



INTERUNIVERSITY PROGRAMME

MASTER OF SCIENCE IN WATER RESOURCES ENGINEERING

Vrije Universiteit Brussel
Katholieke Universiteit Leuven

Global Climate Attribution of Pluvial Floods

A deep learning framework for pluvial flood modelling

by Robbe Buls

Promoter:
Prof. W. Thiery

Supervisors:
S. Lampe
D. Muheki

Master dissertation in partial fulfilment
of the requirements for the degree of

**Master of Science
In Water Resources Engineering**

August 2026

The content of this thesis can only be published with the permission of the promoter.

Preface

This thesis marks the culmination of a wonderful year. A sidebranch of my study career, so to speak. Although the main current has already carried me through quite a few landscapes, this unexpected detour still led me to a few previously unknown river basins. What began as a modest spring of ideas has today flowed into an ocean of new insights.

Writing a thesis rarely proceeds like a straightened canal. Along the way, there are meanders, unexpected obstacles, and occasionally the banks of the data storage are unstable. But ultimately, which kayaker prefers the Albert Canal to the winding Kleine Nete? It is precisely these detours that make the journey worthwhile.

Fortunately, I did not have to navigate this river alone. I could count on a strong support team that helped move the heaviest stones from the current. First and foremost, I would like to thank Professor Wim Thiery, who managed to pull a fascinating topic out of his sleeve at the start of the academic year. Additionally, my sincere thanks go to Seppe Lampe and Derrick Muheki. Their technical expertise, critical eye, and ever-willing assistance ensured that this thesis stayed afloat, even when the current became a bit more turbulent.

Finally, I would also like to thank all the other people who helped shape this journey. Entering a new river basin where the tributaries have known each other for a long time is not something to be taken for granted. Yet, I was welcomed almost immediately into the clear IUPWARE waters. I have often said that I was on an Erasmus exchange in my own country, and yet this muzungu has rarely felt so at home.

Abstract

Climate change is intensifying extreme precipitation events and the associated flood risk. Despite their growing importance, pluvial floods remain underrepresented in climate impact and attribution studies, particularly at the global scale. This thesis aims to develop a global deep-learning framework for pluvial flood modelling and to apply it for climate attribution.

Remote-sensing flood observations from the Global Flood Monitoring (GFM) system were used as target data. The models were forced with a combination of static predictors (elevation, land cover, soil characteristics) and dynamic predictors (precipitation, evapotranspiration, soil moisture). Four machine-learning architectures were developed and evaluated: one artificial neural network (ANN) and three long short-term memory (LSTM) networks with varying input configurations. To assess the influence of climate change, the final model was driven with both factual and counterfactual precipitation datasets.

The best-performing model achieved a ROC-AUC of 0.76 and a PR-AUC of 0.41. Including soil moisture and evapotranspiration substantially improved model performance, highlighting the importance of antecedent hydrological conditions. The model captures physically plausible relationships between flooding and environmental characteristics. Classification performance was considerably stronger than flood-fraction regression, with extreme flood events being systematically underestimated. Climate attribution results indicate a 0.5% increase in global pluvial flood occurrence and extent under factual climate conditions relative to a counterfactual climate without anthropogenic climate change.

Overall, this thesis demonstrates the potential of deep learning for global pluvial flood modelling and climate attribution. Although important challenges remain, particularly in representing extreme flood magnitudes, the developed framework provides a foundation for future flood impact assessments and for quantifying the influence of climate change on pluvial flooding worldwide.

List of Figures

3.1	Flowchart of the thesis.	11
3.2	Theoretical scheme showing the working of an LSTM (Reproduced from Yan (2016)).	14
5.1	Residual distribution plotted against observed flood fraction for flooded pixels. The blue dashed line indicates zero residual error. Color intensity indicates pixel density on a $\log_{10}$ scale.	30
5.2	Mean regression residual evaluated across binned intervals of landcover fractions for flooded pixels. Dashed horizontal lines denote zero error. . . .	31
5.3	Mean regression residual evaluated across binned intervals of environmental variables for flooded pixels. Dashed horizontal lines denote zero error. . . .	32
5.4	The classification evaluated across binned intervals of environmental variables for flooded pixels. The blue line depicts the observations and the orange line depicts the model output	34
5.5	The flooded fraction evaluated across binned intervals of environmental variables for flooded pixel. The blue line depicts the observations and the orange line depicts the model output	35
5.6	Case study around the city of Ivanava (Belarus) at 22/09/2024. The left panel represents the observations, the middle panel depicts the model results and the residuals (modelled-observed) are shown in the right panel. .	38
5.7	Case study around the city of Uranch (Pakistan) at 4/07/2024. The left panel represents the observations, the middle panel depicts the model results and the residuals (modelled-observed) are shown in the right panel. .	39
5.8	Case study around the city of Wanlaweyn (Somalia) at 26/11/2023. The left panel represents the observations, the middle panel depicts the model results and the residuals (modelled-observed) are shown in the right panel.	39
5.9	Daily flooded extent predicted under factual and counterfactual climate conditions. Each point represents one analysis date. The dashed line indicates equal flooded extent in both scenarios.	41
5.10	Distribution of daily flooded extent under factual and counterfactual climate conditions. Histograms are normalized to represent probability densities.	42
5.11	Exceedance curves of daily flooded extent under factual and counterfactual climate conditions. Events are ranked from largest to smallest flooded extent.	42

5.12	Spatial distribution of the relative change in modeled pluvial flood extent between factual and counterfactual climate conditions. Each tile represents the percentage change in total flooded fraction relative to the counterfactual simulation. Red shades indicate larger flooded extents under factual climate conditions, whereas blue shades indicate smaller flooded extents.	43
B.1	Boxplots for the dry and the flooded pixels per fraction of the landcover class forest, short vegetation, paved and wetlands.	61
B.2	Boxplots for the dry and the flooded pixels per the environmental variables elevation, clay content, sand content and bulk density.	62
C.1	Comparison of the ROC-curves for the fluxes-based ANN (panel A), the fluxes-based LSTM (panel B), the ERA5-based precipitation-only LSTM (panel C) and the CHELSA-based precipitation-only LSTM (panel D). The dashed grey line represents the random no-skill baseline	63
C.2	Comparison of the PR-curves for the fluxes-based ANN (panel A), the fluxes-based LSTM (panel B), the ERA5-based precipitation-only LSTM (panel C) and the CHELSA-based precipitation-only LSTM (panel D). The dashed grey line represents the random no-skill baseline. The red marker highlights the selected decision threshold	64
C.3	Comparison of the confusion matrices for the fluxes-based ANN (panel A), the fluxes-based LSTM (panel B), the ERA5-based precipitation-only LSTM (panel C) and the CHELSA-based precipitation-only LSTM (panel D). It is important to note that classification is based on an optimal threshold that is determined for each model separately and therefore differs between the four confusion matrices.	65

List of Tables

3.1	Comparison of the ROC-AUC, PR-AUC, R^2 and Pearson r between the two data loading strategies.	18
3.2	Summary of the performed sensitivity analysis on the hyperparameters. . .	18
3.4	Overview of statistical tests.	20
3.3	Overview and definitions of the classification and regression evaluation metrics.	21
4.1	Producer and user accuracy of the CLMS map in 2020 (European Union’s Copernicus Land Monitoring Service (CLMS), 2026)	24
5.1	Performance evaluation of four machine learning architectures for spatial flood mapping across binary classification and continuous flood fraction regression metrics.	28
5.3	Spearman rank correlation (ρ) between continuous environmental variables and the absolute prediction error in flooded pixels.	32
5.2	Kruskal-Wallis test results evaluating the absolute prediction error across binned landcover fractions and terrain quintiles for flooded pixels.	33
5.4	Mann-Whitney U test results and Rank-Biserial Correlation (RBC) effect sizes comparing dry vs. flooded conditions for observed and predicted states.	36
5.5	Mann-Whitney U test results and Rank-Biserial Correlation (RBC) effect sizes comparing observed vs. predicted conditions within the flooded and dry subsets.	37
5.6	Comparison of observed and modelled flood fraction distributions ($N = 38, 225$).	38
5.7	Comparison of factual and counterfactual flood fraction distributions. . . .	40
A.1	Overview of data sources	60
A.2	Overview of the data after pre-processing	60

Contents

Preface	i
List of Figures	iii
List of Tables	v
Contents	vi
1 Introduction	1
1.1 Impact of floods	1
1.2 ISIMIP	2
1.3 Research objectives	2
2 State of the art	4
2.1 Types of floods	4
2.2 AR6 summary	5
2.3 Flood modelling	7
3 Methods	11
3.1 Methodology overview	11
3.2 LSTM	12
3.3 Model architecture	14
3.4 Evaluation	19
3.5 Climate attribution	20
4 Data	22
4.1 Observed floods	22
4.2 Static data	23
4.3 Precipitation	25
4.4 Historical fluxes	26
5 Results and discussion	28
5.1 Evaluation	28
5.2 Climate attribution	40
6 Limitations	44

7	Future research	46
7.1	Improved analysis with the current setup	46
7.2	Improvements to the model	46
7.3	Landcover change attribution	47
7.4	Regional application	47
8	Conclusions	49
8.1	Model	49
8.2	Climate attribution	49
	Bibliography	50
	Appendices	59
A	Data overview	60
B	Flood dependency on environmental conditions	61
C	Model performance	63

Chapter 1

Introduction

1.1 Impact of floods

Floods accounted for 43% of all recorded natural disasters between 1994 and 2013 and affected nearly 2.5 billion people (CRED and UNISDR, 2015). That makes floods by far the largest driver of human exposure globally. CRED and UNISDR (2015) states that recurrent flooding leads to chronic malnutrition, loss of agricultural production, and long-term poverty, especially in rural areas. Additionally, floods caused more damage to health and education infrastructure than any other type of disaster. Although all income groups experience floods, the impact is disproportionately higher in low-income countries, where construction standards are weaker and preparedness is lower (CRED and UNISDR, 2015). Floods are less deadly than storms or earthquakes, but 55% of all people affected by disasters are affected by floods. Jonkman (2005) found that the flood type explains mortality better than the region. Asian river floods cause by far the most casualties (40%) and affected persons (96%), but flash floods have the highest mortality (3.6%). The main mechanism for that, is that population concentrates in river deltas, coastal areas and urban areas with poor drainage. 1.81 billion people live in areas with significant flood risk, and that these risks correlate strongly with poverty, particularly in Sub-Saharan Africa and South Asia (Rentschler et al., 2022).

Future projections predict that the average annual flood damage in major coastal cities will increase extremely due to socio-economic growth, sea level rise, and land subsidence. Without substantial adaptation, losses will rise to more than 1 trillion USD per year by 2050 (Hallegatte et al., 2013). Even with adaptation, losses increase and cities become more dependent on their protecting infrastructure (Hallegatte et al., 2013). Under a future scenario of 3°C global warming by the end of the century, a child born in 2020 will experience three times as many river floods across their lifetime as one born under pre-industrial climate conditions (Thiery et al., 2021). For people under 40 in 2020, river flood exposure becomes unprecedented with warming above 1.5°C (Thiery et al., 2021). There are, however, large projected regional differences. Regions such as Sub-Saharan Africa, South Asia, and the Middle East and North Africa show the strongest increase in river flood exposure, while there is a smaller but still significant increase in high-income countries (Thiery et al., 2021).

1.2 ISIMIP

The Inter-Sectoral Impact Model Intercomparison Project (ISIMIP) is a framework to support research into the risks and impacts caused by climate change. The main question ISIMIP tries to answer is how climate change affects natural and human systems currently and how it will do that in the future (Rosenzweig et al., 2017). The effects of climate change can be studied over a wide variety of sectors and scales. It is important to identify potential 'hotspots': regions where multiple risks co-occur. Besides, effects can also interact, both in a positive and in a negative way. A wide variety of models, input data, and scenarios make it impossible to compare effects predicted by different studies and on different sectors or scales. Therefore, ISIMIP provides a standardised protocol that includes input data, output format and scenarios. The protocol includes naming conventions, file types, units, grids, etc. The ISIMIP data repository provides input data that can be used by researchers in order to have a consistent baseline to start from. The data provided by ISIMIP are both climate data and socioeconomic data. Both historical observations and projections under multiple scenarios are provided for the two types of data. Additionally for the climate data, detrended or counterfactual observations, which are simulations for the historic period in a world without climate change, are available. This mainly serves impact attribution studies. The goal is to serve both regional and global studies, and intra- and intersectoral research. Therefore, ISIMIP is divided into several sectors, under which agriculture, water and health (The Inter-Sectoral Impact Model Intercomparison Project, 2025). It is a community-driven project with open-source data and a focus on direct impacts on humanity. The goal of this thesis is, therefore, to work within the ISIMIP framework and to contribute to its data repository.

