2016]. Flood return periods are the estimated time interval between floods of a similar size or intensity. This project continues on the work of [Dottori et al., 2016], which is a research conducted by the European Commission, in collaboration with the Joint Research Centre, which resulted in high resolution (250m) flood hazard maps over different return periods (10,20,50,100,200 years). This simulation is based on a two-dimensional hydrodynamic model. It is outside of the scope of this project to go into the details of such models, but it is important to note that this model was compared to a mosaic of flooded areas detected from satellite images. and that there is overlap between these mosaics and the predicted flood return periods. It

[1]Source: `https://thesafegroup.co.uk/blog/types-of-flooding-in-the-uk`

is therefor assumed in this project that it is possible to use the results from this research endeavour as true data.

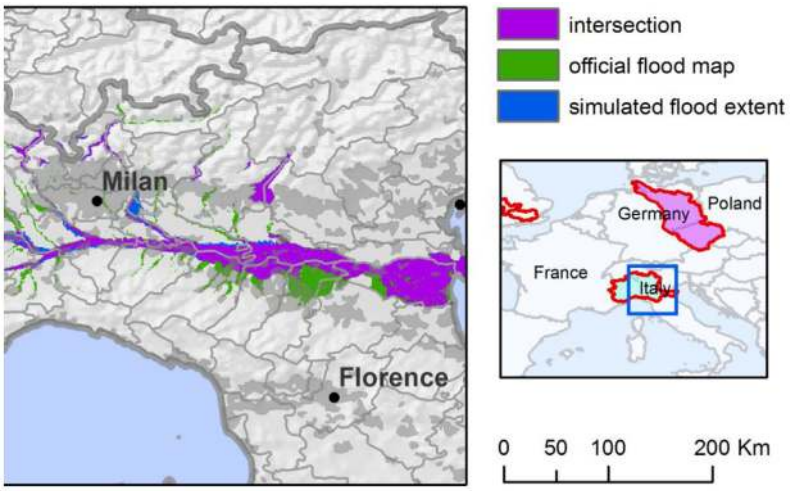


Figure 1: Results from the flooding simulation project [Dottori et al., 2016]

Over the last two decades, data has claimed a bigger role in deciding which assets to invest in. This is also the case for real estate investors, who look at both intrinsic and extrinsic data. Intrinsic values are used most frequently because of their clear connection to the asset. Examples include the amount of square footage of a property or whether the apartment has a balcony. Extrinsic values are harder to quantify. This extrinsic or alternative data consists of attributes of assets that are outside of the intrinsic values, such as the distance to a station from the asset or the temperature development in a certain area. An important subset of these properties is the so-called ESG data.

Over the past years ESG data has gotten more important for investors and real estate funds alike. ESG is an acronym for Environmental, Social and Governmental data, which stands for the categories of alternative data that are taken into consideration. An example of ESG data is as follows: Imagine a scenario where there are two different prospective locations for a super market: a busy street with other established businesses and stores and near an upscale neighbourhood, where profits are fairly certain or a location in a more quiet deprived neighbourhood, where profits are less assured because of the financial situation of the inhabitants of that area. An ESG conscious investor or fund would choose for the latter as it is in line with social consciousness. This type of investing has become more popular over the last years because of a shift from sole focus on profit to more ESG-conscious decisions. However, ESG is not only about being socially conscious, but also about assuring a certain stability in the safety of the assets and investments.

The governmental part of ESG data for example concerns itself with the degree of corruption of a region. The environmental part of ESG data concerns itself with for example the pollution a company or asset imposes on the environment. It however also includes information on risks and adaptability of climate change and climate risks, such as droughts, wildfires and flooding. The following chart shows a division of the occurrence of these environmental disasters in Europe:

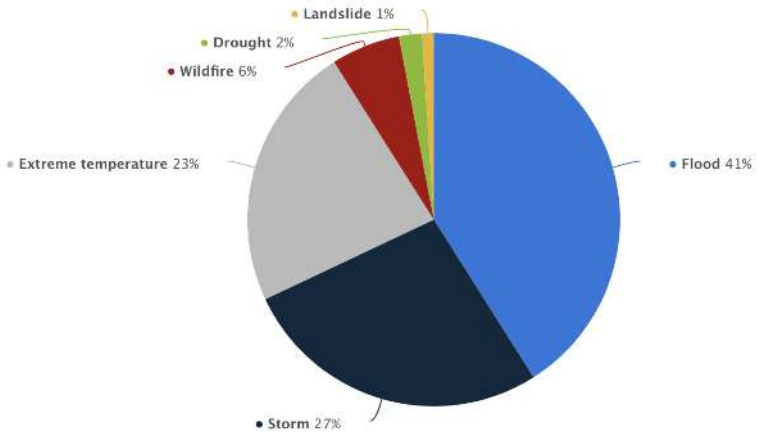


Figure 2: A chart of the most occurring environmental hazards[2].

Figure 2 shows that flooding is the most common natural disaster in Europe, which makes this hazard a particularly concern to European investors. The above mentioned hydrological simulation, achieves results that are comparable to observed flooding return periods [Dottori et al., 2016] but has two downsides: (a) the lack of explainability and (b) the lack of adaptability. Explainability and adaptability of the model are important factors for the resilience of a flooding prediction mode. Explainability is of major importance for planning and risk assessment for for example REIT fund managers, that want to know an answer to the question as to why flooding would occur. An approach that would solve these drawbacks is the implementation of local features. In this context local features are defined as features that arise from the location of the asset itself or the near surroundings. Local features combat inexplainability by attributing features on the location and surroundings to a conclusion of whether a location floods in a certain period and inadaptibility by allowing these local features to be changed and still getting results. Such change could happen over time, such as an increased intensity of rainfall or changes to the surrounding landscape. This project tries to implement such a model that uses local features and tries to answer the question:

[2] https://www.statista.com/statistics/1269886/most-common-natural-disasters-in-europe/

Can risk of flooding be explained by local values of specific locations? To reliably answer this questions, the following sub questions are raised:

1. *Can local features be used to assess a risk of flooding with sufficient accuracy and explainability?*
2. *Which features contribute the most to predicting whether an asset will flood or not?*

The goal of this project is to find features that allow an appropriate answer to the main research question and create a reliable model. It is hypothesized that there are some features that allow for the prediction of a flooding risk, which in turn allows for a relevant report that could be generated, which gives insight as to why the location is prone to flooding or not. It is also hypothesized that flood return period labels can be assigned to the locations with a significant accuracy. This project will be conducted in three steps. Firstly, a literature review will be conducted to index the features that could contribute to a valid flooding explanation, with research backed explanations. Secondly, the architecture and helper functions are build and described: *how was the data retrieved and how is the data processed and the analysis executed?*. Thirdly, the results are compared and a feature analysis is conducted to discover the importance per feature and their respective contribution, which includes a conclusion and discussion on the hypothesis.

It is important to stress again that this project is based on one important assumption: The flooding map is seen as the true data, with the assumption that the simulated data corresponds to historical flooding data. This is not always the case as can be seen in Figure 1, although there is significant overlap and is therefor assumed as such.

2 Feature Research

This chapter puts forward all local factors, that influence the risk of fluvial flooding. These factors were found in a literary review. Per factor, the different features that define the factor are discussed and for all of the features, (a) an explanation is given (b) why it is expected to have an impact, (c) what type of feature it is and (d) what the hypothesized impact is of the feature, if the feature is investigated in the project. An example of such multiple features for one factor is the imperviousness of the ground. This factor can be extracted on solely the location itself, but also for the surrounding area. The features are grouped on their core 'area of effect':

1. **Micro**: data that encompasses elements on the exact location
2. **Meso**: data that requires the nearby surroundings

3. **Macro**: data that encompasses a larger area of effect

For all of the different features a unique identifier inside of these areas of effect is assigned to make quicker distinction possible between the different features. An example of such an identifier is: *Micro 1* for the One day precipitation maximum.

2.1 Micro Feature 1-14: Precipitation

Maximum is for the period 2000-2019 Precipitation encapsulates the rain and snow that falls at a particular location. Precipitation is unfortunately unpredictable and unstable, as there are periods where there is an excess of rain or on the other hand of drought. Precipitation has been stated as one of the leading climate factors of flooding by numerous research endeavours [Xu et al., 2017] and is therefor indispensable in this feature analysis. Pan European precipitation data is however hard to find, with the only high resolution data available being the monthly precipitation per month at a spatial resolution of 250 meters. Flooding however does not occur as a result of continuous precipitation, but rather of intense short durations. Such occurrences are better captured by using extreme precipitation data. This heavy precipitation data can be retrieved from Copernicus [Copernicus CCS, 2020] and was created by the JCS in collaboration with the TU Delft [Muis et al., 2020]. This precipitation data has a spatial resolution of 10x10km and covers the one day precipitation maximum and five day precipitation maximum over a period of 20 years (2000-2019). Consequently, two features can be deduced from this dataset: *one day maximum and five consecutive day maximum.* The one day maximum is a good feature, as some floods occur within one day of extreme rain, while others take multiple days to build up[3], which gives merit to the feature: *five consecutive day precipitation maximum.* this thesis hypothesizes, supported by literature [Titley et al., 2021] that these two features are a major contributing factor to the predictability of the model. The monthly precipitation feature that was described above is also included. This data was extracted from [Fick, 2017] and has a spatial extent period from 1960-2000. This thesis hypothesizes that this feature will be less important than the precipitation maxima, as floods do not occur during average precipitation events, but rather in extreme events. In conclusion this factor entails the following micro features:

- **Micro Feature 1:** One day precipitation maximum
- **Micro Feature 2:** Five consecutive day precipitation maximum
- **Micro Feature 3-14:** Monthly average precipitation over a 40 year period (for each month respectively)

[3]Source from US Environmental Protection Agency: `https://www.epa.gov/climate-indicators/climate-change-indicators-heavy-precipitation`

Soil Type	Bare Soil Imperviousness (%)	Turf Grass Imperviousness (%)
Sand	11	0
Loamy Sand	26	15
Sandy Loam	39	28
Sandy Clay	58	47
Clay	100	100

Table 1: Degrees of Imperviousness from the USDA [Osouli et al., 2017]

2.2 Micro Feature: 15-20: Permeability and Hardening

Permeability is the ability of ground surfaces to absorb water into the ground. When a surface is impervious, the water is unable to infiltrate the ground[4]. High imperviousness is problematic for the absorption of water coming from floods, which is not able to be drained into the ground. As this water will stay above the surface. Permeability of surfaces is influenced by two factors: (a) natural imperviousness and (b) artificial imperviousness. Natural imperviousness depends on the ground type: certain soil types naturally allow water to perforate, while others do not. Table 1 shows different soil types with their respective percentage imperviousness (PIMP), taken from the United States Department of Agriculture [Osouli et al., 2017].

The percentage is calculated by PIMP$= 6.4 * J^{0.5}$ where J is the number of dwellings per hectare [Butler et al., 2018]. Artificial or urban imperviousness entails the imperviousness of the ground by man-made endeavours [Jacobson, 2011]. In some cases the imperviousness increases to 100%, as water is not able to penetrate concrete surfaces: if several houses are build in an area with tarmac roads, the water cannot permeate the ground and the imperviousness percentage increases to 100%. This type of imperviousness is called hardening and is often referenced as one of the contributing factors to urban flooding [Sohn et al., 2020]. Both types of imperviousness are indispensable from this project as they both contribute to the extend of flooding. The last 30 years research on urban flooding has increased; multiple case studies have been conducted such as [Gholami et al., 2010] in the Hajighoshan watershed in Northern Iran, which found by comparing satellite imagery that the increased urbanisation brought hardening, which increased the runoff generation potential. The runoff generation potential is the portion of precipitation or snow and glacial melt that flows across the landscape until it reaches streams. This means that more precipitation will flow on the surface, instead of underground as it is not able to permeate the surface sufficiently and this supports the potential of this feature. Although this case study was conducted in Iran, such

[4]Source from Lecture Series Lumen Learning: `https://courses.lumenlearning.com/geo/chapter/reading-porosity-and-permeability/`

changes are universal and also occur in Europe [Langanke, 2021]. Other studies found the same conclusion: runoff occurs more frequently because of urbanization and small rainfall events of 1–2mm will still cause runoff from impervious surfaces [Ladson, 2019]. It is therefor hypothesized that increased imperviousness in the surrounding areas will increase the chance of flooding. It is important to take the imperviousness over different distances into account as the water could travel far away from the river. Therefor the following features are selected:

- **Micro Feature 15:** Ground type on the location
- **Micro Feature 16:** Leading ground type in a 1 km area
- **Micro Feature 17:** Leading ground type in a 5 km area
- **Micro Feature 18:** Ground imperviousness on the location
- **Micro Feature 19:** Ground imperviousness in a 500m area
- **Micro Feature 20:** Ground imperviousness in a 1 km area
- **Micro Feature 21:** Ground imperviousness in a 5 km area

A pan-European soil type map is provided by the European Environment Agency (EEA) in the form of the Corine European Soil database [EEA, 2003]. The data is available with a resolution of 250 meters or 3" and the different soil types are added to Appendix: Ground Types. Due to restricted open resources a ground map with more precise ground types could not be found with a high enough resolution and this problem is taken up into the Future Work section. The different leading ground type features can be extracted using helper functions that query the surrounding area of the coordinate. The implementation of this helper is explained in the next section: Method. The artificial imperviousness data can be retrieved from the Copernicus database [Copernicus, 2015], Copernicus is a European programme for monitoring the Earth, in which data is collected by Earth observation satellites and combined with observation data from sensor networks on the earth's surface[5]. Again, the surrounding values can be retrieved from these maps using a helper function. The Copernicus data is retrieved from satellite data and has a resolution of 10 meters. This is incredibly precise, but too precise for this project, due to computational constraints. It was therefor downsized to 100x100m. A further elaboration on resolutions that are used in this project will be conducted in the methods section.

2.3 Groundwater

Groundwater plays a role in two types of flooding: (a) groundwater flooding and (b) fluvial flooding. Groundwater flooding occurs when the natural underground

[5]Information from Copernicus About: `https://land.copernicus.eu/about`

drainage system cannot drain rainfall away quick enough, causing the water table; the underground water level, to rise above the ground surface and in its turn creating a flood[6]. Groundwater flooding is a different type of flooding which occurs in a different way from fluvial flooding. This so called water table is also important for fluvial flooding as it is correlated with the amount of water the ground can absorb, before rising above the ground surface and becoming impermeable. It is important however to note that there is a difference between the impact of the ground water level on fluvial flooding and groundwater flooding. This factor as feature can be quantified as the average ground water height at that location. A dataset has not been found for this feature and is thus not included in this research. This project hypothesizes that if this feature were to be included in the model, the groundwater feature would give better insights in the absorption ability of the ground and is hypothesized to contribute, although to a lesser extend than the before mentioned factors to a better prediction on whether a certain area would flood. Especially the extend of this flood: *how far the water would be able to flow*, would be mapped out better. A combination of this feature and the Micro Features 1-20 would give a complete picture of the handling of water of the surrounding area around the location.