1.3 Research objectives

Currently, flood-related data in the ISIMIP data repository are limited to fluvial (river) floods. No corresponding data are currently available for pluvial floods, which are caused by excessive precipitation that cannot run off or infiltrate (see Section 2.1 for a more detailed description). Moreover, most flood-related studies today focus either on river floods (e.g., Arnell and Gosling (2016); Hirabayashi et al. (2013); Nearing et al. (2024)) or on pluvial floods in urban areas (e.g., Samadi et al. (2025); Li and Willems (2020); Liao et al. (2023)). Additionally, despite the growing impacts of pluvial flooding, no global assessment has been done yet on their impacts under a changing climate. Even in the Sixth Assessment Report (AR6) of the Intergovernmental Panel on Climate Change (IPCC), no special attention is given to pluvial floods. The AR6 refers to the effects of climate change on extreme precipitation events as an indication of the impact on pluvial floods (Seneviratne et al., 2021).

Given the strong dependence of pluvial floods on extreme precipitation, the effects of climate change on pluvial flooding are likely to be stronger than those observed for river floods. Therefore, this thesis aims to fill this research gap within the ISIMIP framework and provide data for regional assessments. The goal is to make a machine-learning (ML) model that can predict pluvial floods on a global scale. Following this, using modelled climate scenarios in our ML model, the potential changes in pluvial flood occurrence can be studied, as well as their spatial differences. Three main research objectives are identified:

- Setting up a model under historical observations. Four different models are set up

with varying complexities and inputs;

- Evaluating the performance of the four models by using independent test data. A sensitivity analysis has to be performed in order to investigate what the strengths and weaknesses of the model are;
- Feeding the model with detrended precipitation data in order to assess the impact of climate change on the occurrence and the intensity of pluvial floods.

The second objective provides a quality assessment of the models, which allows to compare the different models amongst each other. The performance of the model on historical observations is assumed to be indicative for the performance for the counterfactual runs and for future runs that could potentially be performed for different climate scenarios. The third objective allows to indicate the global trend in pluvial floods under a changing climate and to compare different regions, land covers, soil types, etc.

Chapter 2

State of the art

2.1 Types of floods

Floods can be classified by the cause of the inundation. The first type is a fluvial flood or a river flood, which occurs when water levels exceed the river banks, causing water to overflow and inundate the river floodplain. The second type is a pluvial flood, which can occur independently of a waterbody and is caused directly by precipitation (Office of Public Works, 2026). Sometimes, pluvial floods are referred to as flash floods. However, in this thesis the term pluvial flood will be used, and a flash flood is assumed to be a very fast form of flood that was triggered by excessive precipitation in a short amount of time, as is done by the National Weather Service (NOAA's National Weather Service, 2024). Besides these two major types of floods, other causes of floods exist as well. Amongst these types are coastal floods, groundwater floods, (glacial) lake outburst floods, and floods caused by rapid snowmelt (Office of Public Works, 2026). CRED and UNISDR (2015) emphasizes that flash floods and acute riverine/coastal floods are increasing, partly due to urbanization and changing hydrological regimes. This makes floods faster, less predictable, and more dangerous.

Pluvial floods, when defined as above, can be caused by two different mechanisms. When the intensity of the precipitation exceeds the ability of the soil to absorb water, this is called infiltration excess overland flow or Hortonian runoff. That ability to absorb water is referred to as the infiltration capacity of the soil. Hortonian runoff is typically linked to situations with heavy rainfall over a short period of time, since there is pore space available for the water to infiltrate, but the precipitation rate exceeds the infiltration capacity. One would expect this type of flood on soils with a low infiltration capacity. The infiltration capacity is determined by soil texture (grain size) and structure (e.g. aggregates, compaction), but to the water content as well (Bharali, 2021). Coarser soils generally have a higher infiltration capacity (FAO et al., 2012). Similarly, soils with good structure, like forest soils or soils with much organic material, have a high infiltration capacity (Bharali, 2021). Structures like plant roots, holes of earthworms or other macropores increase the infiltration capacity (Ward and Robinson, 2000). As Horton's equation explains, the infiltration capacity decreases exponentially during a precipitation event (Guo, 2024; Horton, 1940).

$$f(t) = f_c + (f(0) - f_c)e^{-kt} \quad (2.1)$$

In this equation, $f(t)$ is the infiltration capacity at time t , $f(0)$ is the infiltration capacity at $t=0$, f_c is the infiltration capacity at the moment the soil gets saturated and k

is a constant depending on the type of soil and the vegetation. The exponential decrease in infiltration capacity is caused by the capillary suction of the soils, which is higher for dryer soils. Factors decreasing the infiltration capacity of the soil are crusts, compaction, frozen water and hydrophobic soils (for example occurring after wildfires). Paved soils have an extremely low infiltration capacity. The runoff volume per time unit resulting from Hortonian runoff equals the precipitation minus infiltration capacity.

Another mechanism by which pluvial floods can occur is by saturation excess overland flow or Dunne runoff. In that case, the soil pores are entirely filled and no more water can be stored in the soil. The total pore volume, and therefore also the maximal amount of water that can be stored in the soil, is called the porosity. The difference between the porosity and the soil water content is referred to as the storage capacity. If the soil water content is equal to the porosity (and therefore the storage capacity is zero), Dunne runoff occurs. This type of flood is expected after a longer period of rainfall, since the pores have to be filled with water. Soils with a large number of macropores (typical for forests) have a larger porosity, as well as organic soils like peat soils. Sandy soils have large pores, due to its large grain size. Clay rich soils on the other hand have often many micropores, together leading to a large porosity. An intermediate minimum occurs for silt soils. Rocky soils, compacted soils and paved soils have very low porosities.

2.2 AR6 summary

The Sixth Assessment Report (AR6) of the Intergovernmental Panel on Climate Change (IPCC) states that pluvial floods are directly linked to extreme precipitation, and uses that therefore as a proxy with the assumption of no additional change in surface condition occurred. Therefore, the AR6 does not assess pluvial floods, but refers to the assessment of extreme precipitation. In this thesis, this knowledge gap will be investigated. Below, we summarize the findings of the AR6 on heavy precipitation and river floods.

Observed trends

It is likely that the frequency and the intensity of **extreme precipitation** have increased on a global scale (high confidence for daily intensities) (Seneviratne et al., 2021). The annual maximum daily precipitation (Rx1day) and the five-days precipitation (Rx5day) have increased since the middle of the 20th century over land on the areas with good data (Sun et al., 2021). The share of the precipitation from the wettest days (R95pTOT) has significantly increased (Dunn et al., 2020). However, there is a large regional and seasonal variability in these patterns. In Europe and North-America, there is high confidence in the increase in magnitude and intensity of extreme precipitation since 1950 (Sun et al., 2021). In Asia, there is high confidence for an increase as well, although the spatial heterogeneity is larger here (Sun et al., 2021). For Africa, Central and South America, Australasia and small islands, there is often low to medium confidence for an increase, given the limited data coverage and some inconsistent trends (Sun et al., 2021; Dookie et al., 2019). For the sub-daily extremes, there is a low confidence in the global trend, since the available data series are short and spatially limited (Ali et al., 2021). Nevertheless, many regional studies show an intensification in sub-daily extremes.

There is low confidence in the global **peak flow** trends over the past decades (Seneviratne et al., 2012). A wide variety in different regions occurs, with an increase in North-eastern US, Northwestern Europe and the Amazon (Barichivich et al., 2018; Neri et al.,

2019; Blöschl et al., 2019) and a decrease in parts of the Mediterranean region, Australia, Africa and the Southwest of the US (Tramblay et al., 2020; Wasko and Nathan, 2019; Blöschl et al., 2019). There is high confidence that in snow-rich regions seasonality shifts, with peak discharges earlier in the year (Dudley et al., 2017).

Mechanisms and attribution

There is high confidence for the ‘thermodynamical effect’ as the mechanism explaining the observed changes in **extreme precipitation** (Seneviratne et al., 2021). The effect describes the increase in vapour pressure (the volume of water the air can contain) when the air temperature increases. At a global scale, the rate of increase in extreme precipitation follows approximately the Clausius-Clapeyron relation of 7% per $^{\circ}C$. Additionally, dynamical changes in for example monsoons, tropical cyclones, ENSO, NAO, etc. can locally lead to a reinforcement of the thermodynamical effect, although there is often only low to medium confidence on a regional scale (Nie et al., 2018). Other potential causes are a reduction in aerosols and urbanisation (Samset et al., 2018; Sillmann et al., 2019), which enforces local extreme precipitation (medium confidence). The observed intensification of extreme precipitation (Rx1day, Rx5day) on a global scale cannot be explained by natural variability only, but anthropogenic forcing is, with high confidence, the dominant cause (Paik et al., 2020). Also on continental scale, there is high confidence in the human influence in the data-rich continents (Europe, North America and Asia) (Zhang et al., 2013), but on AR6-region scale, the signal is still weak (low-medium confidence).

For **floods** in general, AR6 concludes that the relation between floods (all types) and heavy precipitation is not always one-on-one, nor between changes in extreme precipitation and changes in floods. Floods are more complex, with different factors in different regions, and at different moments in time other factors that play a more or less important role. The main factors identified to be related to river floods are antecedent soil moisture and snow-water content (Pall et al., 2019; Woldemeskel and Sharma, 2016; Berghuijs et al., 2016). Globally, there is low confidence for anthropogenic influence on the trends in river peak discharges (Seneviratne et al., 2021), although some small-scale studies show high confidence in human impact locally (Teufel et al., 2019; Philip et al., 2018; Cho et al., 2016).

Projections

There is high confidence that **extreme precipitation** events that occur once per year or less will proportionally increase with global temperature (7% per $^{\circ}C$, following the Clausius-Clapeyron curve) (Seneviratne et al., 2021). This scale is mostly independent on the emission scenario or the external forcing (CO_2 , CH_4 , aerosols...) (Pendergrass et al., 2015). For a warming between $1.5^{\circ}C$ and $2^{\circ}C$, it is projected to have a notable increase in events with a return period of 20- and 100 years. CMIP5 even predicts an increase of more than 45% (Kharin et al., 2018). For a warming of $4^{\circ}C$, 10- and 50-year events will become double to three times as abundant as currently (Li et al., 2021). There is high confidence in a non-linear increase, which is stronger for rarer events Fischer et al. (2015). Almost everywhere over land, extreme precipitation events are projected to increase, except for some sub-tropical (semi-)arid regions (Li et al., 2021). Regional differences are determined by a combination of the thermodynamical effect and the dynamical changes (e.g. monsoons, ENSO). There is high confidence in the role of both effects, but lower

confidence in the exact regional outcomes (Seneviratne and Hauser, 2020). Convection-permitting regional climate models (RCMs) project a clear increase in the intensity of sub-daily events in, amongst others Europe, Africa and North America (Helsen et al., 2020; Fowler et al., 2021). These projected increases are often stronger than Clausius-Clapeyron predicts. It must be stated that only a limited amount of studies was performed, but these studies point in the same direction.

On a global scale, more regions are projected to show an increase in **river floods** than regions that show a decrease (medium confidence) (Seneviratne et al., 2021). For the regional patterns, there is often medium to high confidence, with an increase in parts of tropical Africa, South-East and Northern Asia, the Western Amazon, the Andes and high latitudes in North America. In Central- and Eastern Europe, South-West Africa, South- and Central North-America and the Mediterranean on the other side, river peak discharges are projected to decrease (Dankers et al., 2014; Arnell and Gosling, 2016; Hirabayashi et al., 2013).

2.3 Flood modelling

Types of flood models

A common way of modelling floods is by using process-based models or white-box models, where a combination of physical processes within the hydrological cycle are explicitly described by (differential) equations. Typical equations are the Saint-Venant equations for flow in channels or the Richards-equation of soil moisture. The entire system of equations can be solved analytically or by using numerical schemes. The development of such models requires knowledge about the physics of the system and often rely on a number of assumptions and simplifications. Additionally, the implementation of such models asks for realistic values of hydrodynamic or hydrological parameters (Kim et al., 2015). Costabile and Macchione (2015) found that physical models often have problems with short-term predictions. For numerical models, systematic errors are indicated to make predictions unreliable. Process-based models in real-time predictions can suffer from a trade-off between accuracy, resolution and speed (Aghenda et al., 2025).

The counterpart of process-based models are data-driven models or black-box models. These models look for a relationship between inputs and outputs, without any mechanisms taken into account. In these types of models, floods are often considered as stochastic processes. This allows for predicting floods by using probability distributions of the historic observations (often streamflow data) (Kroll and Vogel, 2002). The term flood frequency analysis (FFA) is commonly used. A wide variety of statistical models exists, amongst which autoregressive moving average (AMA), multiple linear regression (MLR), climatology average method (CLIM), quantile regression techniques (QRT) and Bayesian forecasting models (Mosavi et al., 2018). The main advantage of using data-driven methods is that limited physical knowledge on the processes is required and that no simplifications or assumptions have to be done. The architecture of such models is often relatively simple. However, the quality of the model highly depends on the amount and the quality of the data. Especially in ungauged areas, the usage of statistical models requires large assumptions and interpolations. Another potential problem with data-driven models is the 'generalisation problem'. This raises the question whether the model can sufficiently predict situations for which it was not trained. For instance, if changes are brought into the system, would the model still be able to perform as well as it did before

the changes were introduced Solomatine and Ostfeld (2008)? While traditional statistical data-driven models have proven useful, their linear nature often fails to capture the complexity of hydrological systems.