2.4 Meso Feature 1-8: Geographic factors outside of the river

There are several elements that contribute to the flooding risk of a location, that are not substantiated by conducted research, but rather by reasoning. These include the following features:

- **Meso Feature 1:** Distance to river
- **Meso Feature 2:** Relative height to surrounding area (500m)
- **Meso Feature 3:** Relative height to surrounding area (1km)
- **Meso Feature 4:** Relative height to surrounding area (5km)
- **Meso Feature 5:** Depression (500m)
- **Meso Feature 6:** Depression (1km)
- **Meso Feature 7:** Depression (5km)
- **Meso Feature 8:** Relative height to river

[6]Source from Geological Survey Ireland: `https://www.gsi.ie/ga-ie/programmes-and-projects/groundwater/projects/gwflood/Pages/What-is-groundwater-flooding.aspx`

2.4.1 Meso Feature 1: Distance to river

The distance to the river is calculated by finding the distance between the nearest river point and the coordinate. The river map [Dottori et al., 2016] is used for this calculation and this feature is therefor accurate to 250 meters. The distance to the river is an important geographical feature as fluvial flooding usually only occurs close to the river and does not move extremely far land inwards. This feature is hypothesized to be one of the contributing features to the performance power of the model as a large distance would mean that the water can not reach the location. Due to computational restrictions, this feature was only defined for rivers that are within 20 kilometers of the location.

2.4.2 Meso Feature 2-7: Relative height to surrounding area and depression

The relative height to the surrounding area is calculated by using elevation maps. By looking at the local area, it is possible to see whether the water will flow down into the depression if the location were to be in a depression. This feature comes in several different forms: (a) relative height and (b) depression. The relative height is the absolute difference between the height of the location and the height of the surrounding area and the depression value is a boolean value (True or False), encoding whether the location is in a depression (relative height < 0) or not (relative height $>= 0$). This therefor encodes the same value. Both features are created to see whether the relative height difference is the decisive factor or solely being in a depression or not. To get these values the elevation map [Neteler et al., 2022] is used. This map has a resolution of 100x100m horizontally and in the vertical direction an accuracy of approximately 7 meters. It is hypothesized that these features have a rather big impact on the predictive capabilities of the model, however lower than Meso Feature 1 and Meso Feature 8.

2.4.3 Meso Feature 8: Relative height to the nearest river

The relative height to the river is a combination of the results of the previous features. It is hypothesised to be one of the most contributing features as when a location is situated lower than the nearest river, it is prone to flood, but if it is higher than the nearest river, it should not flood at all. The accuracy of this relative height in the vertical direction comes from the elevation map and is 7 meters. This data is a combination of the river map [Dottori et al., 2016] and the elevation map from [Neteler et al., 2022] and therefor the spatial accuracy takes on the shape of the less precise river map.

2.5 Nearby River Properties

Besides the elements that contribute to flooding risk outside of the river, properties inside of the river are also of indispensable importance[7]. Firstly the discharge, the volume of water flowing through a river channel is one of the most important factors for fluvial flooding: if the discharge increases suddenly, the water is not able to stay inside of the river bank and will flow out. The discharge is calculated as follows:

$$Discharge = \sigma * v \tag{1}$$

Where v is the velocity and σ is the cross sectional area. These two properties: cross sectional area and velocity are the determining factor behind the intrinsic river properties that contribute to flooding. The two factors, including their respective features are discussed below. As well as another factor: River Sediment. This yields the following subsections:

1. Cross Sectional Area (σ)
2. Velocity (v)
3. River Sediment
4. Problem: Availability of the features

2.5.1 Cross sectional area (σ)

The cross sectional area is calculated by multiplying the river width with the average river depth. These values are different for every part of the river, but it does give an insight in the cross sectional area on that specific location. In addition to this, the change in cross sectional area is also an important feature: if a river suddenly shrinks, the river can not handle the influx of water and floods. This concludes that the following features should be investigated in future research:

- Width of the river
- Average river height
- Change in cross sectional area of the nearby river

[7]From the Lecture Slides: `https://www2.tulane.edu/~sanelson/Natural_Disasters\riversystems.htm`

2.5.2 Velocity (v)

The velocity of the river depends on three aspects: (a) the long profile, (b) the shallowness of the river and (c) the curvedness of the river. The long profile is a plot of the elevation of the river versus the distance over the whole length of the river. A long profile shows that the elevation decreases more rapidly when moving closer to the source of the river. The shallowness of the river also contributes to the velocity of the water moving through the river. If a river is shallow, an equal amount of water will travel quicker than in deeper rivers, as the cross sectional area of the river decreases and thus the velocity needs to increase. The third feature is the curvedness of the river. If the river meanders significantly, the river slows down because of friction to the river edges. Although the velocity of the river is variable, approximations can be made about the velocity using the following features:

- Distance to origin
- Shallowness of the river (this feature is the same as the average river height)
- The curvedness of the river

2.5.3 River Sediment

River sediment is all of the substances, such as clay or stones that flows through a river, excluding the water itself. Measuring the degree of river sediment or the substances that flow through a river is highly variable and can only be measured on a specific location. It is therefor out of the scope of this thesis to conduct this research.

2.5.4 Availability of the features

Because all of the above mentioned features depend on very detailed geographical data, these features are not implemented. The current river maps that are used in this research have a resolution of 250m, which is too imprecise to derive this data from. The only feature that can be calculated is the distance to the origin, but this feature in itself is not useful and is thus omitted from this project.

2.6 Fluvial Flooding Mitigation Investments

Although most of the current urban developments worsen the impact and intensity of flooding, there are endeavours to decrease the flooding risk. This decrease originates from the creation of waterworks. A prime example of a country, which invests tremendous amount of resources in the construction of waterworks is The Netherlands [Bos and Zwaneveld, 2017]. Examples of mitigation methods that

can be used are the construction of dikes or rerouting rivers [Verol et al., 2019], to allow for a better flow of water (why this is important is portrayed in the intrinsic factors of rivers below). As previously mentioned, there are several ways the river can be 'tamed'. Especially in the future this is getting more and more important, as the amount of discharge of rivers will increase, with an increased precipitation and urbanisation (and thus in its turn more imperviousness). River mitigation strategies come in two different categories (a) Artificial flooding mitigation (b) Non-artificial flooding mitigation. Although both are hard to quantify and not included as features for the model, they are important to mention and therefor discussed.

2.6.1 Artificial flooding mitigation

Artificial flooding mitigation methods include the construction of man-made objects, such as dikes or rerouting a river. These two methods were described briefly in the paragraph above, but make use of terraforming and the addition of other constructions to mitigate the flooding risk. The benefit of using artificial flooding mitigation is that it guarantees safety as a dike would have to break to let water through, unless it breaks. A drawback of this type of mitigation is the practical limits that it imposes: for example, if locks are implemented, water traffic is severely limited [Verol et al., 2019].

2.6.2 Non-artificial flooding mitigation

An alternative and more progressive way of flooding mitigation is to terraform the landscape in a less extreme way. Instead of building barriers to keep out the water, the river is given extra space to flow or the surrounding area is made more permeable to allow water to flow through easier [Verol et al., 2019]. This implementation is more sustainable than the artificial flooding mitigation and fits into the ESG approach: making environmentally and socially conscious decisions.

2.7 Macro Feature 1-4: Other Features and Problems

There are features that are not really relevant on first glance, but could turn out to be valid substitutions for other features that were not able to be investigated. All of these features are rather general, in terms of not being flooding centred, but their relevance is discussed below. This analysis looks into the following features:

- **Macro Feature 1:** GDP
- **Macro Feature 2:** GDP per inhabitant
- **Macro Feature 3:** Government Long Term Vision

- **Macro Feature 4:** Quality of infrastructure

These features could be a more tangible substitute to the Fluvial Flooding Mitigation Investments. A reasoning for this is discussed per separate Macro feature. Note that these correlations are speculated and no supporting evidence was found.

2.7.1 Macro Feature 1-2: Gross Domestic Product (GDP) (per capita)

Mostly poor and under-developed countries are at risk of flooding[8]. This is not only because most of these countries are in climate change prone areas, but also because their water infrastructure and sewage systems are under-developed. Because the fluvial mitigation investments are hard to construct and measure for countries and even harder for more precise areas, an alternative approach could be to look at the GDP of the country or sub region (in this case NUTS3), to get an insight of the wealth of the area. This data is readily available from the KR&A portal for NUTS3 level resolution and is measured in millions of euros (per capita) at current market prices. It is hypothesized that this feature has some impact on the accuracy of the model, although this impact is rather small. This is because of the fact that GDP is not directly correlated to the spending on flooding infrastructure and because flooding avoidance policies are not homogenous across Europe: whilst a country or region may be wealthy, it could have enormous flood risk and still not enough money to put the correct measurements into place or the other way around where a poor region would not be able to pay for it, but because it is far removed from a river it cannot be flooded.

2.7.2 Macro Feature 3: Government Long Term Vision

The government long term vision data was retrieved from the World Economic Forum [World Economic Forum, 2019], on a country resolution and includes a score from 0 to 7 from a survey on the question: *"In your country, to what extent does the government have a long term vision in place?"* This data has a low resolution and is not analytical, so this thesis hypothesizes that this feature will not have a high importance. The Government Long Term Vision could attribute to a more general Fluvial Flooding Mitigation Investment feature as this long term vision could include a water management plan, which would improve the whole flood mitigation in a country, such as in The Netherlands.

[8]From the Zurich Flood Resilience Alliance: `https://floodresilience.net/countries/`

2.7.3 Macro Feature 4: Quality of infrastructure

A good quality of the complete infrastructure in a country could constitute an advanced flooding mitigation infrastructure and could therefor be part of the. This data is retrieved from the World Economic Forum [World Economic Forum, 2018] and shows the quality of the infrastructure of countries from 0 to 7. This data does not have a high resolution and because of the uncertain correlation between this feature and the flooding mitigation feature, this project hypothesizes that this feature will not have a high importance.

2.7.4 Alternative uncountable factors: incidental

There are of course factors which are not quantifiable, but influence the proneness of a certain area to flood significantly. These factors are briefly explained below, including a reasoning as to why they are not quantifiable and a hypothesis how these factors would influence flooding, maybe not in the flooding return periods calculated in the simulation, but in historical flooding occurrences. This is usually because incidental occurrences of the following categories occur:

- Breaking of a dam
- Collapse of dikes
- Destruction of a lock

These factors occur randomly and are thus not countable, although being strongly dependent on the amount of investments that are put into the water management infrastructure. It is also not possible to predict such events, although some factors could lead to their explanation, such as the Macro Features 1-4, as an increased investment and vision into waterworks allows for better quality dams, dikes and locks. This would result in less frequent failure of these construction and these incidental occurrences will decrease.

2.8 Conclusion and Remarks

One can conclude from this literary research that there are a multitude of local features that can be used to predict flooding on a certain location. These features can be broadly categorised into the categories: (a) Micro (b) Meso (c) Macro. These features are expected to give flooding risk predictions for specific locations to a certain degree and this hypothesis will be put to the test. This research will be conducted over the course of the consequent part of the paper. The features displayed in Table 2 were researched and their respective hypothesis are summarized.

Feature	Hypothesised Impact
Micro 1: Max Precipitation 1 Day Period	Has a significant impact
Micro 2: Max Precipitation 5 Day Period	Has a significant impact
Micro 3-14: Average Monthly Precipitation	Not a significant impact
Micro 15-17: Ground Type (point/radius)	Has an impact, but not significant
Micro 18-21: Imperviousness (point/radius)	Has a significant impact
Meso 1: Distance to river	Has an extremely significant impact
Meso 2-4: Relative height to surrounding area	Has a rather big impact
Meso 5-7: Depression	Has a significant impact
Meso 8: Relative height to river	Has an extremely significant impact
Macro 1: GDP	Not a significant impact
Macro 2: GDP per capita	Not a significant impact
Macro 3: Government Long Term Vision	Not a significant impact
Macro 4: Quality of infrastructure	Not a significant impact

Table 2: Hypotheses for the implemented features

There is however an extra remark about the features. The current literary review could not five a definite answer to the question: *what is the furthest extend of river flooding?*. Extra research has to be conducted to answer this question and this would give a better idea on what scales to use for distances of the specific features. Examples of features that could be subject to change are Micro Features 16 and 17, where their respective radius of the area could increase or decrease depending on the outcome.

3 Introduction to Geographic Data

This project is centered around geographic data as all of the data that will be retrieved needs to be available on a pan-European scale. Geographic data or geodata is defined as data and information having an implicit or explicit association with a location relative to Earth. Geodata exists in several forms, but is divided into two general categories: vector data and raster data. As the names suggest, the first type of data is based on vectors and the second one with pixels in a raster. Working with these types of data, however is rather different. Both will be described in more detail down below. As this project mainly exploits the second type of geographic data: raster data, this type of data is explained in more detail

including a very important aspect of working with this data: resolution. Another important aspect of geodata is their projections. Because of the curvature of the Earth, coordinates cannot be put on the map 1:1. To tackles this problem, projections are used. How these projections work and how one can work with them is explained below.

3.1 Vector Data

Vector data consists of points, lines and polygons to encode information. These types of elements have labels assigned to them, which indicate categories or proper names. Points are usually used to indicate cities or points of interest, such as for example the location of a restaurant on Google Maps. Lines are used to create a path between two points. This way, they are used to represent streets or rivers [Lehner and Grill, 2013]. Polygons can be used to represent areas, such as lakes or countries. Vector data is the spatial data most people are familiar with, as it is the format presented in mapping software such as Open Street Maps and Google Maps. The advantage of vector data is that the data image can be enlarged without loss of quality, which makes it more convenient to work with in editing programmes such as QGIS, as zooming in takes less time and more precise analysis can be made. A drawback of this data is that gradual changes such as temperature cannot be encoded in an efficient manner and it is therefor not used to encode such information. Vector data comes in many shapes and forms, but the most used file type is a shapefile. Shapefiles were introduced with the release of ArcView 2 in the early 1990s. A shapefile is a non-topological data structure that does not explicitly store topological relationships [Theobald, 2001]. Shapefiles encode the information as mentioned above and are easily read by Python, using libraries such as Geopandas.