As a compromise between white- and black-box models, often conceptual models or grey-box models are used. These models use simplified concepts of reality instead of complex differential equations. The hydrological cycle is represented as a series of interconnected reservoirs (e.g. the atmosphere, the river, the soil, the aquifer, etc.). In general, the hydrological cycle is represented as a water balance of the different reservoirs, for example in equation 2.2, where $S(t)$ is the storage for a certain reservoir, $P(t)$ is the precipitation, $E(t)$ is the evapotranspiration, and $Q(t)$ is the discharge.

$$S(t + 1) = S(t) + P(t) + E(t) - Q(t) \quad (2.2)$$

These models typically rely on parametrisations of many processes and, therefore, require sufficient calibration based on historical series.

Machine learning in flood modelling

Machine Learning (ML), which are advanced data-based methods, could potentially act as a more robust alternative. These methods have been improving rapidly the last two decades (Mosavi et al., 2017). Aghenda et al. (2025) found barely any paper about the topic before 2010 and sees a strong increase in ML-research for flood modelling since 2020. Also in rainfall forecasting, rainfall-runoff modelling and flood forecasting, these methods have shown their relevance. Machine learning is being applied increasingly, often because classical hydrological and hydrodynamic models are complex, computationally intensive, and difficult to calibrate. Flood processes are highly non-linear, meaning that data-driven models often learn patterns more efficiently than physical-based models (Karim et al., 2023). Early ML studies primarily used simple regression models, Artificial Neural Networks (ANNs) and Support Vector Machines (SVMs) (Abbot and Marohasy (2014); Furquim et al. (2016)). A technique that started gaining interest around 2018–2020 is hybridisation of flood models (Venkatesan and Mahindrakar (2019); Chen et al. (2019); Deng et al. (2022)). Hybrid models are combining the strengths of process-based (hydrodynamic) and data-driven models to improve input, structure, or processing (Perera et al., 2026). In that way, the trade-offs of both can be limited. Often, the data-driven model is an advanced machine learning method.

Since 2020, deep learning (DL) models such as Convolutional Neural Networks (CNNs) and Long Short-Term Memorys (LSTMs) have dominated (Ding et al., 2020). Recent research combines hydrological models with DL architectures (Cui et al. (2021); An and Ouyang (2025)), leading to better generalization and physically consistent behaviour. Deep learning models demonstrate superior performance compared to classical ML models, especially with complex spatio-temporal input (Aghenda et al., 2025).

The modelling of pluvial floods focusses mostly on urban pluvial floods. Generally, pluvial floods are hard to model, because of the very high spatial and temporal rainfall variability that is required and the micro-topography that has strong impacts. Additionally for urban floods, the interaction with the sewer network and the short lag-times that are related with paved surfaces (Cea et al., 2025). Because of the importance of sewer systems in urban pluvial floods, often hydrodynamical models for the sewer systems are used in combination with a rainfall-runoff model (Bulti and Abebe, 2020). Also in the

modelling of urban pluvial floods, ML-techniques found their entrance. Liao et al. (2023) for example developed a 1D-CNN-model that can predict the spatio-temporal evolution of urban pluvial flooding extremely quickly and with high accuracy. Their model is $600\times$ faster than a physical model (SWMM and LISFLOOD-FP) and maintains almost the same accuracy.

LSTM models for flood modelling

In this thesis, a Long Short-Term Memory (LSTM) will be used to model the pluvial floods. In the methods section, a more detailed description of the working of a LSTM will be provided. Here, a short summary of some cases where LSTMs are applied in flood modelling are given. The reason why LSTMs are popular in flood (and river) modelling, is because they have the ability to learn long-term dependencies, which is necessarily to capture storage effects in a catchment (e.g. soil moisture, snowpack, etc.) (Kratzert et al., 2018a). Kratzert et al. (2018a) showed that LSTMs are a useful technique in rainfall-runoff modelling. The LSTM in ungauged settings was proven to outperform a calibrated physical model (SAC-SMA) in gauged settings (Kratzert et al., 2018b).

Flood warning systems often model the streamflow of the rivers (hydrodynamically or data-based) and combine that with an inundation model that translates the streamflow data into a mapping of flood extent or depth (Nevo et al., 2022). ML-methods have been used in flood warning systems both for the streamflow and the inundation modelling (Kabir et al., 2020). In 2017, the Flood Forecasting Initiative was launched by Google. It collects real-time data from different sources on the Google servers and forecasts both river stages and inundation along the river network (Nevo et al., 2022). As target data, river stage measurements from gauging stations are used, while physics-based and conceptual models usually use discharge. For the stage modelling, both an LSTM model and a linear flood model are used. For the inundation modelling, also two different ML-models were used. The LSTM model actually consists of two LSTMs: a sequence-to-one hindcast model and a sequence-to-sequence forecast model. The hindcast reads the entire timeseries until the present state and builds an internal representation of the hydrological state. The outputs of the hindcast model are used as the initial state for the forecast model (Nevo et al., 2022). It is important to note that because of the use of river stage data as target, the model is theoretically only able to predict fluvial floods.

More recently, the Google Floodhub was launched. This is a global LSTM-based model for river floods. The model uses an encoder-decoder LSTM with a hindcast of 365 days and a forecast of 7 days. As input data, six meteorologic variables were used (total precipitation, 2m temperature, surface net solar and thermal radiation, snowfall and surface pressure). The data arise from four different sources: high-resolution forecasts, reanalysis, precipitation gauges and satellite-based precipitation measurements. Apart from the meteorologic data, some static geophysical and anthropogenic characteristics like soil, landcover, actual and potential evaporation, etc. are used. As target data, streamflow data from the Global Runoff Data Centre (GRDC) were used (Nearing et al., 2024). Streamflow was not used as input data itself but only as target data. The model predicts streamflow, and those predicted hydrograms are subsequently converted into flood events by comparing them with return-period thresholds that have been predetermined from historical GRDC observations. A comparison between the Google Floodhub and GloFAS, showed that the AI model predicts extreme river floods up to 5 days in advance equally good or better than GloFAS nowcasts (0 days). It does this completely out-of-

sample, whereas GloFAS is partially calibrated at the same locations (Nearing et al., 2024). Again, it is important to note that the model is trained on streamflow data and can therefore only predict fluvial floods.

Chapter 3

Methods

3.1 Methodology overview

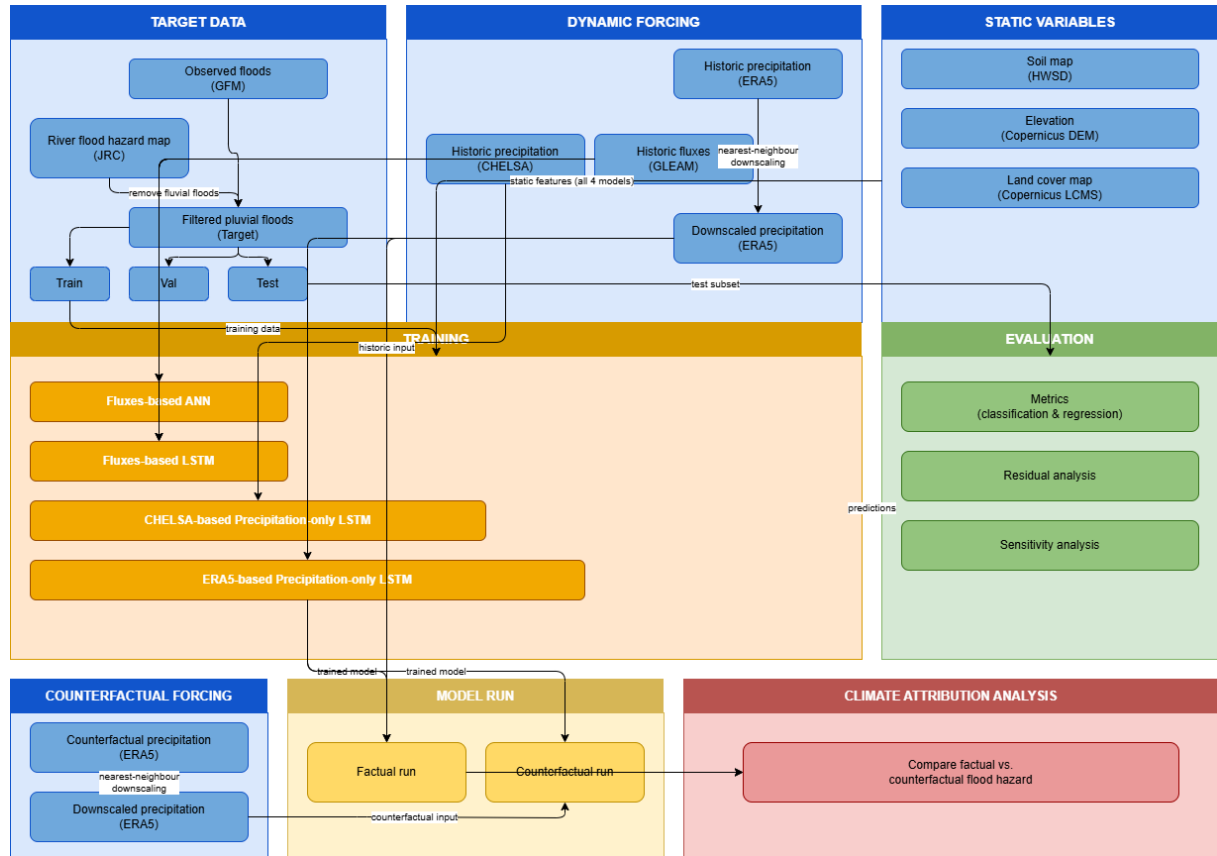


Figure 3.1: Flowchart of the thesis.

Figure 3.1 shows a flowchart of the methodology of this thesis. A detailed description of the different input data is provided in Chapter 4. The target data of observed floods comes from the Global Flood Monitoring (GFM) dataset. Since this dataset contains both pluvial and fluvial floods, we first had to filter out the fluvial floods. The filter was based on a hazard map for river floods from the Joint Research Center (JRC). The pluvial floods, which act as target data, are split into a training, validation and test subset.

As feature data, some static variables of the environment were provided to the model

i.e., elevation (Copernicus DEM), landcover (Copernicus LCMS) and soil characteristics (HWSD). Besides these static features, also a number of time-dynamic features are provided, namely the precipitation, the evapotranspiration and the soil moisture content. In a first phase, an ANN is set up which is provided with the CHELSA precipitation and the modeled evapotranspiration and soil moisture from the GLEAM database. As a potential to increase the model performance, the complexity of the model is increased by going from an ANN to a LSTM. Since counterfactual data are not available for the GLEAM variables, a simplified LSTM has to be trained, solely on the precipitation from the GLEAM dataset. Since the counterfactual precipitation data are not yet downscaled using the CHELSA-algorithm, a simplified representation of the precipitation has to be used as input. Therefore, the CHELSA inputs are replaced by ERA5-data that are downscaled by using a nearest neighbour algorithm. In total four models are trained.

After training, the performance of the four models are evaluated and compared with each other, based on the test subset of the pluvial floods. In order to assess the impact of climate change on pluvial floods the precipitation-only LSTM has to be fed by a counterfactual precipitation dataset. Since the counterfactual dataset has not been downscaled yet by the CHELSA-team, we used the counterfactual precipitation from the ERA5-dataset at a spatial resolution of 0.1° (approximately 9 km). The CHELSA daily dataset is not based on the ERA5-dataset, but on the W5E5-dataset. The model will however been run on the resolution of the original model (30 arcsec). Therefore, the ERA5 data were downscaled with a naive nearest neighbour algorithm. Because of the different downscaling method and the different dataset, the model had to be trained again with this downscaled ERA5-dataset instead of the CHELSA dataset.

A similar approach cannot be done for the fluxes-based LSTM, as there is no counterfactual data available for the GLEAM dataset. Additionally, the counterfactual GLEAM-data would needed to be based on the same inputs as the precipitation data.

3.2 LSTM

The first model developed is an artificial neural network (ANN). An ANN is a computational model, inspired on the human brain, that contains computational nodes. These nodes are referred to as neurons and they process the data. An ANN exists of three types of layers: an input layer, one or multiple hidden layers and an output layer. When training an ANN, the weights of the hidden layers are adapted to the data. In that way, the ANN learns the patterns between inputs and outputs. ANNs have no memory, meaning that they use every input as a separate independent event. As discussed before, the soil water content is likely to be an important variable to predict pluvial floods. Therefore, the model is converted to a more complex type of model with an explicitly stored memory.

Recurrent neural networks (RNNs) are machine learning models that are designed for the processing of timeseries, by saving a certain amount of previous inputs in the internal memory. Traditional RNNs use their hidden state to store this information from previous steps. When training RNNs, the gradient (error correction) is step by step given back from the last to the first layer. In networks with multiple layers, this signal can become progressively weaker due to repeated calculations, until it almost disappears. Since the first layers do not receive any useful information anymore, they do not longer adapt their settings. This causes the learning process to stall and prevents the model from being trained optimally. That is what is referred to as the 'vanishing gradient problem' (Fronk

and Petzold, 2025). The exploding gradient problem is the opposite: during training the gradients become extremely large (Rehmer and Kroll, 2020).

In order to solve these two problems, the Long Short-Term Memory (LSTM) was developed, which is a type of RNN. The original idea for LSTMs was published by Hochreiter and Schmidhuber (1997). LSTMs have internal memory cells, added to the normal timesteps. Therefore, the LSTM can use linear operations instead of multiplication every time step. Consequently, the gradient should be more stable than in other RNNs.