3.2 Raster Data

Raster data uses an image-like file, to encode information with metadata about information such as the used projection (see 3.3. Projections for more information). Every pixel, similar to an image encodes information. Raster data comes in two different types, which are handled in the same way: continuous data and discrete data. Continuous data encompasses data such as average temperature, precipitation or elevation maps. The pixel data flows over into each other, while discrete data encompasses for example categorical maps, such as ground type. The best known use case of raster data used in geodata is satellite imagery: pictures are made from a satellite and metadata is encoded to indicate at what coordinate and location the picture was taken. Raster files are extremely useful in other cases besides just satellite imaging, such as encoding temperature, although this

information is usually extracted from satellite images. Raster data is vital for meteorology, disaster management, and industries where analysing risk is essential[9]. A problem that arises with these types of maps is resolution. Resolution is problematic in two ways: (a) enlarging and (b) stretching. These problems are discussed below.

3.2.1 Resolution

The resolution of a map entails the precision on a horizontal level, meaning the spatial accuracy across latitude and longitude. Having a high resolution is beneficial as more precise data on a specific location can be queried from the file. Very high resolution data is hard to come by as the usual resolution of satellites for non-commercial satellites, such as the NASA Landsat or European Copernicus is around 15m, while very high resolution (VHR) commercial satellites achieve a precision of around 1 meter[10]. Usually raster data files have a resolution of 250x250m or 100x100m, such as the flooding maps [Dottori et al., 2016]. *Appendix Features with Resolution* displays all of the resolutions for the different features.

This lower resolution brings two principal problems: (a) imprecision (b) stretching. If two coordinates are close to each other, comparing the two values will give unsatisfactory results, as both coordinates could be in the same pixel or in adjacent pixels, which does not correctly represent the real representation. An example that will be perpetuated throughout this project is the distance to the river. For this project, the river maps are in raster data format and have a resolution of 250x250m. If the distance to the river is only 100 meters, the distance to the river from a coordinate to the river will be 0, as both of the locations will be in the same pixel. For this feature it could be that this imprecision is detrimental. An additional problem is stretching. The pixels are encoded as square 'blocks', which contain data about that area, but because of the projection the data has to undergo, outer points are stretched out, making for an even more skewed representation, besides the already existing imprecision. Resolution is a very important aspect or restriction of raster data to keep in mind and for this reason the resolution is considered for all features and the method of this project. Working with resolutions aside from the available data having a low granularity also has a trade-off with computational power: consider the flood return period maps with a resolution of 250x250m. To cover all of Europe with this resolution, the total map has a total size of 51836*47087 = 2.44E9 pixels. This file, when loaded into memory expands to 10GB RAM. Resolutions of less than 100x100m are thus unrealistic

[9]From the company website of MGiss: `https://mgiss.co.uk/the-different-types-of-gis-data/`

[10]From Maxar Technologies, a commercial satellite imagery provider: `https://explore.maxar.com/Imagery-Leadership-Spatial-Resolution`

with the current setup and are outside of the scope of this project.

Raster data, in conclusion has many uses and can even replace vector data in most cases, although vector data is still preferred for certain types of data. For this project, only raster data was used, because of the ease of use: once helper functions have been created (see Appendix Helpers). An additional argument for this approach is that combining both vector and raster data could be problematic to work with, as could have been the case for the relative height to river, which could combine a vector river map and a raster elevation map.

3.3 Projections

An important part of geodata is that the data needs to be mapped from a sphere onto a two-dimensional surface. A projected coordinate system is defined on a flat, two-dimensional surface. It has constant lengths, angles and areas across the two dimensions. Geographic coordinate systems do not have this, as they are spheres instead of flat, two-dimensional surfaces. As not always the whole sphere is defined, as is the case for solely European data, a specific transformation function, making use of different projection systems needs to be defined to map the spherical data onto the two-dimensional surface. There are many different projected coordinate systems that can be used: a worldwide coordinate system (WGS-84), a European coordinate system (ETRS-89) or even more precise: Berlin (DHDN / Soldner Berlin). All of these projections use the same raster, but start at different origin locations (the top left point or (0,0) in a raster file) and have different curvatures and distances. The general goal of these projections, when working with coordinates is to project the latitude, longitude coordinates from the world globe onto a two-dimensional plane. For this the WGS-84 projection is used. If however one wants to map the flood maps onto a two-dimensional plane, whilst still respecting the WGS-84 coordinates, it needs to be mapped using the ETRS-89 coordinate system. This is a pan-European coordinate system and as this project is based on European data, it is used frequently for raster files. To find the projection that is used for certain files, the following command can be run to get the metadata of the file and thus the projection:

```
gdalinfo filename.tiff
```

If the projection of the file is known, it can also be found on EPSG.io, where all of the different projections are listed. How the projections are converted, is explained in Appendix Helpers.

3.4 Conclusion

In general geodata comes in two different types, which both have their advantages and disadvantages. For this project raster data are used because they are easy and straightforward to work with. This does mean that the disadvantages, such as resolution size play a big role in the uncertainty of this project and constantly considered. Next up is the Method of the project, here the helper functions that were mentioned above are discussed, in addition to the data acquisition, data cleaning, model pipeline and feature importance analysis approach.

4 Method

This chapter will explain the process of this project, from start of finish. This procedure had the following structure:

- **Helper functions**: the functions that are required to read the data and make sense of it
- **Point Data Acquisition**: the data in the KR&A portal were unbalanced and it did not have enough data for the flood return periods, so new points had to be collected.
- **Adding the features**: to these points, features had to be added and the steps taken to get these features is explained in this subsection.
- **Data cleaning**: this data has to be cleaned, such as removing null values and one hot encoding values
- **Models**: how are the models created and how does this system work

This chapter will explain both the process and considerations that were made, such as the way of extracting the data and data source considerations.

4.1 Helper Functions

All of the geodata features are extracted from raster data and to accommodate for a quicker development process down the line, several helper functions were created. Examples of such helper functions are: getting the location of a WGS-84 point on a projected ETRS-89 map or getting the surrounding values around a coordinate. The following helper functions were defined and created and their respective implementation and pseudo code is available in Appendix Helpers.

- get_reference_point: get the point in the raster array from a WGS-84 coordinate

- get_coordinate_from_point: get the WGS-84 coordinate back, from a point in the raster array
- get_pixel_value: get the value from the point in the raster array from a coordinate
- get_dict_pixel_value: get the value from a bulk of points in the raster array from a coordinate
- get_shortest_distance_non_zero_sparse: get the closest_non_zero point in the surrounding area, if there are any
- get_surrounding_values: get the surrounding values and enact analysis on these points

4.2 Point Data Acquisition

All of the coordinates of the locations for this project were retrieved by getting locations in towns and cities with more than 1000 inhabitants in Europe. This acquisition was selected, because it allows for a spread out approach, where points are queried from all across Europe, whilst still retaining their urban relevance [Jacobson, 2011]. To start off, the coordinates of all of the towns with more than 1000 inhabitants in Europe are required and these coordinates were queried from [OpenDataSoft, 2022]. Then a modified version of the helper *get_closest_point* gets the non zero points of the surrounding area for the flood map with a return period of 200 years and if they exist, gets a sample of at most ten values from the area. For all of these coordinates with a value, the values are once more queried for all of the different flood map return periods (10,20,50,100) and stored in a dataframe. The dataset is balanced by first getting all of the values that also flood for 10 years and removing these entries from the dataframe and doing this gradually up to 200 year return period, to get the points that do flood in that return period but not in the next one. The zero values, meaning no flooding are not acquired in this way, as the *get_closest_point* function retrieves the closest non-zero values instead of also the zero points. The zero values are sampled from the KR&A assets dataset to also account for the urban aspect as well as a relative distribution across Europe. The retrieval of these zero points was an incorrect approach and is taken up into the Future Research: Improved Data.

The following step plan is executed to create the dataset:

1. Get all of the cities and towns in Europe above 1000 inhabitants and their respective coordinates.
2. Query the non zero points in the surrounding area of these points using the *get_closest_point* function in the 200 year flood return period map

3. Add all of these points with their corresponding value into a dictionary
4. Convert the dictionary to coordinates
5. Query the values for the rest of the return periods using these coordinates
6. Collect the values in a dataframe
7. Balance the dataframe, to get the same amount of values for the different return periods.
8. Add a sample from the KR&A dataset of the same size to accommodate for the no flooding values, as these zero points are not retrieved from the *get_closest_point function*.

This approach yields a total of 7618 samples per return period and all of these return periods have a flooding_label attached (0,1,2,3,4,5). The geographical distribution of these flooding_labels is displayed in Figure 3. A more comprehensive image per label, including a map of the river system of Europe can be found in the *Appendix Data Distributions per Flooding Label*. The river maps in these images is only added for illustration purposes as it is a shapefile and therefor not part of the scope of this project.

The no flooding coordinates are coloured purple and these are centered around The Netherlands, The Ruhrgebied and Paris. This is because the bulk of the assets that KR&A covers are situated here.

4.3 Adding the features

This subsection briefly covers the different ways these files are added. To complete the dataset before the features, all of the coordinates get their corresponding country code and NUTS3 code using shapefiles to see whether a location is situated inside of such a region. The NUTS3 code labels a smaller region within provinces or countries and allows for statistical analysis across Europe in a normalized manner. A NUTS3 region usually consists of around 150.000 to 800.000 inhabitants. Consequently all of the features are added using the helper functions or a slightly modified version. Only the relative height to the river is discussed in more detail below as it combines certain functions, which need to be explained in a better manner. The other feature addition is demonstrated in Table 3: Retrieve Features.

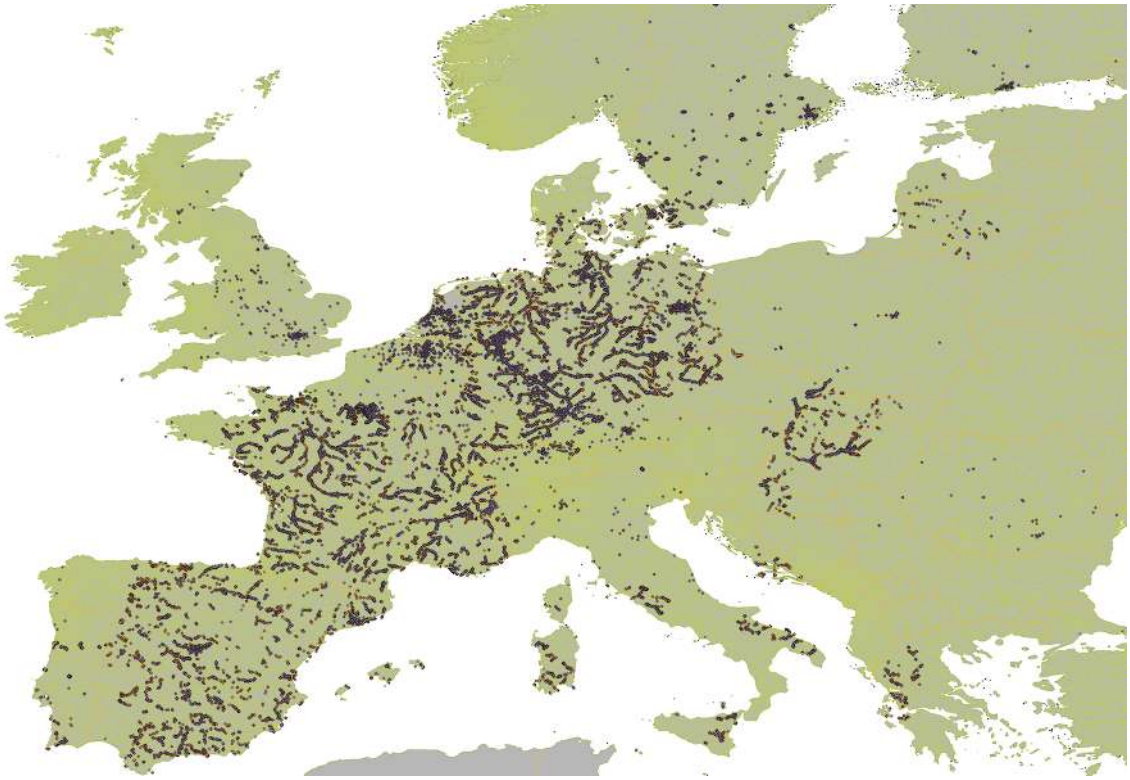


Figure 3: Data Distribution Flooding Values

4.3.1 Relative Height to River

The relative height to river feature is build up of three helper functions: *Get Shortest Distance Non Zero, Get Coordinate from Point* and *Get Pixel Value*. To get the relative height to the closest river, firstly, there has to be a river nearby. If there is, the *Get Shortest Distance Non Zero* helper returns a point from the raster file [Dottori et al., 2016]. This point is then transformed to a coordinate using *Get Coordinate from Point*. For both the coordinate of the data point and the river point, the elevation is retrieved using *Get Pixel Value* on the elevation map [Neteler et al., 2022]. These two values are then compared and the different is then the relative height difference between the point and the closest river.

Feature	Data Source	Added using the following functions
Micro 1-2: Maximum Prec. (One/Five Day Period)	[Copernicus CCS, 2020]	None, nc file query
Micro 3-14: Monthly Prec.	[Fick, 2017]	Get Dict Pixel Value
Micro 15: Ground Type (point)	[EEA, 2003]	Get Dict Pixel Value
Micro 16-17: Ground Type (radius)	[EEA, 2003]	Get surrounding values and max
Micro 18-21: Imperviousness	[Copernicus, 2015]	Get Surrounding Values
Meso 1: Distance to river	[Dottori et al., 2016]	Get Shortest Distance Non Zero
Meso 2-7: Relative Height and Depression	[Neteler et al., 2022]	Get Surrounding Values
Macro 1: GDP	KR&A Portal	Load in for NUTS3
Macro 2: GDP per capita	KR&A Portal	Load in for NUTS3
Macro 3: Govern. Long Term Vision	[World Economic Forum, 2019]	Load in for NUTS3
Macro 4: Quality of overall infrastructure	[World Economic Forum, 2018]	Load in for countries

Table 3: Retrieve Features

4.4 Data Cleaning

After all of the data points are queried and all of the features are retrieved, the data has to be cleaned and one-hot encoded before being fed into the models. The following steps were conducted in this part:

1. Map the null values
2. Fill the null values
3. Change the datatype
4. Change non-sensible data

This section will first talk about the null values and how they are replaced per feature. The second part will talk about other changes to the data such as hot encoding the categorical features.

4.4.1 Null Values

The following table shows the amount of null values per category and the method that was used to remove them.

Feature	N NaNs	Method to remove values
Micro 1: One Day Maximum Prec.	204	Mean of other values
Micro 2: Five Day Maximum Prec.	987	Mean of other values
Micro 3-14: Average Monthly Prec.	1	Mean of other values
Micro 15-17: Ground Type	4	Set to most occurring category
Micro 18-21: Imperviousness	4	Mean of the other values
Meso 1: Distance to river	25891	Fill to maximum distance, as NaN was set if no river was found
Meso 2-4: Relative Height	8	Set to mean of relative height
Meso 5-7: Depression (Height)	8	Set to most occurring category
Meso 8: Relative Height to river	25892	Set to 0
Macro 1-2: GDP and GDP/inh	577	Mean of other values
Macro 3: Gov. Long Term Vision	7	Mean of other values
Macro 4: Quality of overall infra.	19	Mean of the other values

Table 4: Null Values and Methods of replacement

The most common way of filling up the data was taking the mean or getting the most common category. A reasoning for the different features are described in *Appendix Motivation for Null Values*.