The input for the LSTM should be a sequence of vectors $[x[1], x[2], \dots, x[t]]$ containing the model input features for all timesteps T . LSTMs work with gates that determine the amount of information that can flow through at a certain moment. Each gate takes the input of the current time step $x[t]$ and the hidden state of the previous time step $h[t-1]$. Each gate calculates a value between zero and one (using a sigmoid function σ) and uses that value to determine which information to pass. In total there are three gates: the input gate, the forget gate and the output gate. Additionally, the candidate information to be added is not a real gate, but has a similar working. Each gate receives three sets of parameters: a weight matrix for the inputs W , a weight matrix for the hidden state U and a bias-vector b . W determines how strong each input feature (e.g. precipitation, elevation ...) influences the gate. W therefore learns the physical sensitivities of the system. U determines how strong the previous hidden state influences the gate. If U reacts strongly to a specific internal state, the forget gate can learn that that state must slowly decay. One could say U is the internal hydrological structure that the model has learned itself. Finally b determines the baseline of a gate. Figure 3.2 depicts a schematic overview of an LSTM.

The input gate $i[t]$ determines how much new information can enter the model. It determines the amount of the inputs that is relevant to update the internal hydrological state. Since this should be a value between 0 and 1 (enter nothing = 0, enter everything = 1), a sigmoid function σ is used.

$$i[t] = \sigma(W_i x[t] + U_i h[t-1] + b_i) \quad (3.1)$$

The forget gate $f[t]$ determines how much of the old memory is saved. It states the information of the previous state that is still valid in the next timestep. Also this is scaled with a sigmoid function (forget everything = 0, remind everything = 1).

$$f[t] = \sigma(W_f x[t] + U_f h[t-1] + b_f) \quad (3.2)$$

The cell input $g[t]$ contains the candidates for the update: the new information that theoretically could be added to the memory. It determines the potential change in hydrological state based on the new inputs. This one is not scaled with a sigmoid function (between 0 and 1), but with a hyperbolic tangent $\tanh$ (between -1 and 1).

$$g[t] = \tanh(W_g x[t] + U_g h[t-1] + b_g) \quad (3.3)$$

The output gate $o[t]$ determines how much of the internal memory becomes visible in the output. It computes what part of the internal hydrological state is used to predict the next state. This is again scaled by using a sigmoid function.

$$o[t] = \sigma(W_o x[t] + U_o h[t-1] + b_o) \quad (3.4)$$

The cell state $c[t]$ is the core of the LSTM and gives the situation of the cell at a certain moment. It is updated as described in equation 3.5: the old memory is multiplied

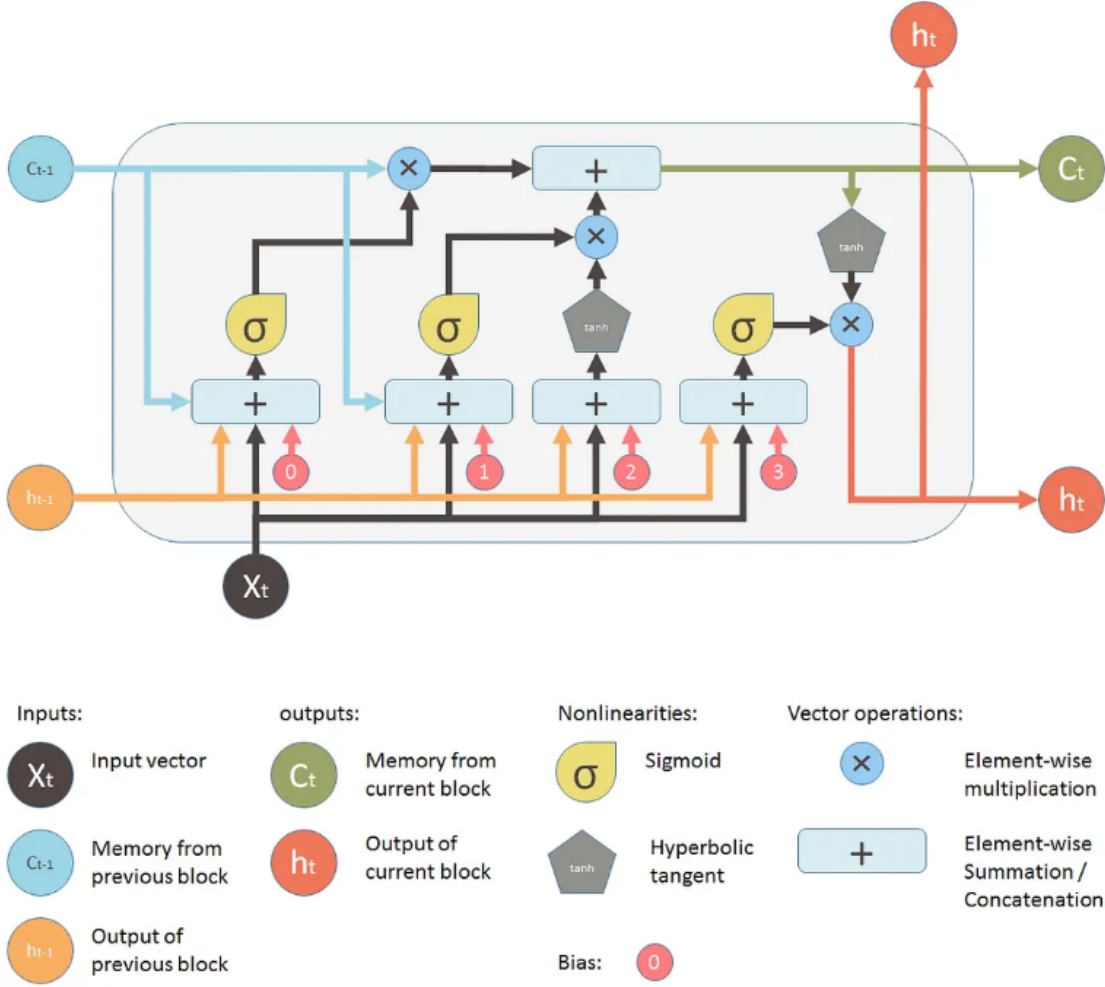


Figure 3.2: Theoretical scheme showing the working of an LSTM (Reproduced from Yan (2016)).

with the forget gate and the new information is multiplied with the input gate. The new state is a combination of what is remembered from the previous timestep and the new information in the current timestep. The symbol $\odot$ stands for the Hadamard product, which is a vector operation in which each element is the product of the corresponding elements in the two matrices.

$$c[t] = f[t] \odot c[t - 1] + i[t] \odot g[t] \quad (3.5)$$

Finally the hidden state $h[t]$ is the output of the LSTM going to the next timestep. The visible output is a filtered version of the internal state.

$$h[t] = o[t] \odot \tanh(c[t]) \quad (3.6)$$

3.3 Model architecture

All data were stored in different folders on the Hydra-cluster of the Vlaams Supercomputer Center (VSC). The downloads of the different input data were performed by using different forms of API (Application Programming Interface). The downloaded data appeared

in different formats (tiff, Zarr, NetCDF, etc.), at different resolutions and in different spatial extents. In order to save storage space and to ease programming, all data were transformed to a homogeneous format (tiff) and resolution (1 km). Since the entire globe at kilometre-scale requires enormous tiff-files, the globe was cut in 516 tiles. These tiles were homogeneous over all input variables.

Due to large differences in filenames and folder structure, first a parquet file containing the location of all input files was created. This uniform catalog contains only metadata that makes it easier to finally program the data loader. The next step is to make a data-frame containing (1) all flooded pixels (at the moment of flooding) and (2) a sample of non-flooded pixels. Reading one data-frame makes the training much faster than opening and closing the original raster files continuously. To sample in a strategic way, a distinction was made between pixels that flooded somewhere in the observation period and pixels that never flooded. For the pixels that flooded somewhere in the dataset, a random sample of 5% (of the pixel-date combinations) was taken. For the pixels that never flooded, only 0.1% of the pixel-date combinations was sampled. This data-frame only contains metadata. To prevent a time-consuming opening and closing of tiffs during the training phase, all dynamic and static predictor variables were extracted from the raster datasets and stored in tile-specific HDF5 files. During training, the data loader reads the required time series directly from these HDF5 files rather than repeatedly accessing the original raster files. The architecture of the LSTM-code consists of three big components: the data loader, the Lightning data module and the effective LSTM-model.

The data loader essentially combines all inputs into one consistent dataset for all pixels and all timesteps. The dataset is first initialised, where for each pixel and each day the tile, x, y, date and flood value is added. For each day, a timeseries of 30 days before the target day is built. Missing values are not discarded but explicitly encoded using a binary mask feature, indicating the presence or absence of valid observations. This allows the model to learn how to handle incomplete data rather than excluding such samples entirely. As input dimensions, the fluxes-based ANN and LSTM has three input variables (precipitation, evapotranspiration and soil moisture), with each their value and mask. The precipitation only models (CHELSA-based and ERA5-based) receive only one variable (precipitation), which has a value and a mask. That gives 2 features per day, over a timeframe of 30 days plus the 8 static variables.

Finally, the input variables were standardised using z-score normalisation. This ensures comparable feature scales and stabilises the training process. These steps are done for both the static and the dynamic variables. For the precipitation only, a log-transformation was applied, since precipitation typically has a high amount of zeros and a certain amount of very high values. Transforming it makes these extreme values less dominant and prevent extremely high or low z-scores.

In the data module, the data is split into training, validation and test sets. The training set is used to adapt the weights in the LSTM in order to minimise the error (often referred as ‘loss’). The validation set is used during the process of training. After each round of training (often called ‘epoch’), the model is tested on these validation data. Since the model has not seen these data during the adaptation of the weights, the data are assumed to be independent and can therefore be used to assess how well the model performs on new data. Based on this results, the decision is taken if training should be stopped or continued. Finally, the test set is only used once the entire training is finished. The goal of the test set is to have an independent estimation of the model performance. If the

evaluation points at a poor performing model, the model is not trained again, but the input data and model architecture has to be reconsidered. 70% of the data becomes part of the training set, 15% is used for validation and 15% is used for testing the model.

The split is based on the moment of observation. The period between 2016 and 2021 is used to train the model. The year 2022 is used as validation set. Finally, the years 2023 and 2024 are used as independent test set. By using a stratified shuffle split, the flood-ratio in all three sets remains comparable. Despite there was a first sampling when making the data-frame, there are much more pixels dry than flooded, there is an extreme class imbalance. Using this skewed distributed data, the training would be biased toward a negative result, causing the model to nearly always predict no-flood. To prevent this, a weighted random sample was taken, where each data point is assigned a weight that is inversely proportional to the frequency of its class in the dataset. This grouping strategy prevents spatial and temporal leakage between training and evaluation sets, ensuring that the model is evaluated on truly unseen spatial-temporal conditions.

The effective LSTM-model requires a timeseries of dynamical features. The LSTM produces a hidden representation for each timestep. An attention layer subsequently learns which timesteps are most informative and computes a weighted combination of these hidden representations. The dynamical data is combined with the static features. The model uses a multi-task learning (MTL) approach to maximize predictive power. MTL is a strategy in machine learning where a model is trained to solve multiple problems with one shared neural network. In ‘normal’ machine learning methods, the model optimises one metric. MTL forces the model to find a representation that fits multiple goals. In this case, there are two tasks: one that has to classify whether there will occur a flood or not (binary) and one regression that tells which fraction of a pixel is flooded. During the training step, the losses of both tasks are summed. In this way, the LSTM is forced to look for patterns that minimise both error margins.

The main reason why the MTL was implemented, is because floods are rare events. A binary signal (flood or not a flood) is easier to learn. An additional benefit of having two tasks is that the regression takes information from the classification. The classification guides the regression to the flooded areas. Finally, having two tasks prevents overfitting of the model. Since two patterns have to be identified, the final model is likely to be more robust.

Model tuning

The architecture of the model as described is the result of an iterative process of trial-and-error. In this section, a brief description of the most important tests and experiments is provided.

The initial **loss function** for the classification head was a simple Binary Cross-Entropy (BCE) loss function. A BCE predicts the distance between the predicted probabilities and the observed binary labels (dry or flooded). This BCE was replaced by a focal loss function, which reduces the relative weight of easy-to-classify samples (always dry pixels) during training, allowing the model to learn more from the more difficult samples (pixels that sometimes flood). For the regression header, a Gaussian negative log-likelihood was used, where the model predicts both an expected value and an uncertainty per sample (Nix and Weigend, 1994). This approach was chosen to enable the model to predict extreme flooding events with more confidence, rather than systematically predicting the

average. After a first implementation of this Gaussian negative log-likelihood loss function, it appeared that the predicted variance exploded due to numerical instability. In order to reduce that artifact, the log-variance was limited to a fixed interval. Another test that was performed, was to perform a logit transformation on the flooded fraction. The idea was that this would make the distribution closer to the Gaussian distribution, since that is not limited between 0 and 1. However, it significantly reduced the regression performance (R^2 reduced from -0.18 to -1.48). That was possibly caused by small errors in logit space that are exaggerated in the original fractional scale after backtransformation by using the Sigmoid function. Therefore, this transformation was not implemented and the regression head works directly on the original values between 0 and 1.