4.4.2 Other Changes

The following modifications were implemented to improve the quality of the data:

- Non sensual data was replaced by data that made sense; if a value was too big, it was set to the maximum believable value that the feature could take on.
- Data types were changed where it made sense: strings or integers for discrete features and floats for continuous features.
- The categorical features: ground type (point/1km/5km) *(Micro 15-17)* were one hot encoded for the Neural Network model.

4.4.3 Conclusion

4.4.1 demonstrates that Null values were replaced instead of removed. This was done to maintain the evenly distributed shape of the dataset and because the alteration of the null values are substantiated. This is a choice that was made

during the project and it could be investigated in future research whether this assumption is correct or that these null values should be removed.

4.5 Machine Learning Models

In this project three different algorithms were considered and compared: (a) Logistic Classifier, (b) Random Forest Classifier and (c) Neural Network. The models were trained on three different y_label sets: A Binary model, which indicates whether there will be flooding or not in a period of 20 years, a Quaternary model, which indicates the flooding labels that are used in the KR&A portal: (i) no flooding, (ii) low, (iii) medium, (iv) high and a Senary model, which entails all of the flood return periods. This chapter explains the different classifiers. It is important to also stress the design choice of the Binary model: the return period of 20 years was taken because it is a period that gives insight in long term investments of properties, whilst still retaining its relevance. In addition to this is the maximum precipitation data acquired over a period of 20 years, which fits the return period perfectly.

4.5.1 Logistic Classifier

The Logistic Classifier or LC model in short, was implemented using the Python library sklearn. The data was split up into a training and test set by the sklearn extension train_test_split, with a ratio of 0.66 and the independent variables were separated from the predictors. This yields 30.472 training variables and 15.236 test variables. Using the LogisticRegression from sklearn, the model was fitted and then evaluated, for each of the different data types. The model was evaluated on three metrics: accuracy. Consequently, a feature correlation matrix was extracted to give insights in what features contribute the most to the correct prediction.

4.5.2 Random Forest Classifier

The first machine learning algorithm applied to the data was the Random Forest algorithm. The principal reason for using a random forest algorithm is its explainability: besides being able to correctly predict whether an area will flood from its local features, one want to know which features contribute to this event. A Random Forest is a supervised learning algorithm that makes use of bagging to reduce the variance of a model. Bagging takes random samples from the dataset and uses these to create different models, from which the average is taken. The Random Forest model trains several Decision Trees from different parts of the data and takes the average of these trees as the model. This reduces the variance and also decreases the sensitivity to the data, as outliers are quickly neglected by taking

the average of thousands of different decision Trees. Decision Trees are built by minimizing the sum of squared errors at the nodes of the tree. Because there would be high correlation between the trees if the split data were to create complete decision trees and average them, a Random Forest model requires an additional step: instead of taking all of the features, Random Forests take a subset of the features and use these to generate different Decision Tree, with each a different subset of the features. This way the correlation between the Decision Trees decreases and after all of these trees have been created, the average is taken from these trees and the Random Forest model is trained, just like with regular bagging. Figure 4 shows the first two nodes of a Random Forest. This example shows that the explainability is extremely high and consequently a feature importance matrix can be generated from the trained Random Forest model to get an insight in how the features contribute to the model.

Figure 4: Example from the project of the first nodes of a Decision Tree generated by the RF Classifier

The model was created and fitted using the sklearn DecisionTreeClassifier. The data was split up into a training and test set by the sklearn extension *train_ test_ split*, with a ratio of 0.66 and the independent variables were separated from the predictors. This yields 30.472 training variables and 15.236 test variables. sklearn allows for the creation of a text based decision tree output, which can be used to

see which considerations were made to come to the right answers.

4.5.3 Neural Network

A Neural Network was also implemented to predict whether flooding will occur for all of the different models. Neural Networks are based on the way the human brain works, with nodes and connections between these that transmit data to other nodes, if the activation is high enough. This approach yields a non-linear model, which in its turn can be trained using back-propagation. Explainability, however is problematic with neural networks and feature importance is hard to implement as the network itself makes features from the inputs given to it. One solution is to train the network by continuously removing certain features, but because of the stochastic behaviour of Neural Networks, this does not yield satisfactory results. There are also solutions for this problem, but these are outside of the scope of this project. For this research the neural network is only trained to give an indication on the performance of such a model on this data. Neural networks are build up out of layers, which are connected to each other. This kind of fully dense connected neural network was implemented using Tensorflow and Keras. Keras is a library on top of Tensorflow, which allows for quicker development of neural networks and by leveraging the Sequential function from Keras. The models use the following loss functions and optimizers:

Model	Loss Function	Optimizer
Binary	BinaryCrossentropy	Adam
Quaternary	CategoricalCrossentropy	Adam
Senary	CategoricalCrossentropy	Adam

Table 5: Loss Functions and Optimizers for the Neural Networks of the different models

4.6 Evaluation

The scores for all of the models will be measured in accuracy. This accuracy score was retrieved by both the test set and the validation set, which consists of all of the KR&A assets. This validation set is extremely skewed as can be seen in Figure 5, to the no flooding side. It is however added as an indication of the performance of the model in real world cases and gives insight to the data of the contractor. The distribution of the flooding labels found in the KR&A dataset are displayed in Figure 5. Important to note that the no flooding label is shrunk down ten times to also demonstrate the distribution of the other flooding labels. besides the accuracy of the model is explainability also an important aspect of the models; it is

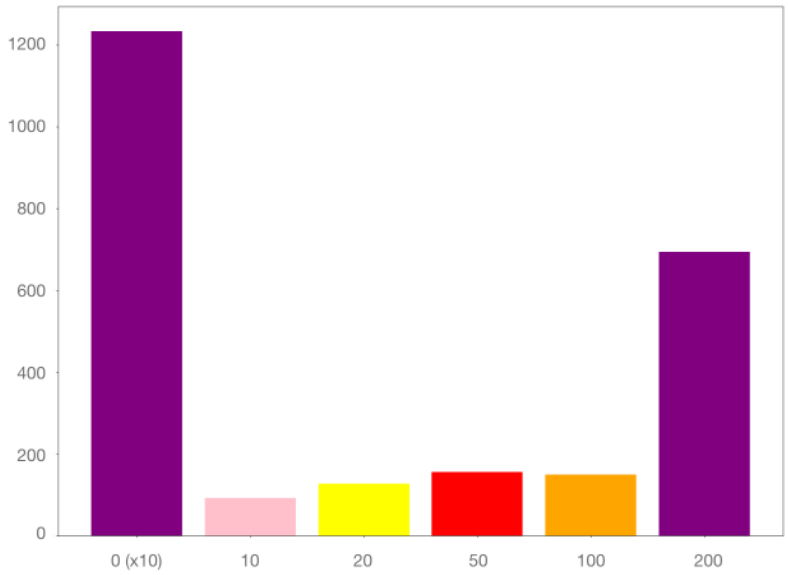


Figure 5: Distribution of the KR&A Validation Set Points over the labels

a contributing factor in selecting which of the models performs the best. Although explainability cannot be quantified, it is possible to state that one model is more explainable than the other.

5 Results

In this section the results of the project are outlined for the three different models: (a) Binary, (b) Quaternary and (c) Senary for the different classifiers: (a) Logistic Classifier, (b) Random Forest and (c) Neural Networks. This section reports the predictive power of the models, including the corresponding feature importance matrix, if applicable. The Senary model results are displayed, but an analysis of these results is moved to the *Appendix Complete Results Senary Model*.

5.1 Binary Model

The binary model returns a prediction on whether an asset will flood or not in a return period of 20 years (0 or 1). The baseline for this model is thus 50%, as a random prediction would be right 50% of the time on an evenly distributed dataset. This subsection presents the results for the different machine learning methods, with a small discussion about what is portrayed in these results. The

results of the Binary model are displayed in Table 6.

Classifier	Test Set Accuracy (%)	KR&A Set Accuracy (%)
Logistic Classifier	91.73	89.99
Random Forest	97.51	97.19
Neural Network	92.06	94.81

Table 6: Accuracies for the Binary Model

5.1.1 Logistic Classifier

An accuracy of 91.73% was found for the test set and an accuracy of 89.99% was found for the KR&A validation set. This is significantly higher than the baseline that was hypothesised for a random prediction. The ten most significant feature importances are displayed in Table 7 in descending order. The relative importance for the remainder of the features is included into the *Appendix All Feature Importances*.

Feature Index	Feature	Coefficient
Micro 19	Imperviousness (500m)	-0.034
Micro 20	Imperviousness (1km)	-0.033
Micro 18	Imperviousness (point)	-0.030
Micro 2	Max. Precipitation 5 Day Period	0.022
Micro 7	Monthly Precipitation (May)	0.017
Micro 21	Imperviousness (5km)	-0.016
Micro 10	Monthly Precipitation (Aug)	-0.015
Micro 9	Monthly Precipitation (Jul)	-.0.009
Micro 4	Monthly Precipitation (Feb)	0.009
Meso 8	Relative Height to River	0.008

Table 7: Feature Importances for Binary Model: Logistic Classifier

Table 7 shows that there is a significant negative relation to the imperviousness features *(Micro 18-21)*, meaning that a higher imperviousness results in a no flooding prediction. Additionally, some of the monthly precipitation features *(Micro 7,9,10)* have a predictive power, but usually cancel each other out. The maximum precipitation in a 5 day period *(Micro 2)* also has a positive relation, meaning that with an increased maximum precipitation, the location is more likely to be subject to flooding.

5.1.2 Random Forest Classifier

The Random Forest Classifier yields an accuracy of 97.51% for the test set and an accuracy of 97.19% for the KR&A validation set. This score far exceeds the baseline and also the logistic classifier model. The highest feature importances of this classifier are displayed in Table 8.

Feature Index	Feature	Importance
Micro 19	Imperviousness (500m)	0.217
Micro 20	Imperviousness (1km)	0.180
Micro 18	Imperviousness (point)	0.135
Micro 21	Imperviousness (5km)	0.092
Macro 2	GDP/inh	0.067
Macro 3	Government Long Term Vision	0.045
Macro 1	GDP	0.033
Micro 9	Monthly Precipitation (Jul)	0.021
Micro 10	Monthly Precipitation (Aug)	0.019
Micro 8	Monthly Precipitation (Jun)	0.016

Table 8: Feature Importances for Binary Model: Random Forest Classifier

Table 8 demonstrates that the imperviousness features *(Micro 18-21)* significantly contribute to the predictive power of this classifier. Other features that contribute to some degree are the GDP features (*Macro 1,2)* and the government long term vision feature *(Macro 3)*. Although these are not as significant as the imperviousness features, they have a combined impact of approximately 15%. The other features are not significant and do not contribute to the performance power in any remarkable way.

5.1.3 Neural Network

The Artificial Neural Network yields an accuracy of 92.06% for the test set and an accuracy of 94.81% for the KR&A validation set. This is lower than the Random Forest Classifier and a bit higher than the Logistic Classifier. The relative feature importance is non-trivial to derive from a neural network, requiring the application of e.g SHAP [Lundberg, 2018], which is out of scope for this thesis. Hence, for this section a Table with feature importance is not presented.

5.2 Quaternary Model

As previously discussed, the Quaternary model returns a prediction on the flooding label of a certain location, as defined by the KR&A portal. KR&A classifies no

flooding as no flooding, low risk as a flood return period of 200 years, medium risk as a flood return period of 50 years and high risk as a flood return period of 10 years. This means that there are four classes in total (thus the name Quaternary). The baseline of this model is thus 25%, as a random prediction would be right 25% of the time. This subsection presents the results for the different machine learning methods, with a small discussion about what is portrayed in these results. Table 9 demonstrates the accuracies of the different classifiers for the Quaternary model.

Classifier	Test Set Accuracy (%)	KR&A Set Accuracy (%)
Logistic Classifier	44.93	83.77
Random Forest	73.81	90.77
Neural Network	46.26	84.43

Table 9: Accuracies for the Quaternary Model

5.2.1 Logistic Classifier

The Logistic Classifier yielded an accuracy of 44.93% for the test set and an accuracy of 83.77% for the KR&A validation set. The latter has a high accuracy as the dataset is rather skewed. This would also indicate that the classifier has a tendency to choose the no flooding label (0). Although this bias seems to exist, the classifier significantly outperforms the baseline of 25%. Table 10 shows the correlations for this model.

Feature Index	Feature	Coefficient
Micro 18	Imperviousness (point)	0.015
Micro 19	Imperviousness (500m)	0.010
Meso 8	Relative Height to River	-0.009
Micro 2	Max. Precipitation 5 Day	-0.009
Micro 20	Imperviousness (1km)	0.008
Micro 9	Monthly Precipitation (Jul)	0.005
Micro 21	Imperviousness (5km)	0.004
Micro 1	Max. Precipitation 1 Day	-0.004
Micro 10	Monthly Precipitation (Aug)	0.004
Macro 2	GDP/inh	0.003

Table 10: Feature Importances for Quaternary Model: Logistic Classifier

The imperviousness features *(Micro 18-21)* yield a big importance to the predictive power of the model, as well as the relative height to river *(Meso 8)* and the maximum precipitation in 5 consecutive days *(Micro 2)*, which yield a negative correlation. The correlations are not as strong as for the Binary model.

5.2.2 Random Forest Classifier

The Random Forest Classifier yields an accuracy of 73.81% for the test set and an accuracy of 90.77% for the KR&A validation set. This score far exceeds the baseline and also the Logistic Classifier. The highest feature importances for this model are summarized in Table 11.

Feature Index	Feature	Importance
Micro 19	Impervousness (500m)	0.083
Micro 20	Imperviousness (1km)	0.077
Micro 21	Imperviousness (5km)	0.050
Micro 18	Imperviousness (point)	0.047
Meso 4	Relative Height (5km)	0.045
Meso 3	Relative Height (1km)	0.040
Macro 2	GDP/inh	0.037
Meso 2	Relative Height (500m)	0.037
Micro 9	Monthly Precipitation (Jul)	0.034
Micro 1	Max. Precipitation 1 Day	0.033

Table 11: Feature Importances for Quaternary Model: Random Forest Classifier

Table 11 shows that the imperviousness features *(Micro 18-21)* significantly contribute to the predictive power of the model. In addition to *Micro 18-21*, the relative height features *(Meso 2-4)* also have a relatively high importance. For the other features, including all of the different average monthly precipitations, the feature importance is low and similar. As one can already see in the top ten features and even better in the complete feature importance list, displayed in *Appendix All Feature Importances*.