The **splitting of the data** in training, validation and test data was performed, as earlier described, based on time. All years between 2015 and 2021 are used for training data, the year 2022 is used for evaluation and the years 2023 and 2024 are set aside as an independent test set. Originally, a split based on a tile-month combination was performed. This however caused independency problems for flood events that cross the boundary of the month, certainly since a period of 30 days was taken into account. To limit calculation time and RAM storage, not the entire validation and test set were used, but a representative sample of 200 000 pixel-day combinations. This sample is drawn proportionally over the entire available timeframe, so that the class distribution in the sample corresponds to the actual, global distribution. Based on this, it was determined that approximately 19% of the pixel-day combinations in the validation and test sets are classified as flooded, which was used as the global reference prevalence throughout the rest of the study. Note that this is not the prevalence in the observations, as a sample was already drawn earlier when creating the data frame.

During the training, an **evaluation** has to be performed in order to determine if the training can be stopped and to prevent overtraining. In a first phase, the F1-score (which is a combination of precision and recall) was used. However, due to the large class imbalance, the F1-score results in a suboptimal configuration since the F1 neglects the correct identification of the non-flooded pixels (true negatives). Therefore, a model that classifies nearly all pixels as flooded can be selected. As a better alternative, the Youden's J statistic is calculated. Youden's J is defined as the difference between the true positive rate (the proportion of observed floods the model correctly simulates) and the false positive rate (the frequency at which the model incorrectly identifies a dry pixel as flooded).

In the original training setup, for computational and memory reasons, a limited number of tiles were randomly selected and fully loaded for each training period. As a result, the model was exposed to only a fraction of the global climatic variability during the entire training. The **data loading** infrastructure was adapted such that a parallel reading is performed where only the sampled pixels were loaded per tile instead of the entire tile. That allowed that pixels from 150 tiles can be used instead of only pixels from 20 tiles. The loaded data were also reused across multiple consecutive training epochs within the same period, instead of loading a new selection for each epoch. This change yielded the greatest performance improvement in the entire development trajectory, which confirms that the spatial diversity of the training data is crucial for the model performance. Ideally, pixels from all 516 tiles would be used, but this is computationally unfeasible. An overview of the most important metrics is given in table 3.1. A clear improvement in all metrics can be observed.

Table 3.1: Comparison of the ROC-AUC, PR-AUC, R^2 and Pearson r between the two data loading strategies.

Metric	20 tiles per period	150 tiles per period
ROC-AUC	0.636	0.764
PR-AUC	0.274	0.406
R^2	-0.145	0.038
Pearson r	0.09	0.219

Hyperparameters are predefined variables in a machine learning model that drive the learning process of a model. An overview of the most important hyperparameters in this model is provided here. The learning rate determines the step size taken during optimization to update model weights. The class ratio is the percentage of the pixels in the training set that is taken flooded. The time window is the amount of days that is kept in memory. The regression loss can be weighted, where more heavy floods are penalised more than small floods. Regularisation, in the form of weight decay, adds a penalty proportional to the magnitude of the model weights to the loss function, discouraging overly large weights and therefore reducing the risk of overfitting. Model depth refers to the number of LSTM layers, where each additional layer allows the model to capture increasingly abstract temporal patterns. Model width refers to the number of hidden units per LSTM layer, which determines the amount of information the model can retain at each time step. The classification loss function determines how the discrepancy between predicted and observed flood occurrence is penalised during training. Besides focal loss, a weighted BCE loss, where a fixed class weight compensates for the imbalance between flooded and non-flooded pixels, was tested as an alternative. The focal loss parameter α is the class-balancing hyperparameter that addresses the raw frequency imbalance between classes.

To investigate the sensitivity of the model to different design choices and to avoid the reported configuration performing well only by chance, some tests were performed where one hyperparameter was changed. These models were evaluated separately based on ROC-AUC and PR-AUC. Table 3.2 gives an overview of the different runs and the resulting AUC-values.

Table 3.2: Summary of the performed sensitivity analysis on the hyperparameters.

Variant	Change	ROC-AUC	PR-AUC	Conclusion
Reference	/	0.636	0.259	/
Learning rate scheduler	Reduce learning rate on plateau	0.610	0.259	Slightly negative
Class ratio training	Increase from 15 % to 20 %	0.616	0.253	Slightly negative
Time window	Shorten from 30 to 15 days	0.612	0.256	Slightly negative
Weighted regression loss	Larger floods weighted more heavily	0.590	0.232	Clearly negative
Regularisation	Added weighted decay	0.628	0.276	Equivalent
Model depth	Increase LSTM layers from 2 to 3 layers	0.615	0.277	Slightly negative
Model width	Increase the number of hidden units from 128 to 192	0.516	0.208	Clearly negative
Classification loss	BCE with class weight instead of Focal Loss	0.547	0.211	Clearly negative
Focal loss parameter	$\alpha = 0.25$	0.623	0.273	Optimum at $\alpha = 0.5$
Focal loss parameter	$\alpha = 0.75$	0.610	0.249	Optimum at $\alpha = 0.5$

None of the tested alternatives yielded an improvement over the reference configuration. For all tested values of hyperparameters, barely any improvement was observed. That indicates that the model’s performance at that time was not limited by the values hyperparameters.

3.4 Evaluation

In order to evaluate the models, the independent test set is used to compare the model results with the observations. The model first classifies the pixels in flooded and non-flooded and later applies a regression for the fraction of the pixel that is flooded to all pixels that are classified as flooded. Both heads (classification and regression) have to be evaluated. For the classification, the pixels are grouped in true positive (TP) (observed and modelled to be flooded), true negatives (TN) (observed and modelled to be dry), false positives (observed to be dry, modelled to be flooded) and false negatives (FN) (observed to be flooded, modelled to be dry). Based on these numbers, a range of classification metrics can be calculated. For the regression, the observed flooded fraction in pixel i (y_i) is compared with the modelled fraction ($\hat{y}_i$). Table 3.3 contains an overview of the used metrics, their formula and their meaning.

Machine learning is often referred to as a black box: the user knows the in- and outputs, but has no insights in the logic and the decision making within the model. In order to get an idea on what the model drives and what its strengths and weaknesses are, the model results are compared over the different static attributes. The model residual is compared over the different environmental variables provided to the model. Secondly, a sensitivity analysis is performed, where a visual analysis is supported by a statistical test that indicates if the occurrence of pluvial floods is related to certain environmental conditions or not. All statistical tests are non-parametric because both flood occurrence and environmental variables violate the normality assumptions required for parametric tests. Moreover, the analyses involve skewed distributions and ordinal ranking rather than normally distributed residuals. For all tests, a significance level of $\alpha = 0.05$ is used. The p-value indicates if there is a difference, but not what the size of that difference is. Therefore, effect sizes are introduced. Given the large size of the test set, these effect sizes are extra important, since the p-value tends to rapidly decrease for large sample sizes (Lin et al., 2011).

To determine if the quality of the model depends on the environmental conditions, two different tests are performed. The Kruskal-Wallis test compares the distribution of the absolute model error across groups of a certain variable. Since we deal with continuous variables, they first have to be split in quantiles. As effect size, ϵ^2 , which is similar to R^2 in a regression, was used. It expresses which part of the variance between the groups is explained by the model and varies between 0 and 1. Secondly, the Spearman correlation test checks if there is a monotonic trend in the relation between the absolute error and the environmental variable. The Spearman correlation coefficient ρ is used both as test statistic and as effect size.

The Mann-Whitney U-test is used to check whether the distribution of an environmental variable (e.g. elevation) differs significantly between two independent groups. A t-test cannot be used in this case, since data such as elevation or soil characteristics are rarely normally distributed. The Mann-Whitney U-test is used for two different comparisons in this thesis: to compare the distribution of environmental variables between flooded and non-flooded pixels within both the observations and the model output. The effect size traditionally used together with the Mann-Whitney U-test is the rank biserial correlation (RBC), which measures the strength and direction of the relationship between a binary variable (flood or non-flood) and a ranked variable. This RBC varies between -1 and +1, where 0 indicates no relationship, while -1 or +1 indicates a total separation where all ranks in one group are lower or higher than the other.

Finally, the Wilcoxon signed-rank test was performed. Just like the Mann-Whitney

U-test, it compares the distribution of two groups, but now for groups that are not independent. We use the Wilcoxon signed-rank test to compare the distributions of the flooded pixel fraction for the observations and the model output. With the test, we check whether there is a systematic bias between the observed and predicted flood fraction within the same pixel. The effect size used is still called the same rank biserial correlation (RBC) as in the Mann-Whitney U-test, although it is now a variant suitable for matched pairs.

Table 3.4 provides an overview of all tests.

Table 3.4: Overview of statistical tests.

Statistical test	Data	H0	Statistic	Effect size
Kruskal-Wallis test	≥ 3 groups	Equal distributions	H	ϵ^2
Spearman correlation (r_s)	2 continuous variables	No monotonic relationship	ρ	ρ
Mann-Whitney U-test	2 independent groups	Equal distributions	U	RBC
Wilcoxon signed-rank test	2 dependent groups	Equal distributions	W	RBC

3.5 Climate attribution

The final goal of the model is to determine the impact of climate change on the occurrence and intensity of pluvial floods. Therefore, two model runs were performed: one with an observed precipitation dataset (factual) and one with the same precipitation dataset from which the effect of climate change is removed (counterfactual). The counterfactual dataset was created using the ATTRICI method, which removes the climate change signal from observed climate data while conserving daily variability and internal climate dynamics (Mengel et al., 2021). A generalized linear model (GLM) was used to quantify the relationship between precipitation and global mean temperature change. By setting the GMT anomaly to zero, the warming trend was removed, and quantile mapping was applied to translate observed precipitation values into counterfactual values while preserving their relative rank and spatio-temporal co-variability (Mengel et al., 2021). Comparing both on different aggregation levels (global, regional) indicates what the effect of climate change on pluvial floods is. Relative changes can be calculated as in equation 3.7.

$$\Delta = \frac{factual - counterfactual}{counterfactual} \cdot 100 \quad (3.7)$$

Table 3.3: Overview and definitions of the classification and regression evaluation metrics.

Metric	Formula	Meaning
Classification (Binary)		
Precision	$\frac{TP}{TP+FP}$	Proportion of predicted flooded pixels that are actual floods.
Recall	$\frac{TP}{TP+FN}$	Proportion of actual flooded pixels correctly detected by the model.
F1-score	$2 \times \frac{Precision \times Recall}{Precision + Recall}$	Harmonic mean of precision and recall, balancing both metrics.
Specificity	$\frac{TN}{TN+FP}$	Proportion of actual dry pixels correctly identified as dry.
Critical Success Index (CSI)	$\frac{TP}{TP+FN+FP}$	Ratio of correctly predicted floods to all pixels where a flood occurred or was predicted.
False Alarm Ratio (FAR)	$\frac{FP}{TP+FP}$	Proportion of predicted floods that turn out to be dry.
Cohen's Kappa	$\frac{p_o - p_e}{1 - p_e}$	Measures agreement between predictions and observations, corrected for chance agreement.
Matthews Correlation Coefficient (MCC)	$\frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$	Balanced metric for binary classification, robust against class imbalance.
ROC-AUC	$\int_0^1 TPR(FPR) dFPR$	Area under the Receiver Operating Characteristic curve: measures ability to distinguish between classes.
PR-AUC	$\int_0^1 Precision(Recall) dRecall$	Area under the Precision-Recall curve: focus on the flooded class performance.
Log-loss	$-\frac{1}{N} \sum_{i=1}^N [c_i \log(p_i) + (1 - c_i) \log(1 - p_i)]$	Penalizes incorrect predictions based on the probability of the prediction.
Brier Score	$\frac{1}{N} \sum_{i=1}^N (p_i - c_i)^2$	Mean squared error of the predicted probabilities compared to actual binary outcomes.
Regression (Flood Fraction)		
Mean Absolute Error (MAE)	$\frac{1}{N} \sum_{i=1}^N y_i - \hat{y}_i $	Average absolute magnitude of the prediction errors.
Root Mean Squared Error (RMSE)	$\sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2}$	Square root of the average squared errors: higher penalty on large errors.
Residual	$\frac{1}{N} \sum_{i=1}^N (\hat{y}_i - y_i)$	Measures systemic overestimation (positive) or underestimation (negative) by the model.
Coefficient of Determination (R^2)	$1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}$	Proportion of variance in the dependent variable predictable from the independent variable.
Pearson r	$\frac{\sum (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum (y_i - \bar{y})^2 \sum (\hat{y}_i - \bar{\hat{y}})^2}}$	Measures the linear correlation between actual and predicted flood fractions.

TP : True Positives; FP : False Positives; TN : True Negatives; FN : False Negatives; p_o : observed agreement; p_e : expected agreement;

TPR : True Positive Rate; FPR : False Positive Rate; N : total number of pixels; c_i : actual binary class (1 or 0) for pixel i ;

y_i : observed fraction for pixel i ; $\hat{y}_i$: predicted fraction; $\bar{y}$: mean of observed fractions; $\bar{\hat{y}}$: mean of predicted fractions

p_i : predicted probability for pixel i

Chapter 4

Data

4.1 Observed floods

Global Flood Partnership

The Global Flood Partnership (GFP) is a global network that brings experimental and operational flood-related products together to support early warning, scenario analysis, hydrological monitoring, and impact assessment. Its portfolio ranges from meteorological and hydrological forecasting systems to model-based and satellite-based flood mapping. Within the early-warning component, systems such as the Global Flood Awareness System (GloFAS) provide ensemble streamflow forecasts up to 30 days ahead, indicating the probability of exceeding flood thresholds. These forecasts are also used to plan satellite acquisitions in advance. Scenario analysis builds on these predictions by translating them into potential inundation extents, either through hydrodynamic modelling (e.g., JRC inundation maps) or satellite-based flood scenarios, including historical and probabilistic flood extent datasets. Hydrological monitoring products, such as the Global Flood Monitoring System (GFMS), provide short-range flood predictions based on precipitation forecasts and routing models. During flood events, GFP also delivers near-real-time satellite-derived inundation maps and river discharge estimates (Alfieri et al., 2018).

Remote sensing observations

Within this broader GFP framework, the Global Flood Monitoring (GFM) system provides the operational, satellite-based flood extent information used in this thesis. GFM is part of the Copernicus Emergency Management Service. GFM uses a combination of algorithms to analyse satellite imagery from the Sentinel-1 constellation: the LIST algorithm (Chini et al., 2017), the DLR algorithm (Martinis et al., 2015) and the TUW algorithm (Bauer-Marschallinger et al., 2022). Sentinel-1 are two polar-orbiting satellites perform SAR measurements day and night, providing accurate images for flood detection even in poor weather (Torres et al., 2012). The VV-polarisation of the images was used, since this has a high sensitivity for water surfaces (Schumann and Moller, 2015). In total three algorithms are applied to the Sentinel 1 images. The ensemble of these three algorithms was used as consensus map. The downloaded output of GFM is based on this ensemble (Copernicus Emergency Management Service, 2025).

Splitting pluvial and fluvial floods

The flood observations in the GFM dataset contain floods of all types (coastal floods, river floods and pluvial floods). Since the model has to be trained only on the observed pluvial floods, a pre-processing step needed to be taken to split up the dataset into a pluvial and a fluvial dataset. Therefore, the global river flood hazard maps of the Joint Research Catalogue (JRC) of the European Union are used. For seven different return periods, ranging from T10 to T500, they developed inundation maps with the water depth for the entire globe (excluding Greenland, Antarctica and some smaller islands). The hydrodynamic model used was LISTFLOOD-FP, while the hydrological model LISTFLOOD was used to create the inputs (Baugh et al., 2024).

For the purpose of this thesis, a water depth of at least 10 cm on the inundation map of the 100 year return period was chosen as a mask for the areas where floods are assigned to be fluvial. Floods occurring outside these mask are marked as pluvial floods. This approach is relatively simplistic and neglects pluvial floods in floodplains. Especially along rivers crossing cities, some pluvial floods will be neglected. The 10 cm threshold was introduced to avoid noise on the training data.

4.2 Static data

Elevation

The topographic information used in this thesis is based on the Copernicus DEM, which is a digital surface model (DSM) of the globe. It therefore also contains buildings, vegetation, etc. The Copernicus DEM is based on the TanDEM-X mission between 2011 and 2015. The DEM was published in 2016 (Wessel et al., 2018). TanDEM-X, standing for TerraSAR-X add-on for Digital Elevation Measurement, uses interferometric synthetic aperture radar (InSAR) to measure the elevation of the earth surface. This active technique uses the differences in phase between two radar signals to derive the elevation differences. Therefore, two satellite twins (TerraSAR-X en TanDEM-X) are used at 250 m separate from each other. Since both satellites measure the same location nearly simultaneously, weather and moving objects are approximately the same and the accuracy of the depth measurements increases. This technique is referred to as single-pass interferometry. A bistatic mode was used, which means that one satellite uses a radar to send the signal and both satellites measure the reflected signal. Because both satellites receive exactly the same signal, altitude differences can be derived from the phase differences between the two radar images (Krieger et al., 2007).

The global Copernicus DEM is freely available at 30 and 90 m resolution (European Space Agency, 2025). The raw radar data is then processed into an elevation map via various steps, such as precisely aligning the images, calculating phase differences, and converting that phase into elevation. Subsequently, all segments are combined into a single global DEM. The final TanDEM-X DEM complies with the international HRTI-3 standard. This means that the absolute height error is less than 10 meters (with 90% certainty) and that the relative error lies between 2 and 4 meters, depending on the steepness of the terrain (Krieger et al., 2007). Wessel et al. (2018) found a very low average error (+0.17m based on kinematic GPS tracks and +0.12m GPS bench marks) and a very high absolute accuracy ($RMSE < 1.4m$ globally). For low vegetation this accuracy is around 1.1m, while urban areas (1.4m) and forests (1.8m) have a higher

RMSE. For slopes below 10° , the 90% confidence interval (LE90) gives a value of 2.6m, while steeper slopes have a LE90 value of 5.5m.

Land cover

The Copernicus Land Monitoring Service (CLMS) is part of the Copernicus-program of the European Union that provides geographical information about land cover and land use and the changes in that. Besides that, also some biophysical variables (like vegetation status or energy fluxes) are provided. The Global Dynamic Land Cover is one of the products from CLMS. Between 2015 and 2019, an annual global land cover map was provided at 100 m resolution at three classification levels. For the year 2020, a detailed global land cover map with a resolution of 10 m is available. This map will be used in this thesis. The map has an overall accuracy of $80.10 \pm 0.20\%$ (European Union’s Copernicus Land Monitoring Service (CLMS), 2026). The accuracy fluctuates largely between different landcover classes. Table 4.1 contains an overview of the producer accuracy (PA) and the user accuracy (UA) for all landcover classes. While the PA measures how well a certain land cover type is classified by CLMS, the UA indicates the probability that a pixel on the map effectively represents the landcover that is observed in the real world. Classes like tree cover, mangroves, water bodies and snow and ice are identified with relatively high accuracy, while crop- and grassland and built-up area are much more difficult to classify correctly.

Table 4.1: Producer and user accuracy of the CLMS map in 2020 (European Union’s Copernicus Land Monitoring Service (CLMS), 2026)

Landcover	PA(%)	UA(%)
Tree cover	90.90	83.39
Shrubland	67.90	56.39
Grassland	64.53	82.44
Cropland	86.22	87.04
Herbaceous wetland	73.62	77.20
Mangroves	100.00	100.00
Moss and lichen	83.66	61.34
Bare/sparse vegetation	86.79	89.33
Built-up	77.83	59.87
Permanent water bodies	92.23	89.13
Snow and ice	99.92	85.64

The land use map contains 11 classes of land use. The amount of classes determines to a large extent the degrees of freedom when training the model. Therefore, reducing the number of classes will reduce the degrees of freedom and the ambiguity, so it also reduces the required amount of observations. All classes where natural or managed vegetation with short roots or where no roots occur (shrubland, grassland, cropland, moss and lichen and bare/sparse vegetation), are reclassified under the name "short vegetation". The class 'tree cover' is kept as it was, and will be referred to as "forest". The classes of 'herbaceous wetlands' and 'mangroves' are merged into the class "wetland". Finally, the built-up class was kept as it was and is referred to as "paved". The classes 'snow and ice', 'unclassifiable' and 'no data' are neglected. The class of 'permanent water bodies' was neglected as well, since this class was used as mask in the flood detection algorithm. For the four new classes, a binary map was produced on the original resolution of 10m. These maps were

rescaled and retiled, forming a separate map for each class containing the percentage of a pixel taken by that land use.

Soils

The information on the soils is based on the Harmonized World Soil Database (HWSD) by the Food and Agriculture Organisation (FAO) and the Applied System Analysis (IIASA). The soil map has a spatial resolution of approximately 1 km. The original database was established in 2008, but the renewed version of 2023 was used (FAO et al., 2012). The HWSD is a harmonized compilation of regional and national soil databases, standardized into a common classification and attribute system (Nachtergaele et al., 2010). Since the HWSD is a merge of different regional and continental soil maps, the accuracy differs largely between regions. The HWSD contains a raster file and an attribute database containing soil parameters for each pixel. The spatial resolution of 1km might be considered as relatively coarse for soil characteristics.

Three variables were taken from the HWSD, namely the bulk density, the percentage sand and the percentage clay. The bulk density is the mass of the soil particles divided by the volume they occupy. The HWSD used the USDA classification for soil textures, making clay the fraction of the soil with a grain diameter less than 0.002 mm and sand the fraction with a grain diameter between 0.05 and 2.00 mm (FAO et al., 2008). For upscaling, the average value of this three parameters was used.

4.3 Precipitation

Historic precipitation

The historic precipitation data is gathered from the CHELSA-daily dataset. CHELSA daily contains global climate data on kilometer scale since 1979. Variables included in CHELSA daily are total cloud cover, near-surface relative humidity, precipitation, surface air pressure, surface downwelling shortwave flux, near-surface wind speed, daily mean, maximum and minimum near-surface air temperature and air temperature lapse rate. In this thesis, only the precipitation will be used.

The global, high resolution dataset of meteorologic variables was made by using a topographic informed downscaling algorithm called CHELSA V2.1. The original datasets are too coarse to capture local processes such as convection. Conversely, a dynamic downscaling would be too computational expensive. CHELSA V2.1 uses W5E5 as input, which is the observational reference dataset used in ISIMIP3. It is a daily, 0.5° global dataset that combines WFDE5 over land and ERA5 over the ocean (Karger et al., 2023). ERA5 is a climate reanalysis dataset produced by the ECMWF. ERA5 has an hourly output on a horizontal resolution of 31 km at 137 model levels up to 1 hPa (Hersbach et al., 2020). WFDE5 is the bias-corrected ERA5 dataset. CHELSA V2.1 uses a 30 acsec elevation map (GMTED2010) and a land-ocean masking (GSHHG and MOA) to downscale the W5E5 dataset. Additionally, lapse rates, wind fields and cloud fractions were taken from ERA5 to improve the downscaling (Karger et al., 2023). The precipitation has been downscaled by using an orographic correction by means of a windward/leeward index and boundary layer height. Wind fields are used to estimate moisture transport and uplift, while boundary layer height corrections account for atmospheric stability. Valley

exposition is considered to capture rain shadow effects. Within each cell, a flux-conserving redistribution has been applied (Karger and Zimmermann, 2021).

CHELSA-W5E5 systematically improves the performance of W5E5 for temperature, precipitation, and radiation, especially in mountainous terrain. For the precipitation specifically, the global bias reduces slightly. Mainly in orographic areas, the bias reduction is significant. On flat terrain, there is sometimes an improvement, but smaller than in more elevated areas. The main limitation lays in the convective precipitation, which does not improve compared to the W5E5 dataset. This is because only the windward/leeward index and the boundary layer height are used to downscale. Consequently, no convection is included, making the tropics and flat terrain to remain a difficult area. The extremes (95-percentile) increase in mountainous areas, but in flat areas there is a decrease (Karger et al., 2023).

Counterfactual precipitation

In order to assess the impact of climate change on the occurrence of pluvial floods, a comparison has to be made between the current climate and the theoretical climate for the same period if there would have been no climate change. Therefore, the precipitation data therefore have to be detrended to remove the impact of climate change on the precipitation. Such datasets are referred to as 'counterfactual' or 'detrended'. The W5E5 dataset contains a counterfactual dataset, but this is not yet downscaled by the CHELSA V2.1 algorithm. Therefore, the proper analysis can currently not be performed. The evaluation of the models with the CHELSA forcing are therefore used as a reference for which accuracy potentially be reached.

As a temporary solution, the precipitation is replaced by a counterfactual forcing at a resolution of 0.1° . However, the model still runs at the 30 arcsec resolution. Therefore, the 0.1° data is downscaled by a very simple nearest neighbour algorithm, where the value of each 30 arcsec pixel is taken equal to the value of the closest pixel of 0.1° resolution. The data at 0.1° resolution exist for the W5E5 dataset, but to limit data storage, the ERA5 dataset is used, for which the data were already downloaded. It is a combination of a numerical weather model (ECMWF IFS Cy41r2), a large amount of observations (radiosondes, satellites, GNSS-RO, etc.) and an advanced data assimilation system (Hersbach et al., 2020).

4.4 Historical fluxes

As described earlier, the first two versions of the model (an ANN and a LSTM) are, besides by the precipitation, also fed by the evapotranspiration and the soil moisture. The information for these fluxes comes from the GLEAM model.

The Global Land Evaporation Amsterdam Model (GLEAM) is a model that describes the different components of the evapotranspiration based on satellite observations. The core of GLEAM is the Penmann equation, which calculates the potential evaporation based on satellite observations of the air temperature at the surface, the leaf-area index (LAI), the wind speed and the vapour pressure deficit. To make the step from potential to actual evaporation for bare soils and transpiration by plants, GLEAM uses a stress factor which is determined by using machine learning, trained on global measurements of eddy-covariance and sap flow. By using a waterbalance, GLEAM is able to output not only the

different parts of the evapotranspiration, but it also calculates the soil moisture within the root zone. To minimise errors on this water balance, satellite images of the surface soil moisture are included. Apart from this, three specific fluxes are computed as well: the interception loss, the evaporation from open water and sublimation and evaporation from snow and ice (Global Land Evaporation Amsterdam Model, 2025).