5.2.3 Neural Network

The Artificial Neural Network yields an accuracy of 46.26% for the test set and an accuracy of 84.43% for the KR&A validation set. This is significantly lower than the Random Forest Classifier and a bit higher than the Logistic Classifier. The relative feature importance is non-trivial to derive from a neural network, requiring the application of e.g SHAP [Lundberg, 2018], which is out of scope for this thesis. Hence, for this section a Table with feature importance is not presented.

5.3 Senary Model

Due to the fact that the Senary model is not the core focus of this project, an overview is given of only the overall model results in terms of accuracy percentage.

A more extensive analysis, per model, including discussion, is presented in Ṫable 12 shows the final results of the classifiers for the Senary model.

Classifier	Test Set Accuracy (%)	KR&A Set Accuracy (%)
Logistic Classifier	24.54	74.03
Random Forest	58.51	88.99
Neural Network	24.38	47.45

Table 12: Accuracies for the Senary Model

5.4 Comparing the models

To compare all results, an overview is presented in Table 13.

Model	Logistic Classifier (%)	Random Forest (%)	Neural Network
Binary	91.73	97.51	92.06
Quaternary	44.93	73.81	46.26
Senary	24.54	58.51	24.38

Table 13: Model Accuracy's on Test Set

These results show that the Random Forest Classifiers perform the best out of the three classifiers for all of the different models. For all of the models that are able to be investigated using feature importance, it is also observed that all of the imperviousness features *(Micro 18-21)* contribute the most to the predictive power of the models, followed by the relative height features *(Meso 2-4)*. A comparison of the scores between the Logistic Classifier and the Neural Network shows an almost identical picture.

6 Discussion

In this section we discuss the research outcomes with respect to three general categories: (a) performance of the models (b) the quality of the currently implemented features (c) future improvements to the method and the data.

6.1 Performance of the models

The following three aspects are discussed to quantify the performance of the models: (a) the predictive performance of the classifiers (b) the predictive power of the classifiers on the KR&A assets (c) the explainability of the methods.

6.1.1 Predictive performance of the classifiers

The accuracy of the models decreases when more classes are introduced. This is logical, as there are more options the model can choose from. The extend to which the accuracy decreases for the different models is however significant: where the random forest retains a rather high predictability according to the accuracy metric, the other two classifiers' accuracy significantly decrease when more classes are introduced. The fact that the logistic classifier under performs with more than two classes was as expected because of the nature of a logistic classifier that creates a boundary between the classes instead of deconstructing the classification problem step by step. The low performance of the neural network was not hypothesized. However, neural networks work with probabilities while a tree-based classifier works with features and memorizes deterministic rules. This lets tree-based models out perform neural networks.

6.1.2 Predictive power of the classifiers on the KR&A assets

The predictive power of the classifiers on the KR&A assets seems to be strong at first glance, as a high accuracy is retained. However, a bias within this validation dataset, as discussed in Section 4.6, is potentially affecting the performance positively. The results of the KR&A assets are therefor not a good metric to indicate the predictive performance of the models for a pan European analysis, but for the assets in the KR&A portal it is a good representation as future assets are likely to be in these areas of interest.

The KR&A validation set retains its high accuracy, but because the distribution is skewed, this validation set accuracy is not a good metric to compare the models with.

6.1.3 Explainability of the methods

The degree of explainability of the models is apparent when looking at the feature importance matrices: the Logistic Classifier is explainable with its coefficients and the Random Forest Classifier is explainable from the feature importances scores. The Neural Network however has low explainability, because the network tends to create its own features from the inputs, which in its turn makes it hard to quantify the feature importances. The explainability is the same for the Binary, Quaternary and Senary models, where the logistic classifier and random forest classifier are explainable and the neural network is not.

6.2 Importance of the features

Although feature importance of logistic classifier models is possible[11], the Random Forest generated Decision Tree is most commonly used to do a feature importance analysis. All of the feature importances for the different models are displayed in *Appendix All Feature Importances*. This section discusses the importance for all of the features and whether this is according to the hypothesised impacts, described in *Chapter 2: Feature Research*. Because a total of 6 models are trained that explain the features, the Random Forest Classifier for the Binary and Quaternary model are taken as the golden standard and the other models are used to substantiate claims made for the different features.

6.2.1 Micro 1: One day precipitation maximum

Maximum precipitation in a one day period has no significant contribution, with an importance score of 0.0106 and 0.0334 for the Binary and Quaternary Random Forest classifier respectively. This is not in line with the hypothesis, which stated that this would have a significant contribution. A possible explanation for this is the fact that river or flash flooding usually occurs when there is excessive rain for a longer period of time[12].

6.2.2 Micro 2: Five day precipitation maximum

The five day precipitation maximum has no significant contribution with an importance score of 0.0105 and 0.0320 respectively for the Binary and Quaternary RF classifier. This is not in line with the hypothesis.

6.2.3 Micro 3-14: Average monthly precipitation

Because there are 12 average monthly precipitation, the combined contribution is significant with 0.142 and 0.654 respectively for the Binary and Quaternary model. This is because these are 12 features combined. If the average monthly precipitation features are looked at individually, the months June, July and August *(Micro 8-10)* have a moderately high feature importance, ranging between 0.016 and 0.021 for the Binary model and between 0.032 and 0.033 for the Quaternary model. The latter observation is somewhat in line with the hypothesis, which stated that the average monthly precipitation would not contribute in a major

[11]Tutorial on implementing feature importance: `https://machinelearningmastery.com/calculate-feature-importance-with-python/`

[12]Source from US Environmental Protection Agency: `https://www.epa.gov/climate-indicators/climate-change-indicators-heavy-precipitation`

way. This is not the case, but the average precipitation features *Micro 3-14* have a similar importance to the extreme precipitation features *(Micro 1,2)*.

6.2.4 Micro 15-17: Ground Type (point/1km/5km)

The ground type feature does not contribute, with an average importance score of 0.0032 and 0.0125 for the Binary and Quaternary model respectively. This is not in line with the hypothesis which stated that the ground type would be a significant contributing factor as the ground type would give an indication of the imperviousness *(Micro 18-21)* (see *3.6.10*) and thus would contribute to explaining the flooding.

6.2.5 Micro 18-21: Imperviousness (point/500m/1km/5km)

Imperviousness has the leading feature importance score with an average score for the four different features of 0.156 and 0.064 for the Binary and Quaternary model respectively. A possible reason for this incredible score is the fact that imperviousness is only defined for cities, which is where most of the no flooding labels reside (see Figure 3). This feature is however less relevant for the Random Forest classifier of the Quaternary model. This seems to be an outlier as these features are again the leading features for the other classifiers in terms of feature importance. It was hypothesized that this feature would have a major importance, but the significance it yields in for example the Binary model was not hypothesized. Further research has to be conducted on this as a possible source of this problem could have arisen from the method of extracting the no flooding properties; this improved approach is put forward in Section 6.3: Future Research.

6.2.6 Meso 1: Distance to river

The distance to the river has a very low feature importance with a score of 0.007 and 0.023 for the Binary and Quaternary model respectively. This is against the hypothesis, which stated that this feature would have one of the major feature importances. A possible explanation for the result could be the fact that distances to river are imprecise because of the resolution of the river maps (resolution of 250x250m). This resolution makes it hard to get the exact distance to the river for two reasons. Firstly, rivers are seldom 250 meters in width, while the raster image of the rivers does indicate this. Secondly, because of the 250 meter resolution, more precise measurements are not possible. This is especially problematic for points that are located very closely to the river, which in turn increases the variance. A possible solution is put forward in *6.3. Future Work*. Another explanation can be seen in the amount of NaN values for this feature. Because of computational constraints, it was not possible to get the distance to the nearest river if the river

was more than 20 kilometers removed from the location. It should be investigated whether this feature has a bigger impact if the nearest river was defined for all of the locations, as currently these values were filled by defining them as being 20 kilometers removed from the river.

6.2.7 Meso 2-4: Relative Height (500m/1km/5km)

The relative height does not contribute to the model in a significant way with an average importance score of 0.009 and 0.041 for the Binary and Quaternary model respectively. This is not in line with the hypothesis, which stated that this feature would contribute significantly, because water tends to flow to lower points. The reason why this happens could be explained because of the inaccurate height maps, although this is highly unlikely. A possible improved hypothesis could state that this feature would only make sense for certain points to be more prone to higher flooding if the surrounding area floods as well, instead of having a higher risk in general. Instead of the relative height as a local minimum, the distance to the river would make more sense as a feature, as this would signify a global minimum.

6.2.8 Meso 5-7: Depression (500m/1km/5km)

The depression does not contribute anything to the prediction, with an average importance score of 0.001 and 0.0052 for the Binary and Quaternary model respectively. This is not in line with the hypothesis, as an indication of a depression would mean that from the local surrounding areas, water would flow into this local minimum. A case could be made that these features are a weaker version of the features *Meso 2-4*: relative height. The results, although rather weak do show that it is not the boolean value: depression that matters, but rather the relative height difference. This feature has therefor explained the relaive height features *Meso 2-4* in greater detail and showed that the absolute height difference between the two points is of bigger importance than whether the difference is negative or positive.

6.2.9 Meso 8: Relative height to river

The relative height to the closest river has a low feature importance with a score of 0.007 and 0.020 for the Binary and Quaternary model respectively. This is against the hypothesis, which stated that this feature would be the biggest contributor to the explainability of the model, because a point cannot flood if the point is relatively high above the river. A possible reason why this feature does not have the intended effect is the quality of the data. In *3.2.1*, resolutions are discussed, which are a limiting factor, especially if the distances to the nearest river are relatively low. With the elevation map having a resolution of 100m and the river

map 250m, assets that are relatively close will have a very similar height, whilst from more precise elevation maps, the opposite would be true. This could be a possible explanation to why this feature is unreliable.

6.2.10 Macro 1: GDP

The GDP feature has an average feature importance with 0.033 and 0.031 for the Binary and Quaternary models respectively. The hypothesis for this feature stated that *Macro 1* would have not have a high impact although this feature allegedly indirectly correlated to the amount of investments in flood mitigation architecture. This is the case and it could indicate that the correlation between this feature and the Fluvial Flood Mitigation Architecture factor does not exist. However, further research should indicate whether this is indeed not possible to do.

6.2.11 Macro 2: GDP per capita

The GDP feature has an above average feature importance with 0.067 and 0.037 for the Binary and Quaternary models respectively. The hypothesis for this feature stated that *Macro 2* would have not have a high impact although this feature allegedly indirectly correlated to the amount of investments in flood mitigation architecture. The other classifiers support the average feature importance of the Quaternary Random Forest model, which means that the same reasoning as for the GDP feature *Macro 1* can be used.

6.2.12 Macro 3: Government Long Term Vision

The government long term vision has a below average feature importance with a score of 0.046 and 0.018 for the Binary and Quaternary models respectively. This is in line with the hypothesis, which stated that this feature would not have a significant impact.

6.2.13 Macro 4: Quality of overall infrastructure

The quality of the overall infrastructure has a below average feature importance, with an importance score of 0.014 and 0.012 for the Binary and Quaternary model respectively. This is in line with the hypothesis, which stated that the quality of the overall infrastructure was not localised enough to have a significant impact, as developments to water works has to be constructed on a regional level to have an impact.

6.3 Future Work

This section presents proposed improvements to the current approach and give other possible additions to improve the data quality and gain better insights in the features that contribute to the prediction. The workflow of the project does not need an improvement. To reiterate the process, which consists of four steps: (i) selecting the features, (ii) finding data sources to create these and (iii) cleaning the data and finally (iv) training models on the data. There are however three parts that could be improved upon: (a) improved data (b) adding new investigated features and (c) moving from flooding simulations to historical flooding data. These three parts will be zoomed in on in the next three subsections.

6.3.1 Improved Data

For the collection of data, three improvement points are indicated: (a) overcoming resolution constraints of geographic raster data, (b) selection of no flooding data points, (c) alternative datasource for the ground soil features *(Micro 15-17)*.

Overcoming Resolution Contraints
There are some problems with the collection of the data. A recurring problem is the resolution of certain geographic raster data. Most importantly, the river maps. A possible future improvement is to better accommodate for the river properties features using shapefiles of the river maps instead of the raster images, that were used in this project [Theobald, 2001]. This would make it possible to get an infinitely more precise estimation of the distance from the location to the middle of the river. A downside of this method which needs to be addressed is the fact that the lines representing the river are one-dimensional and thus have no width. However, to estimate a river's width up to a resolution of 1 meter, satellite images could be used (see section 3.1.1 Resolutions for more information). This improvement would yield a better distance to river feature *(Meso 1)*, but would also improve the relative height to the river *(Meso 8)*. The current problem with this feature, as described in *6.2.11* is that the resolution of the river maps is 250x250m, which means that the location of the nearest river is not exact and thus the query to the elevation map is also incorrect. Better river maps would solve this.

Selection of No Flooding Data Points
Another possible future improvement is a better selection of no flooding points, preferably from the surrounding areas of the other flooding labels. Currently, the no flooding points are centred around big metropolitan areas such as Paris and the Rurhgebied and do not share a lot of the features that the other points have. This may have skewed the results, especially the role of the imperviousness *(Micro*

18-21) on the accuracy. Due to time constraints, it was not possible to implement a new method of retrieving surrounding no flooding values and is thus proposed for future research.

Alternative Data Source for ground soil features
Another feature that could have an improved origin of the data is the ground type features *(Micro 15-17)* and its ground type categories. The *Appendix Ground Types* shows that the categories also include tree cover, which is not the ground type itself, but rather the observation from a satellite. Real ground data could give a better idea on how impervious the ground actually is.

6.3.2 Adding new investigated features

In the feature research of this project several features were identified that were not able to be implemented either through time constraints or due to processed data not being available for an implementation. These are the following features:

- Width of nearest river
- Average river height
- Change in cross sectional area of the nearby river
- Distance to the origin of the river
- Curvedness of the river

All of these features can be implemented using satellite imagery and using this satellite imagery to get data about the river. Using computer vision, rivers can be identified and more precise river maps can be created. By implementing these features using the above mentioned approach, river properties are incorporated which are one of the most contributing feature categories according to[13].

6.3.3 Changing y_labels to historical flooding data

As previously discussed, the assumption was made that the flood simulations [Dottori et al., 2016] are mapped 1:1 to historical flood return periods. This is not exactly the case however and a possible solution to this inconsistency could be to use historical flooding maps to use real world *y_label* data in a certain time frame. Alternatively, to using historical flooding maps, the use of further improved flooding simulation models can also lead to more representative results.

[13]From Lecture Slides: `https://www2.tulane.edu/~sanelson/Natural_Disasters/riversystems.htm`

7 Conclusion

The research conducted in this project first investigated several academic sources that indicated such possible local features. Secondly, the data for these features, where possible was retrieved and grouped. This data was used to train three different classification models: logistic classifier, random forest classifier and a neural network. This research was conducted to answer the research question: Can risk of flooding be explained by local values of specific locations? To answer this research question, the following questions had to be answered. Firstly, *Can local features be used to assess a risk of flooding with sufficient accuracy?* Yes, both an answer on the question on whether a location will flood based on the local features as which flooding risk label, defined by KR&A can be predicted to a sufficient accuracy. Secondly, *which features contribute the most to predicting whether an asset will flood or not?*. According to this project, the imperviousness *(Micro 18-21)*, relative height *(Meso 2-4)* and monthly precipitation in July and August *(Micro 8-9)* play a major role in the prediction, yielding a high feature importance weight. To answer the main research question, according to these sub answers: The risk of flooding can be explained to some degree, whilst predicting whether flooding occurs at that location is certainly possible, with the best performing classifier having an accuracy of 97.51%. However there were some problematic facets about the dataset. These are described above and have to be tackled in future research to come to a more definite answer. Most importantly, the manner of acquisition of the no flooding labels should be done more carefully. It is however important to address the main assumption of this project one more time, that the flooding simulation corresponds 1:1 with historical flooding return maps, which is not the case. Although predicting flooding is not an exact science and historical occurrences do not entail future outcome, it is possible to predict flooding with the help of local features of a location. In conclusion we can say that the research question was answered and that local features of locations are able to make predictions on whether a location will flood. More difficult, however is to give an answer to the degree of risk that exists on these places, such as the labels that were defined in the KR&A portal.

References

[Bankoff and Lee, 1983] Bankoff, S. G. and Lee, S. C. (1983). Critical review of the flooding literature.

[Bos and Zwaneveld, 2017] Bos, F. and Zwaneveld, P. (2017). Cost-benefit analysis for flood risk management and water governance in the netherlands: An overview of one century. *CPB Background Document | August*.

[Butler et al., 2018] Butler, D., Digman, C., Makropoulos, C., and Davies, J. W. (2018). *Urban drainage*. Crc Press.

[Chen et al., 2005] Chen, A., Hsu, M., Chen, T., and Chang, T. (2005). An integrated inundation model for highly developed urban areas. *Water science and technology*, 51(2):221–229.

[Copernicus, 2015] Copernicus (2015). Imperviousness dataset. Data retrieved from: `https://land.copernicus.eu/pan-european/high-resolution-layers/imperviousness`.

[Copernicus CCS, 2020] Copernicus CCS (2020). Extreme precipitation risk indicators for europe and european cities from 1950 to 2019. Data retrieved from: `redacted`.

[Dottori et al., 2016] Dottori, F., Salamon, P., Bianchi, A., Alfieri, L., Hirpa, F. A., and Feyen, L. (2016). Development and evaluation of a framework for global flood hazard mapping. *Advances in water resources*, 94:87–102.

[EEA, 2003] EEA (2003). Soil type dataset. Data retrieved from: `https://www.eea.europa.eu/data-and-maps/data/soil-type`.

[Fick, 2017] Fick, S.E., R. H. (2017). Worldclim 2: new 1km spatial resolution climate surfaces for global land areas. *International Journal of Climatology 37 (12)*, pages 4302–4315. Data retrieved from: `https://www.worldclim.org/data/worldclim21.html`.

[Gholami et al., 2010] Gholami, V., Mohseni Saravi, M., and Ahmadi, H. (2010). Effects of impervious surfaces and urban development on runoff generation and flood hazard in the hajighoshan watershed. *Caspian journal of environmental sciences*, 8(1):1–12.

[Jacobson, 2011] Jacobson, C. R. (2011). Identification and quantification of the hydrological impacts of imperviousness in urban catchments: A review. *Journal of environmental management*, 92(6):1438–1448.

[Ladson, 2019] Ladson, A. (2019). Using wsud to restore predevelopment hydrology. In *Approaches to Water Sensitive Urban Design*, pages 209–228. Elsevier.

[Langanke, 2021] Langanke, T. (2021). Imperviousness and imperviousness change in europe.

[Lehner and Grill, 2013] Lehner, B. and Grill, G. (2013). Global river hydrography and network routing: baseline data and new approaches to study the world's large river systems. *Hydrological Processes*, 27(15):2171–2186.

[Lundberg, 2018] Lundberg, S. (2018). Shap (shapley additive explanations).

[Muis et al., 2020] Muis, S., Apecechea, M. I., Dullaart, J., de Lima Rego, J., Madsen, K. S., Su, J., Yan, K., and Verlaan, M. (2020). A high-resolution global dataset of extreme sea levels, tides, and storm surges, including future projections. *Frontiers in Marine Science*, 7:263.

[Neteler et al., 2022] Neteler, M., Haas, J., and Metz, M. (2022). Copernicus Digital Elevation Model (DEM) for Europe at 100 meter resolution (EU-LAEA).

[OpenDataSoft, 2022] OpenDataSoft (2022). All Cities with a Population > 1000. Data retrieved from `https://public.opendatasoft.com/explore/dataset/geonames-all-cities-with-a-population-1000/information/?disjunctive.cou_name_en&sort=name`.

[Osouli et al., 2017] Osouli, A., Bloorchian, A. A., Grinter, M., Alborzi, A., Marlow, S. L., Ahiablame, L., and Zhou, J. (2017). Performance and cost perspective in selecting bmps for linear projects. *Water*, 9(5):302.

[Sohn et al., 2020] Sohn, W., Kim, J.-H., Li, M.-H., Brown, R. D., and Jaber, F. H. (2020). How does increasing impervious surfaces affect urban flooding in response to climate variability? *Ecological Indicators*, 118:106774.

[Tellman et al., 2021] Tellman, B., Sullivan, J., Kuhn, C., Kettner, A., Doyle, C., Brakenridge, G., Erickson, T., and Slayback, D. (2021). Satellite imaging reveals increased proportion of population exposed to floods. *Nature*, 596(7870):80–86.

[Theobald, 2001] Theobald, D. M. (2001). Understanding topology and shapefiles. *ArcUser Online*.

[Titley et al., 2021] Titley, H. A., Cloke, H. L., Harrigan, S., Pappenberger, F., Prudhomme, C., Robbins, J., Stephens, E., and Zsótér, E. (2021). Key factors influencing the severity of fluvial flood hazard from tropical cyclones. *Journal of Hydrometeorology*, 22(7):1801–1817.

[Verol et al., 2019] Verol, A. P., Battemarco, B. P., Merlo, M. L., Machado, A. C. M., Haddad, A. N., and Miguez, M. G. (2019). The urban river restoration index (urrix)-a supportive tool to assess fluvial environment improvement in urban flood control projects. *Journal of Cleaner Production*, 239:118058.

[Ward and Winsemius, 2018] Ward, P. J. and Winsemius, H. (2018). *River Flood Risk*. PBL Netherlands Environmental Assessment Agency.

[World Economic Forum, 2018] World Economic Forum (2018). Competitiveness rankings. Data retrieved from: https://reports.weforum.org/global-competitiveness-index-2017-2018/competitiveness-rankings.

[World Economic Forum, 2019] World Economic Forum (2019). Gci 4.0: Government long term vision. Data retrieved from https://govdata360.worldbank.org/indicators/h5b74eef6.

[Xu et al., 2017] Xu, X., Wang, Y.-C., Kalcic, M., Muenich, R., Yang, Y.-C., and Scavia, D. (2017). Evaluating the impact of climate change on fluvial flood risk in a mixed-use watershed. *Environmental Modelling & Software*, 122.

Appendix

A Ground Types

Category Index	GLC Global Class (According to LCCS terminology)
1	Tree Cover, broadleaved, evergreen >15% tree cover
2	Tree Cover, broadleaved, deciduous, closed
3	Tree Cover, broadleaved, deciduous, open (15-40% tree cover)
4	Tree Cover, needle-leaved, evergreen
5	Tree Cover, needle-leaved, decidious
6	Tree Cover, mixed leaf type
7	Tree Cover, regularly flooded, fresh water (& brackish)
8	Tree Cover, regularly flooded, saline water, (daily variation of water level)
9	Mosaic: Tree cover / Other natural vegetation
10	Tree Cover, burnt
11	Shrub Cover, closed-open, evergreen (Examples of sub-classes at reg. level *: (i) sparse tree layer)
12	Shrub Cover, closed-open, deciduous (Examples of sub-classes at reg. level *: (i) sparse tree layer)
13	Herbaceous Cover, closed-open (Examples of sub-classes at regional level *: (i) natural, (ii) pasture, (iii) sparse trees or shrubs)
14	Sparse Herbaceous or sparse Shrub Cover
15	Regularly flooded Shrub and/or Herbaceous Cover
16	Cultivated and managed areas
17	Mosaic: Cropland / Tree Cover / Other natural vegetation
18	Mosaic: Cropland / Shrub or Grass Cover
19	Bare Areas
20	Water Bodies (natural & artificial)
21	Snow and Ice (natural & artificial)
22	Artificial surfaces and associated areas

Table 14: A List of the Ground Types

B Data Distributions per Flooding Label

This section provides all of the data distributions per flooding label.