In this thesis, the total evaporation will be used. This is the sum of the transpiration by plants, the evaporation from bare soil, the interception loss, snow and ice sublimation and open water evaporation. Also the root-zone soil moisture will be included in the model. Although this does not entirely captures the total storage within a soil, it will be used as an approximation for that.

Chapter 5

Results and discussion

5.1 Evaluation

Model performance comparison

Table 5.1 contains the evaluation metrics for the four different models.

Table 5.1: Performance evaluation of four machine learning architectures for spatial flood mapping across binary classification and continuous flood fraction regression metrics.

Metric	Fluxes-based ANN	CHELSEA-based LSTM	ERA5-based LSTM	Fluxes-based LSTM
Classification Metrics				
Precision	0.2897	0.2527	0.2435	0.3114
Recall	0.8340	0.7727	0.8530	0.8176
F1-score	0.4300	0.3809	0.3788	0.4510
Specificity	0.5169	0.4601	0.3738	0.5728
Critical Success Index (CSI)	0.2739	0.2352	0.2337	0.2911
False Alarm Ratio (FAR)	0.7103	0.7473	0.7565	0.6886
Cohen's Kappa	0.2043	0.1304	0.1159	0.2408
Matthews Correlation Coefficient (MCC)	0.2773	0.1858	0.1895	0.3069
ROC-AUC	0.7492	0.6706	0.6635	0.7636
PR-AUC	0.4193	0.3207	0.2877	0.4064
Log-loss	0.4629	0.4880	0.4976	0.4509
Brier Score	0.1462	0.1558	0.1601	0.1426
Regression Metrics				
Mean Absolute Error (MAE)	0.1514	0.1491	0.1526	0.1500
Root Mean Squared Error (RMSE)	0.2008	0.2050	0.2072	0.2014
Residual	-0.0064	-0.0256	-0.0180	-0.0120
Coefficient of Determination (R^2)	0.0440	0.0034	-0.0187	0.0376
Pearson r	0.2157	0.1733	0.1328	0.2193

The best model created has a F1 of 0.4510, a MCC of 0.3069, a PR-AUC of 0.4064 and a R^2 of 0.0376. These metrics suggest the general performance of the model is rather modest and results from the model should be interpreted with caution. However, it is clear the model does learn some patterns and has a certain predictive power. Given the limited amount of information provided (8 static and 3 dynamic variables), the model can be considered as relatively simple. The current performance and the increased accuracy when adding the fluxes supports a large potential for future improvements.

The fluxes-based LSTM is the best model on almost all classification metrics (highest precision, F1 score, specificity, CSI, MCC, and ROC-AUC). Memory helps to define the flood class. For the classification, the choice to increase the complexity from an ANN to a LSTM is justified. Interestingly, the fluxes-based ANN achieves slightly better figures on the regression metrics than the fluxes-based LSTM ($R^2 = 0.0440$ vs 0.0376). The ANN is slightly less biased in terms of continuous fraction forecasting.

Adding the evapotranspiration and the soil moisture increases the model accuracy. All classification metrics point in the direction of an improved model. The F1 increase from 0.3809 to 0.4510, the CSI from 0.2352 to 0.2911 and the ROC-AUC from 0.6706 to 0.7636. In general, also the regression head gained performance. R^2 increased from 0.0034 to 0.0376. Adding information about these water fluxes significantly increases the accuracy, suggesting they play a crucial role in the occurrence of pluvial floods.

The loss in accuracy caused by replacing the CHELSA-downscaled precipitation by a nearest neighbour approach is overall very limited. F1 only decreases from 0.3809 to 0.3788, and the recall even increased (from 0.7727 to 0.8530). Also the MCC scores slightly better for the ERA5 variant (0.1895 instead of 0.1858), although everything remains within the noise of a single training session, and no strong conclusion can be drawn from it. However, it shows that the decrease in accuracy is not the same for all metrics. The decrease in spatial resolution of the input data seems to primarily affect the regression, more than the classification. The R^2 is slightly negative, making that predicting the average flooded fraction would overall have a slightly better result than the current model.

The model accuracy was expected to decrease when the CHELSA-input was replaced by the nearest neighbour downscaling of the ERA5 data. However, the difference in performance is relatively limited, indicating that the simplified model can be used to perform a climate attribution.

Model residual analysis

As the model based on the ERA5 precipitation data is the model used for the climate attribution analysis, this is the model that is used in the rest of this thesis.

In figure 5.1, the absolute residual is plotted on the y-axis against the observed flood fraction on the x-axis using a hexbin density plot.

A clear negative linear trend is visible in the residuals relative to the observed values. At low observed flood fractions (< 0.2), the residuals are positive. The model overestimates the flood when the observed flood is small. At high observed flood fractions (> 0.4), on the other hand, the residuals are negative. The model therefore significantly underestimates the flood when a major flood is observed. The model tends towards moderate or low predictions to minimize the global loss function. Pluvial floods depend strongly on local microtopography and extreme precipitation (convective showers). If the input data (e.g. DEM or precipitation data) is scaled too coarsely, the model smooths out the extremes. The orange coloring is located in the top-left corner at an observed fraction around 0.05 - 0.20. This confirms that the vast majority of the evaluation data consists of pixels with slight to moderate flooding, and that the model tends to slightly overestimate for those most frequent categories.

Flooded pixels are divided into 5 groups. The land cover groups are organized into bins per 25% cover of a pixel. For each group, the median of the variable is plotted against the mean residual of the model in figure 5.2.

The observed pattern for the fraction forest and the fraction paved are relatively similar: both show a very sharp peak for small fraction that are just above zero, but then flatten out to fairly small errors. At very low fractions of short vegetation, there is a strong underestimation. Around 0.7-0.8, the bias is almost zero, after which a slight underestimation occurs again at very high fractions. The residuals are negative across the entire range of the wetlands, which indicates that the model continuously underestimates flooding in wetlands.

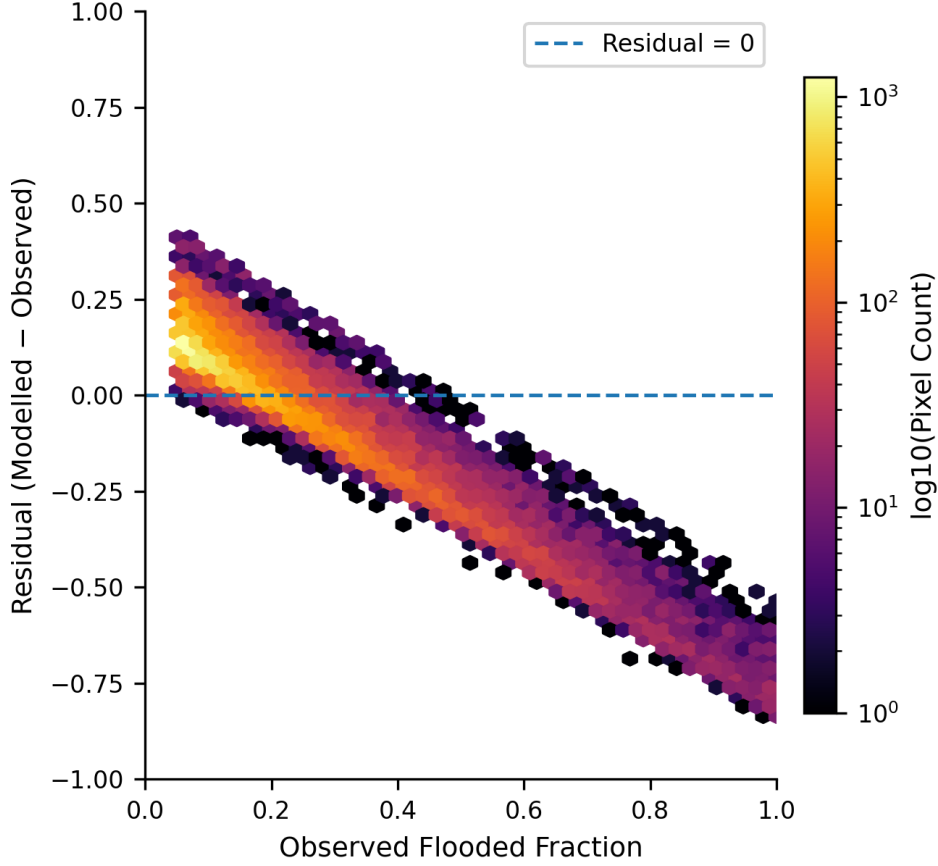


Figure 5.1: Residual distribution plotted against observed flood fraction for flooded pixels. The blue dashed line indicates zero residual error. Color intensity indicates pixel density on a $\log_{10}$ scale.

Figure 5.3 depicts the same as figure 5.2, but now for the elevation and the soil characteristics. For these variables, the bins are quintile-based.

At low heights ($< 200m$), the residual is close to 0 (with a small peak). As the height increases ($> 300m$), the residual drops and the model underestimates the flooded fraction. The 'expected' low-altitude pluvial flood is therefore modeled relatively well. The model struggles more with floods at higher altitudes.

For clay content, the line rises from negative (underestimation at low clay contents 5%) to positive (overestimation at high clay contents above 27%). With the sand content, we see the opposite: the line actually drops from positive (overestimation at low sand content below 27%) to strongly negative (underestimation at high sand content above 60%). Due to the high infiltration capacity of sandy soils, fewer floods are expected in sandy soils. There are two situations that deviate from the logic: the model predicts significant flooding despite a high sand content, and a sandy soil did flood unexpectedly in reality. These two cases account for the downward curve. Clay soils, on the other hand, infiltrate hardly any water, meaning that the deviations from the logic are a dry prediction and a dry observation of clay soil. That causes the rising curve. These observations mean that the model understands the logic of infiltration capacity, but exaggerates the effect of soil texture on infiltration.

Bulk density shows a rather erratic pattern, but with a clear underestimation at low

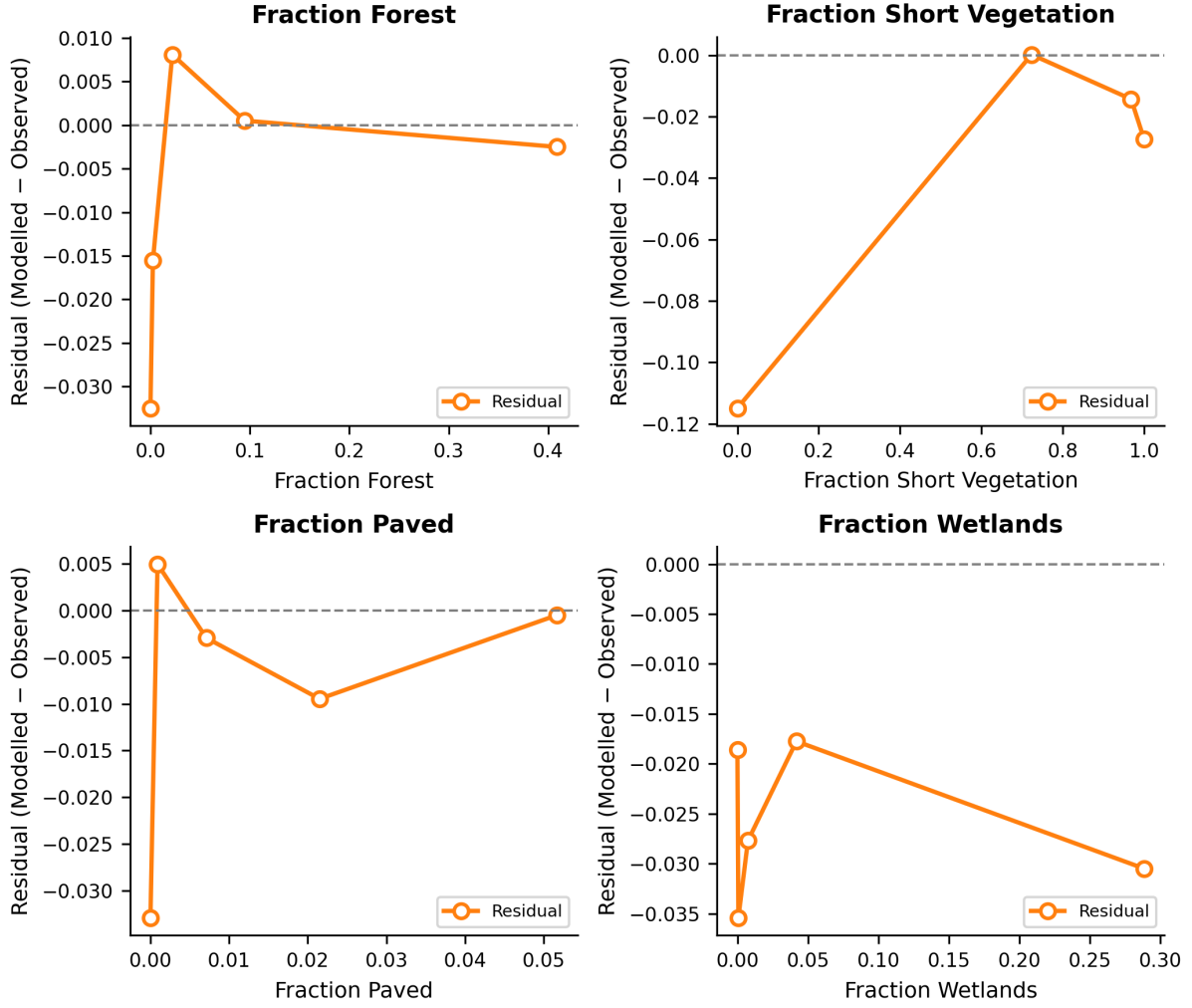


Figure 5.2: Mean regression residual evaluated across binned intervals of landcover fractions for flooded pixels. Dashed horizontal lines denote zero error.

density (< 0.8) and at a density around $1.55g/cm^3$. Low bulk density is often associated with organic soils or forest soils with high porosity. The erratic nature suggests that the relationship between soil density and the flood fraction is non-linear and possibly interacts with land cover.