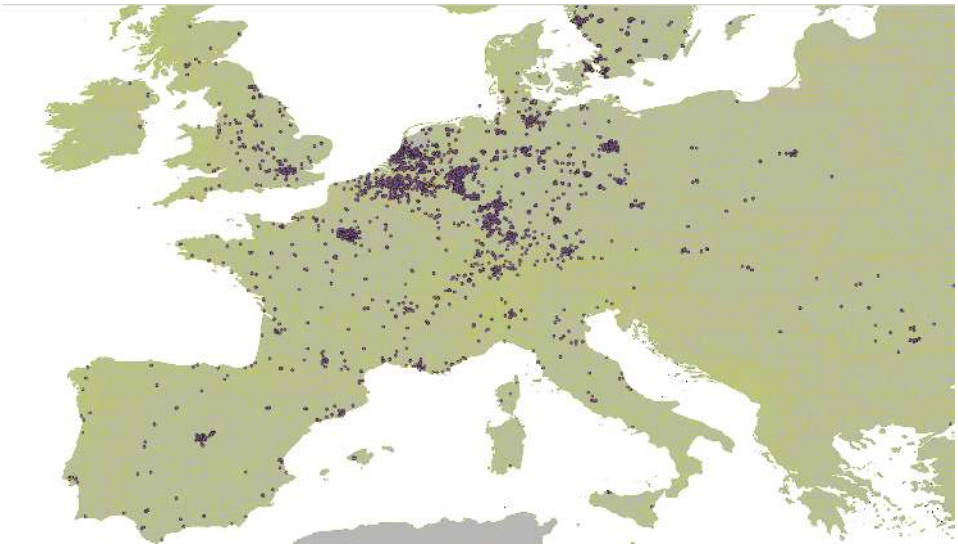


Figure 6: Data Distribution Flooding Value: 0

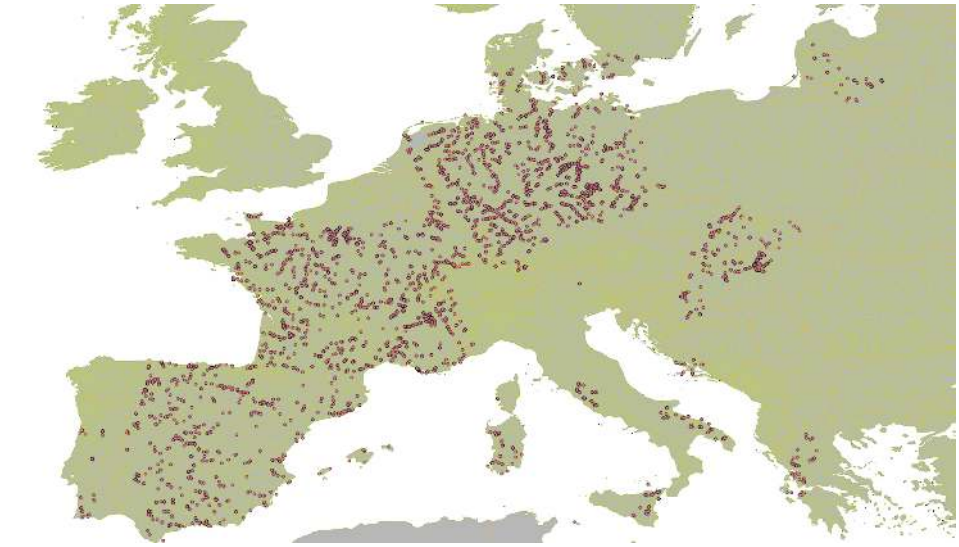


Figure 7: Data Distribution Flooding Value: 1

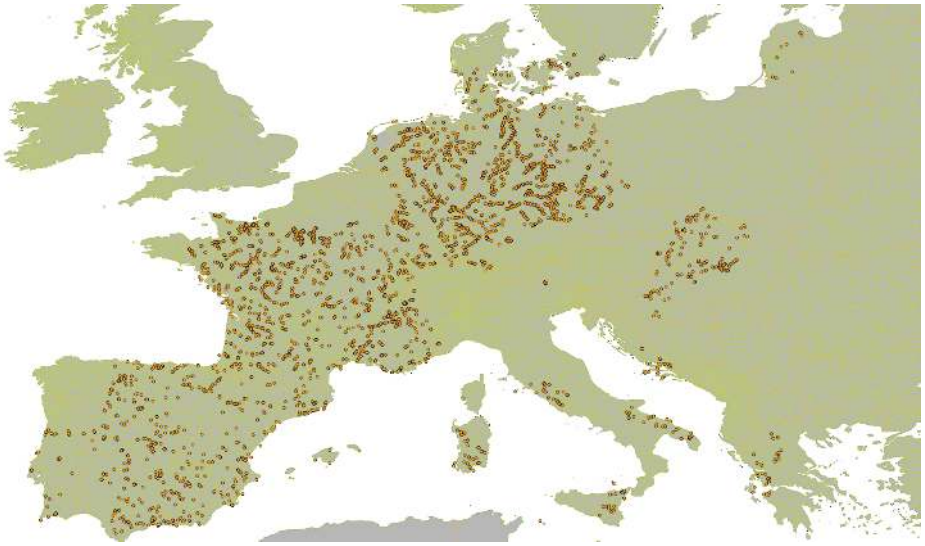


Figure 8: Data Distribution Flooding Value: 2

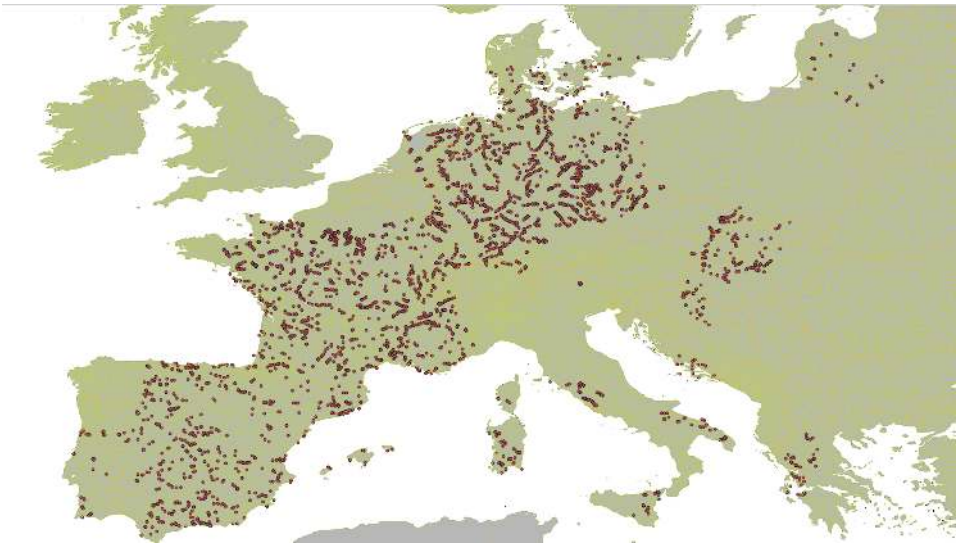


Figure 9: Data Distribution Flooding Value: 3

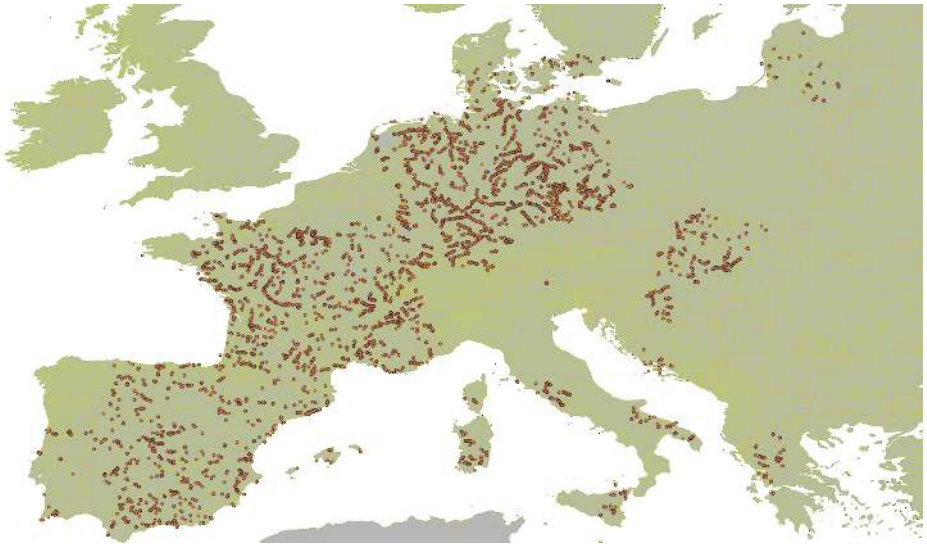


Figure 10: Data Distribution Flooding Value: 4

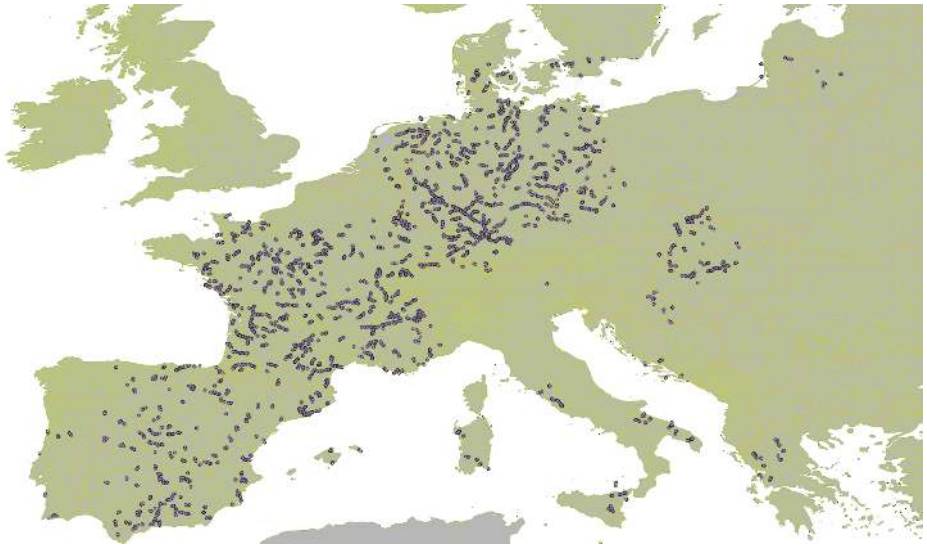


Figure 11: Data Distribution Flooding Value: 5

C Features with Resolution

Feature Index	Feature	Spatial Resolution
Micro 1	One day Maximum Prec.	2000x2000m
Micro 2	Five Day Maximum Prec.	2000x2000m
Micro 3-14	Average Monthly Prec.	1000x1000m
Micro 15-17	Ground Type	250x250m
Meso 1	Distance to River	250x250m
Meso 2-7	Relative Height and Depression	100x100m
Meso 8	Relative Height to River	250x250m
Macro 1	GDP	NUTS3 level
Macro 2	GDP per capita	NUTS3 level
Macro 3	Govern. Long Term Vision	Country level
Macro 4	Quality of overall infrastructure	Country level

Table 15: Features with their corresponding resolution

D Helpers

These are all of the helpers that were used in this project:

D.1 Get Reference Point (c2p)

This helper function uses the projection for the specific data to project WGS84 coordinates to the mapped data. The following pseudocode was used for this helper:

```
a = get the spatial reference of the global coordinate system
b = get the spatial reference of the projected coordinate system (wkt)
tf_c2p = osr.CoordinateTransformation(b,a)
point = tf_c2p.TransformPoint(coord[0], coord[1])
return point[0], point[1]
```

This function requires three properties: (coord, fp, wkt) the coordinate, the file location and the projection string. As explained in Chapter 3.3, the projection string can be found on EPSG.io. This helper is used in almost all of the different

other helpers and is thus indispensable. Note that c2p is the name of the function in the project, but a more natural title is chosen for the explanation of this function. The pseudocode above shows that the projected coordinate system is compared to the global coordinate system and a point is then transformed according to this transformation and returned.

D.2 Get Coordinate from Point (p2c)

This helper function uses the projection for the specific data to project points from the mapped data to WGS84 coordinates. The following pseudocode was used for this helper:

```
a = get the spatial reference of the global coordinate system
b = get the spatial reference of the projected coordinate system (wkt)
tf_p2c = osr.CoordinateTransformation(a,b)
coord = tf_p2c.TransformPoint(point[0], point[1])
return coord[0], coord[1]
```

This function requires three properties: (point, fp, wkt) the point, the file location and the projection string. As explained in Chapter 3.3, the projection string can be found on EPSG.io. This helper is used in almost all of the different other helpers and is thus indispensable. Note that p2c is the name of the function in the project, but a more natural title is chosen for the explanation of this function. The pseudocode above shows that the projected coordinate system is compared to the global coordinate system and a coordinate is then transformed according to this transformation and returned.

D.3 Get pixel value

This helper function uses the previously defined c2p function and returns the value from that pixel. The pseudocode for this function is as follows:

```
px, py = c2p
ds = gdal.Open(fp)
origin, pixel_size = ds.GetGeoTransform
new_px = (px-origin[0])/pixel_size[0]
new_py = (py-origin[1])/pixel_size[1]
data_array = ds.GetRasterBand(1).ReadAsArray()
x = data_array[new_px, new_py]
return x
```

The function first gets the corresponding points for the coordinate and then loads in the geodata file (a raster file). Then it queries the points from this array and returns this.

D.4 Get Dict Pixel Value

Because most of the time, one wants to retrieve all of the values for all of the coordinates from a map, a function that does the above thing in bulk is preferred. This yields the following pseudocode:

```
origin, pixel_size = ds.GetGeoTransform
data_array = ds.GetRasterBand(1).ReadAsArray()
point_list = dict()
for coord in coords:
    px, py = c2p
    new_px = (px-origin[0])/pixel_size[0]
    new_py = (py-origin[1])/pixel_size[1]
    x = data_array[new_px, new_py]
    point_list[coord] = x
return point_list
```

Here the map is loaded in first and then for all of the coordinates the points are retrieved, which in turn are queried from the map and then added to a dictionary. This dictionary is returned. Important is that the coordinates should have a unique id attached to them, so that it is easier to map these values to, for example a dataframe with all of the other features.

D.5 Get Shortest Distance Non Zero (Sparse)

This helper retrieves all of the surrounding values that are non zero in a certain area. The pseudo code is defined as follows:

```
long_dist = sqrt(distance^2)
short_dist = long_dist
x_search = point-long_dist:point + long_dist
y_search = point-long_dist:point + long_dist
search_field = grid[x_search, y_search]
non_zero_vals = nonzero(search_field)
for elem in non_zero_vals:
    dist = get_distance()
    if dist < short_dist:
        short_dist = dist
    return short_dist
```

The grid defined here is the complete loaded in data_array of the map. From this map, a subarea is queried, which is then queried to see whether there are non zero values. If these exists, a loop tries to find the closest point and this distance is

returned. This function is the basic to functions that return which point this is and its location and extra additions. One can also see that the furthest points, such as the top left are further removed from the point than the distance allows. This does not matter in this case, as the distance found is then a bit longer than the allowed threshold, which does not bother the research. This would be a problem for the next helper, which will be explained below.

D.6 Get Surrounding Values

This helper gets the surrounding values and returns the following statistics: whether the coordinate is higher or lower than the surrounding values, what the relative height is between the points and the coordinate. The pseudo code is defined as follows:

```
py, px = c2p
ds = gdal.Open(fp)
origin, pixel_size = ds.GetGeoTransform
new_px = (px-origin[0])/pixel_size[0]
new_py = (py-origin[1])/pixel_size[1]
horizontal = new_px-distance:new_px+distance+1
vertical = new_py-distance:new_py+distance+1
area = data_array[horizontal, vertical]
middle = math.ceil(area.shape[0]/2)
real_area = get_real_surrounding_area(area, middle_point)
average = mean(real_area)
if average > 0:
    depression = 1
else:
    depression = 0
return average, depression
```

The area is not taken as a whole, but rather only a part is retrieved, using the *get_real_surrounding_area* function. This function gets only the points that are maximally removed by the distance diagonally. This is done to make sure that only the data surrounding the point in a circle is retrieved.

E All Feature Importances

Feature Index	Feature	Correlation
Micro 1	Max. Precipitation 1 Day	0.00628
Micro 2	Max. Precipitation 5 Days	0.01868
Micro 3	Monthly Precipitation (Jan)	0.00347
Micro 4	Monthly Precipitation (Feb)	0.00627
Micro 5	Monthly Precipitation (Mar)	-0.00152
Micro 6	Monthly Precipitation (Apr)	0.00464
Micro 7	Monthly Precipitation (May)	0.00829
Micro 8	Monthly Precipitation (Jun)	-0.00683
Micro 9	Monthly Precipitation (Jul)	-0.01093
Micro 10	Monthly Precipitation (Aug)	-0.01024
Micro 11	Monthly Precipitation (Sep)	-0.00429
Micro 12	Monthly Precipitation (Okt)	0.00260
Micro 13	Monthly Precipitation (Nov)	0.00214
Micro 14	Monthly Precipitation (Dec)	0.00339
Micro 15	Ground Type (point)	0.00313
Micro 16	Ground Type (1km)	0.00321
Micro 17	Ground Type (5km)	0.00312
Micro 18	Imperviousness (point)	-0.03264
Micro 19	Imperviousness (500m)	-0.03005
Micro 20	Imperviousness (1km)	-0.02451
Micro 21	Imperviousness (5km)	-0.01161
Meso 1	Distance to River	0.00003
Meso 2	Relative height (500m)	-0.00108
Meso 3	Relative height (1km)	-0.00108
Meso 4	Relative height (5km)	-0.00263
Meso 5	Depression (500m)	0.00006
Meso 6	Depression (1km)	0.00012
Meso 7	Depression (5km)	0.00012
Meso 8	Relative Height to River	0.01515
Macro 1	GDP	-0.00001
Macro 2	GDP/capita	0.00049
Macro 3	Government L.T. Vision	0.00014
Macro 4	Quality of Infrastructure	0.00113

Table 16: Binary Model Logistic Classifier Correlations

Feature Index	Feature	Importance
Micro 1	Max. Precipitation 1 Day	0.0106
Micro 2	Max. Precipitation 5 Days	0.0105
Micro 3	Monthly Precipitation (Jan)	0.0093
Micro 4	Monthly Precipitation (Feb)	0.0141
Micro 5	Monthly Precipitation (Mar)	0.0075
Micro 6	Monthly Precipitation (Apr)	0.0096
Micro 7	Monthly Precipitation (May)	0.0123
Micro 8	Monthly Precipitation (Jun)	0.0160
Micro 9	Monthly Precipitation (Jul)	0.0212
Micro 10	Monthly Precipitation (Aug)	0.0189
Micro 11	Monthly Precipitation (Sep)	0.0088
Micro 12	Monthly Precipitation (Okt)	0.0089
Micro 13	Monthly Precipitation (Nov)	0.0073
Micro 14	Monthly Precipitation (Dec)	0.0076
Micro 15	Ground Type (point)	0.0034
Micro 16	Ground Type (1km)	0.0031
Micro 17	Ground Type (5km)	0.0029
Micro 18	Imperviousness (point)	0.1346
Micro 19	Imperviousness (500m)	0.2179
Micro 20	Imperviousness (1km)	0.1795
Micro 21	Imperviousness (5km)	0.0916
Meso 1	Distance to River	0.0069
Meso 2	Relative height (500m)	0.0087
Meso 3	Relative height (1km)	0.0082
Meso 4	Relative height (5km)	0.0114
Meso 5	Depression (500m)	0.0009
Meso 6	Depression (1km)	0.0008
Meso 7	Depression (5km)	0.0009
Meso 8	Relative Height to River	0.0067
Macro 1	GDP	0.0330
Macro 2	GDP/capita	0.0669
Macro 3	Government L.T. Vision	0.0455
Macro 4	Quality of Infrastructure	0.0144

Table 17: Binary Model Random Forest Classifier Feature Importances

Feature Index	Feature	Correlation
Micro 1	Max. Precipitation 1 Day	-0.4027
Micro 2	Max. Precipitation 5 Days	-0.8923
Micro 3	Monthly Precipitation (Jan)	-0.0468
Micro 4	Monthly Precipitation (Feb)	-0.1156
Micro 5	Monthly Precipitation (Mar)	0.1063
Micro 6	Monthly Precipitation (Apr)	-0.0880
Micro 7	Monthly Precipitation (May)	-0.1480
Micro 8	Monthly Precipitation (Jun)	0.2666
Micro 9	Monthly Precipitation (Jul)	0.4914
Micro 10	Monthly Precipitation (Aug)	0.3800
Micro 11	Monthly Precipitation (Sep)	0.2152
Micro 12	Monthly Precipitation (Okt)	-0.0467
Micro 13	Monthly Precipitation (Nov)	-0.0235
Micro 14	Monthly Precipitation (Dec)	-0.0573
Micro 15	Ground Type (point)	-0.0806
Micro 16	Ground Type (1km)	-0.0788
Micro 17	Ground Type (5km)	-0.0799
Micro 18	Imperviousness (point)	1.4776
Micro 19	Imperviousness (500m)	1.0306
Micro 20	Imperviousness (1km)	0.8399
Micro 21	Imperviousness (5km)	0.4066
Meso 1	Distance to River	-0.0082
Meso 2	Relative height (500m)	-0.1995
Meso 3	Relative height (1km)	-0.0989
Meso 4	Relative height (5km)	-0.2859
Meso 5	Depression (500m)	-0.0011
Meso 6	Depression (1km)	-0.0016
Meso 7	Depression (5km)	-0.0003
Meso 8	Relative Height to River	-0.9232
Macro 1	GDP	0.0006
Macro 2	GDP/capita	0.3042
Macro 3	Government L.T. Vision	0.0030
Macro 4	Quality of Infrastructure	0.0277

Table 18: Quaternary Model Logistic Classifier Correlations

Feature Index	Feature	Importance
Micro 1	Max. Precipitation 1 Day	0.0334
Micro 2	Max. Precipitation 5 Days	0.0320
Micro 3	Monthly Precipitation (Jan)	0.0286
Micro 4	Monthly Precipitation (Feb)	0.0283
Micro 5	Monthly Precipitation (Mar)	0.0278
Micro 6	Monthly Precipitation (Apr)	0.0297
Micro 7	Monthly Precipitation (May)	0.0328
Micro 8	Monthly Precipitation (Jun)	0.0323
Micro 9	Monthly Precipitation (Jul)	0.0339
Micro 10	Monthly Precipitation (Aug)	0.0327
Micro 11	Monthly Precipitation (Sep)	0.0289
Micro 12	Monthly Precipitation (Okt)	0.0298
Micro 13	Monthly Precipitation (Nov)	0.0293
Micro 14	Monthly Precipitation (Dec)	0.0287
Micro 15	Ground Type (point)	0.0136
Micro 16	Ground Type (1km)	0.0130
Micro 17	Ground Type (5km)	0.0100
Micro 18	Imperviousness (point)	0.0473
Micro 19	Imperviousness (500m)	0.0832
Micro 20	Imperviousness (1km)	0.0775
Micro 21	Imperviousness (5km)	0.0499
Meso 1	Distance to River	0.0226
Meso 2	Relative height (500m)	0.0368
Meso 3	Relative height (1km)	0.0398
Meso 4	Relative height (5km)	0.0455
Meso 5	Depression (500m)	0.0053
Meso 6	Depression (1km)	0.0051
Meso 7	Depression (5km)	0.0050
Meso 8	Relative Height to River	0.0199
Macro 1	GDP	0.0308
Macro 2	GDP/capita	0.0369
Macro 3	Government L.T. Vision	0.0178
Macro 4	Quality of Infrastructure	0.0121

Table 19: Quaternary Model Random Forest Classifier Feature Importances

Feature Index	Feature	Correlation
Micro 1	Max. Precipitation 1 Day	-0.0008
Micro 2	Max. Precipitation 5 Days	-0.0020
Micro 3	Monthly Precipitation (Jan)	0.0007
Micro 4	Monthly Precipitation (Feb)	0.0004
Micro 5	Monthly Precipitation (Mar)	0.0012
Micro 6	Monthly Precipitation (Apr)	0.0005
Micro 7	Monthly Precipitation (May)	0.0004
Micro 8	Monthly Precipitation (Jun)	0.0020
Micro 9	Monthly Precipitation (Jul)	0.0027
Micro 10	Monthly Precipitation (Aug)	0.0023
Micro 11	Monthly Precipitation (Sep)	0.0017
Micro 12	Monthly Precipitation (Okt)	0.0008
Micro 13	Monthly Precipitation (Nov)	0.0009
Micro 14	Monthly Precipitation (Dec)	0.0007
Micro 15	Ground Type (point)	-0.0002
Micro 16	Ground Type (1km)	-0.0002
Micro 17	Ground Type (5km)	-0.0002
Micro 18	Imperviousness (point)	0.0057
Micro 19	Imperviousness (500m)	0.0039
Micro 20	Imperviousness (1km)	0.0032
Micro 21	Imperviousness (5km)	0.0016
Meso 1	Distance to River	-0.0001
Meso 2	Relative height (500m)	-0.0008
Meso 3	Relative height (1km)	-0.0004
Meso 4	Relative height (5km)	-0.0013
Meso 5	Depression (500m)	0.0000
Meso 6	Depression (1km)	0.0000
Meso 7	Depression (5km)	0.0000
Meso 8	Relative Height to River	-0.0033
Macro 1	GDP	0.0000
Macro 2	GDP/capita	0.0022
Macro 3	Government L.T. Vision	0.0001
Macro 4	Quality of Infrastructure	0.0000

Table 20: Senary Model Logistic Classifier Correlations

Feature Index	Feature	Importance
Micro 1	Max. Precipitation 1 Day	0.031
Micro 2	Max. Precipitation 5 Days	0.030
Micro 3	Monthly Precipitation (Jan)	0.031
Micro 4	Monthly Precipitation (Feb)	0.029
Micro 5	Monthly Precipitation (Mar)	0.031
Micro 6	Monthly Precipitation (Apr)	0.032
Micro 7	Monthly Precipitation (May)	0.035
Micro 8	Monthly Precipitation (Jun)	0.033
Micro 9	Monthly Precipitation (Jul)	0.033
Micro 10	Monthly Precipitation (Aug)	0.033
Micro 11	Monthly Precipitation (Sep)	0.031
Micro 12	Monthly Precipitation (Okt)	0.032
Micro 13	Monthly Precipitation (Nov)	0.031
Micro 14	Monthly Precipitation (Dec)	0.031
Micro 15	Ground Type (point)	0.016
Micro 16	Ground Type (1km)	0.014
Micro 17	Ground Type (5km)	0.010
Micro 18	Imperviousness (point)	0.028
Micro 19	Imperviousness (500m)	0.061
Micro 20	Imperviousness (1km)	0.060
Micro 21	Imperviousness (5km)	0.045
Meso 1	Distance to River	0.029
Meso 2	Relative height (500m)	0.053
Meso 3	Relative height (1km)	0.057
Meso 4	Relative height (5km)	0.064
Meso 5	Depression (500m)	0.053
Meso 6	Depression (1km)	0.007
Meso 7	Depression (5km)	0.008
Meso 8	Relative Height to River	0.025
Macro 1	GDP	0.024
Macro 2	GDP/capita	0.027
Macro 3	Government L.T. Vision	0.012
Macro 4	Quality of Infrastructure	0.010

Table 21: Senary Model Random Forest Classifier Feature Importances

F Complete Results Senary Model

As previously discussed, the senary model returns the flood return period for all of the different flood return periods: no flooding, 10 years, 20 years, 50 years, 100 years and 200 years. This means that there are six classes (0,1,2,3,4,5). The baseline of this model is thus 16.66%, as a random prediction would be right one sixth of the time. This subsection presents the results for the different machine learning methods, with a small discussion about what is portrayed in these results. The Senary model results are displayed in Table 22.

Classifier	Test Set Accuracy (%)	KR&A Set Accuracy (%)
Logistic Classifier	24.54	74.03
Random Forest	58.51	88.99
Neural Network	24.38	47.45

Table 22: Senary Model Results

F.1 Logistic Classifier

The logistic Classifier yielded an accuracy of 24.54% for the test set and an accuracy of 74.03% for the KR&A validation set. The latter has a high accuracy as the dataset is rather skewed. This would also indicate that the classifier has a tendency to choose the no flooding label (0). Although this bias seems to exist, the classifier outperforms the baseline of 16.66%. The feature importances are displayed in Table 23.

Feature Index	Feature	Correlation
Micro 18	Imperviousness (point)	0.00475
Micro 19	Imperviousness (500m)	0.00333
Micro 20	Imperviousness (1km)	0.00273
Meso 8	Relative Height to River	-0.00273
Micro 9	Monthly Precipitation (Jul)	0.00241
Micro 10	Monthly Precipitation (Aug)	0.00207
Micro 2	Max. Precipitation 5 Days	-0.00198
Macro 2	GDP/capita	0.00180
Micro 8	Monthly Precipitation (Jun)	0.00178
Micro 11	Monthly Precipitation (Sep)	0.00154

Table 23: Senary Model Logistic Classifier Correlations

F.2 Random Forest Classifier

The Random Forest Classifier yielded an accuracy of 58.51% for the test set and an accuracy of 88.99% for the KR&A validation set. This score far exceeds the baseline and also the Logistic Classifier. The corresponding feature importance plot is displayed in Table 24.

Feature Index	Feature	Importance
Meso 4	Relative Height (5km)	0.064
Micro 19	Imperviousness (500m)	0.061
Micro 20	Imperviousness (1km)	0.060
Meso 3	Relative Height (1km)	0.057
Meso 2	Relative Height (500m)	0.053
Micro 21	Imperviousness (5km)	0.045
Micro 7	Monthly Precipitation (May)	0.035
Micro 8	Monthly Precipitation (June)	0.033
Micro 9	Monthly Precipitation (July)	0.033
Micro 10	Monthly Precipitation (August)	0.033

Table 24: Senary Model Random Forest Classifier Feature Importances

The above-mentioned feature importance analysis shows that the imperviousness features significantly contribute to the predictive power of the model, although the total predictive power of the model is evenly distributed across all of the features.

F.3 Neural Network

The Artificial Neural Network yields an accuracy of 24.38% for the test set and an accuracy of 47.45% for the KR&A validation set. This is significantly lower than the Random Forest Classifier and a bit higher than the Logistic Classifier. The relative feature importance is non-trivial to derive from a neural network, requiring the application of e.g SHAP [Lundberg, 2018], which is out of scope for this thesis. Hence, for this section a Table with feature importance is not presented.

G Motivation for Null Values

Here the motivation for the choices on the replacement of the null values are discussed. Table 25 shows the Null value choices again for completeness.

Feature	N NaNs	Method to remove values
Micro 1: One Day Maximum Prec.	204	Mean of other values
Micro 2: Five Day Maximum Prec.	987	Mean of other values
Micro 3-14: Average Monthly Prec.	1	Mean of other values
Micro 15-17: Ground Type	4	Set to most occurring category
Micro 18-21: Imperviousness	4	Mean of the other values
Meso 1: Distance to river	25891	Fill to maximum distance, as NaN was set if no river was found
Meso 2-4: Relative Height	8	Set to mean of relative height
Meso 5-7: Depression (Height)	8	Set to most occurring category
Meso 8: Relative Height to river	25892	Set to 0
Macro 1-2: GDP and GDP/inh	577	Mean of other values
Macro 3: Gov. Long Term Vision	7	Mean of other values
Macro 4: Quality of overall infra.	19	Mean of the other values

Table 25: Null Values and Methods of replacement

G.1 Micro 1-2: One/Five Day Maximum Precipitation

The mean of the values is taken because the assumption is made that this feature would then have an 'average' result, which means that the feature would not contribute to the predictive power if the mean is taken. For other features, however there would be an impact as these could still sway the results for the entries that have Null values for the *Micro 1 or 2* feature.

G.2 Micro 3-14: Average Monthly Precipitation

The mean of the values is taken because the assumption is made that this feature would then have an 'average' result. Especially because there is only 1 NaN value, it would not have a high impact.

G.3 Micro 15-17: Ground Type

The most occurring category is chosen to keep this feature neutral for the locations that do not have this feature defined. This feature in its turn does not alter the current parameters of the model, which it has formed around other locations.

G.4 Micro 18-21: Imperviousness

Set the imperviousness to the mean of the imperviousness values to keep this feature neutral for the locations that do not have this feature defined. This feature in its turn does not alter the current parameters of the model, which it has formed around other locations.

G.5 Meso 1: Distance to River

There are a lot of NaN values because of the way that this feature is retrieved: if there is not a river in a 20 kilometer range, it returns NaN. Because flooding does not occur at more than 20 kilometer outside of the river, the NaN values are set to a distance of 20 kilometers. This was done to make the computation of this feature viable.

G.6 Meso 2-4: Relative Height

Set the relative height to the mean of the relative heights to keep this feature neutral for the locations that do not have this feature defined. This feature in its turn does not alter the current parameters of the model, which it has formed around other locations.

G.7 Meso 5-7: Depression (Height)

Set the depression to the most occurring category: 1 to keep this feature neutral for the locations that do not have this feature defined. This feature in its turn does not alter the current parameters of the model, which it has formed around other locations. As there are only 8 values, this decision does not bear big consequences.

G.8 Meso 8: Relative Height to River

This feature gets the NaN values from the Distance to River *(Meso 1)* feature and is set to 0, as this feature tries to explore whether the relative height to the river contributes to the predictive power of the models and what its impact is. If this value is set to 0, it negates this feature, whilst still retaining the influence of the other features.

G.9 Macro 1-2: GDP and GDP per capita

GDP and GDP per capita are set to the mean of their respective location counterparts to keep this feature neutral for the locations that do not have this feature

defined. This feature in its turn does not alter the current parameters of the model, which it has formed around other locations. These NaN values come from NUTS3 regions, which do not have these values defined.

G.10 Macro 3: Government Long Term Vision

Set the government long term vision to the mean of the defined government long term vision features to keep this feature neutral for the locations that do not have this feature defined. This feature in its turn does not alter the current parameters of the model, which it has formed around other locations.

G.11 Macro 4: Quality of overall infrastructure

Set the quality of overall infrastructure to the mean of the defined quality of overall infrastructure features to keep this feature neutral for the locations that do not have this feature defined. This feature in its turn does not alter the current parameters of the model, which it has formed around other locations.

G.12 Final Remarks

In hindsight after reanalyzing the data, it is clear that there are a few points that are outside of continental Europe and located in locations such as The Canary Islands. These values should have been removed as these have several missing values for the features. This miscalculation does not have any major consequences as these coordinates are approximately 0.5% of the data.