To investigate if the absolute model error differs significantly between different environmental conditions, the Kruskal-Wallis test is used. Table 5.2 shows the results of the test.

Model performance varies across all environmental conditions, but in general the ϵ^2 -values are small. The model performance differs the most (but still limited) with the forest cover ($\epsilon^2 = 0.0343$), the percentage clay ($\epsilon^2 = 0.0250$), the fraction of short vegetation ($\epsilon^2 = 0.0227$) and the elevation ($\epsilon^2 = 0.0213$). The forest fraction therefore explains 3.4% of the variation in the absolute error. Consequently, the model's margin of error fluctuates most strongly in forested areas. All other variables have barely any influence on the model performance, despite being all of them significant concerning the p-value.

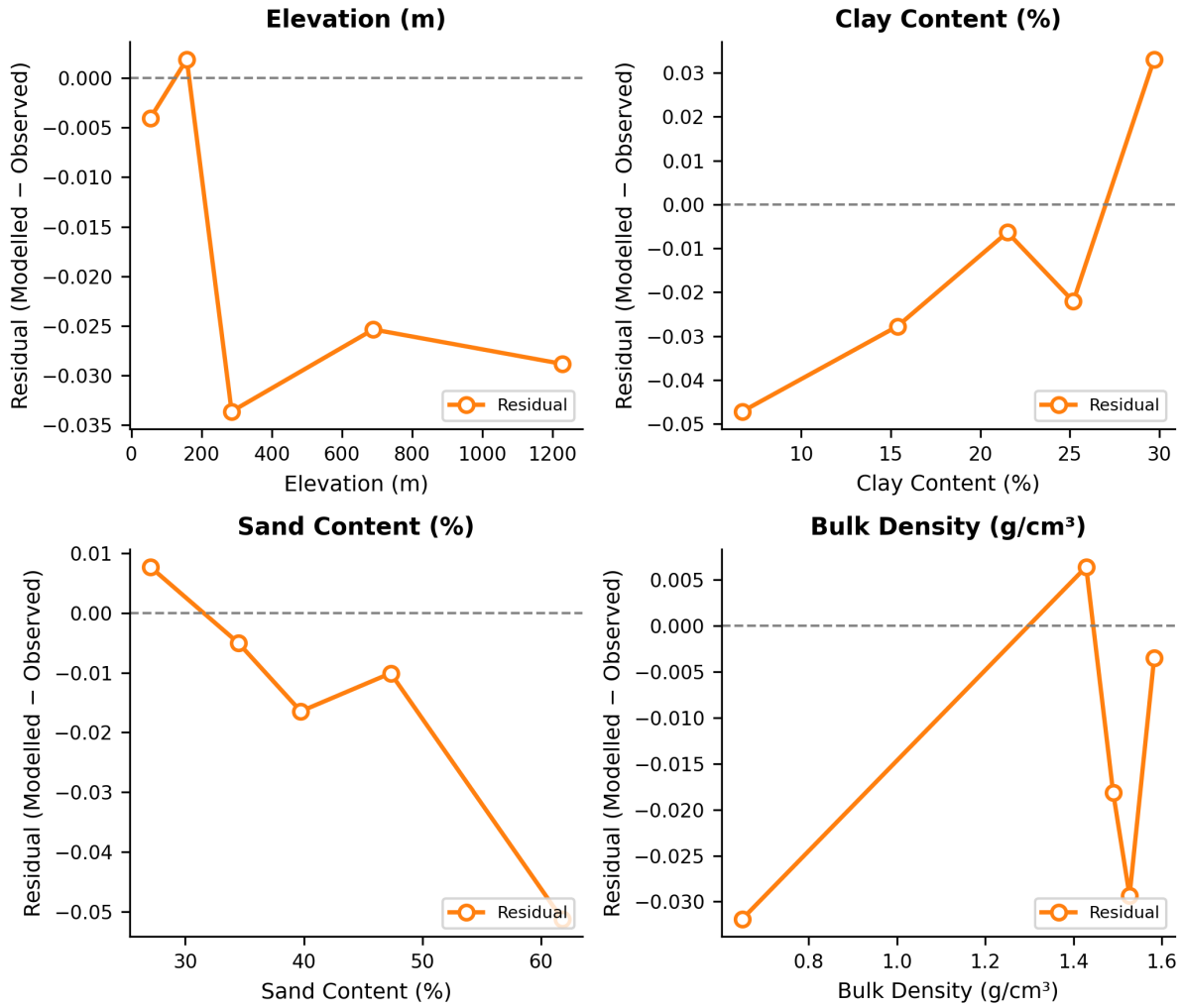


Figure 5.3: Mean regression residual evaluated across binned intervals of environmental variables for flooded pixels. Dashed horizontal lines denote zero error.

Table 5.3: Spearman rank correlation (ρ) between continuous environmental variables and the absolute prediction error in flooded pixels.

Variable	Spearman ρ	p-value
Elevation	-0.1334	$< 10^{-16}$
Percentage clay	0.1657	$< 10^{-16}$
Percentage sand	-0.0743	$< 10^{-16}$
Bulk density	0.0365	$5.492 \cdot 10^{-11}$
Fraction forest	-0.0695	$< 10^{-16}$
Fraction short vegetation	0.0541	$7.077 \cdot 10^{-16}$
Fraction paved	-0.0248	$2.012 \cdot 10^{-4}$
Fraction wetland	0.0054	0.4188

Table 5.2: Kruskal-Wallis test results evaluating the absolute prediction error across binned landcover fractions and terrain quintiles for flooded pixels.

Variable	H-statistic	p-value	Epsilon squared (ϵ^2)
Elevation	816.841	$< 10^{-16}$	0.0213
Percentage clay	898.040	$< 10^{-16}$	0.0250
Percentage sand	331.751	$< 10^{-16}$	0.0090
Bulk density	317.813	$< 10^{-16}$	0.0097
Fraction forest	774.243	$< 10^{-16}$	0.0343
Fraction short vegetation	507.426	$< 10^{-16}$	0.0227
Fraction paved	8.088	0.0442	0.0002
Fraction wetland	69.315	$3.166 \cdot 10^{-14}$	0.0029

The clay fraction shows the strongest linear relationship with the model error ($\rho = 1657$). The higher the percentage of clay in the soil, the greater the absolute error of the model. Elevation, on the other hand, shows a negative correlation, which means that lower-lying areas lead to larger model errors. With less sand ($\rho = -0.0743$) and less forest ($\rho = -0.0695$), the model error increases slightly. More short vegetation ($\rho = -0.0541$) and higher soil density ($\rho = -0.0365$), on the other hand, lead to a slightly larger model error. Wetland has no significant correlation with the size of the model error ($\rho = 0.4188$). Therefore, the amount of wetland present in the cell does not affect the accuracy of the prediction. An important note is that the amount of wetland pixels in the test set is limited. Although paved surface has a small ρ value due to the large N, the ρ of -0.0248 is negligible.

Sensitivity analysis

Figure 5.4 shows the observed flood frequency versus the flood frequency estimated by the model across the entire dataset (including dry pixels).

The observed incidence is lowest at sea level (around 0.01), rises steadily to a clear peak around 1000 - 1200 m (to about 0.30), and drops back below 0.2 at even higher altitudes. The model exhibits exactly the same dynamics at the beginning: a rise from sea level with a peak around 1000 m. Above 1000 m, however, the estimated probability remains stable on a high plateau (above 0.30) instead of decreasing along with it. The model recognizes that medium-sized mountain ranges carry an increased risk, but is too cautious above 1000 m by keeping the probability high.

The clay percentage shows a rather erratic pattern with a clear peak in flood probability around 25 - 28% clay, after which the probability decreases at very heavy clay soils ($> 30\%$). The sand percentage has a similar incidence with some fluctuations between 30 and 55% sand and a constant rise starting from 60% sand. The model overestimates across the entire range but follows nearly every dip and peak of the clay observation line, only more flattened. It also identifies the peak around 25% clay. The model has learned the non-linear relationship between clay content and flood occurrence well but overestimates flood occurrence. For the sand percentage, the model shows a much flatter line. The rise starting from a sand percentage of 60% is also not followed. The distribution for bulk density remains fairly stable between 0.15 and 0.2, but shows a striking sharp peak starting from $1.4g/cm^3$. The model structurally overestimates this (values just below

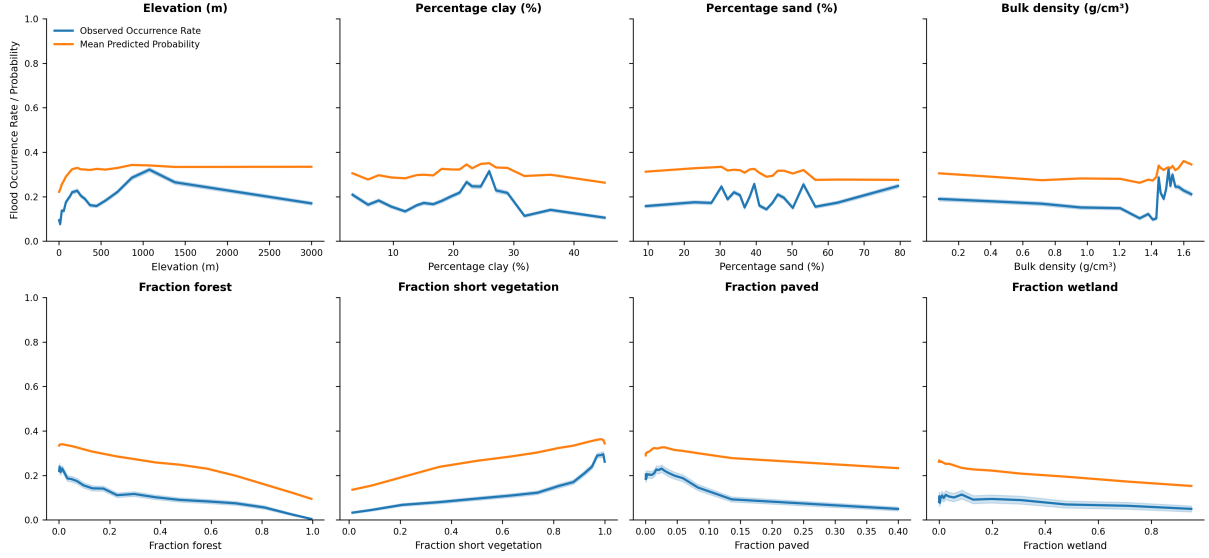


Figure 5.4: The classification evaluated across binned intervals of environmental variables for flooded pixels. The blue line depicts the observations and the orange line depicts the model output

0.3), but does pick up the peak. A high bulk density implies compacted soil. The model demonstrates here that it is highly sensitive to this specific physical soil threshold.

The forest fraction shows a steady, continuous decrease from approximately 0.25 at 0% forest to nearly 0 at 100% closed forest cover. The model overestimates, but does show a perfectly proportional decrease from approximately 0.35 to approximately 0.10. The short vegetation fraction gradually increases from approximately 0.05 at no low vegetation to 0.30 at 100% low vegetation. The model increases in exactly the same manner from approximately 0.15 to 0.35. Forests take up water (sponge effect) and reduce flood risks. Low vegetation offers less sponge effect than forest and is often found in flatter, flood-prone areas, especially agricultural land. The model simulates the hydrological damping effect of forests (and the absence thereof with short vegetation) well, although the overestimation remains. For the wetlands, the observations fluctuate around 0.10 across the entire range, while the model shows a gentle downward trend.

A surprising observation occurs in the paved fraction, where there is initially a small peak with very light paving, after which the incidence decreases as the paving increases to > 0.15 . The model responds identically with a peak and a gradual decline thereafter. That is surprising, because more flooding would actually be expected on paved surfaces. It should be noted that the fraction of paved surfaces nowhere exceeds 0.4. This could indicate that (large) cities, where pluvial flooding is a risk, are not present in the dataset. This potentially has to do with the way pluvial flooding was filtered out of all observed floods, namely using the river flood risk map. Since cities are often located in the floodplain of rivers, it is possible that many cities were filtered out.

Figure 5.5 shows the observed flooded pixel surface area fraction versus the predicted fraction, filtered exclusively for actually flooded pixels.

For the elevation, the lines lie almost directly on top of each other at a fraction between 0.20 and 0.25. For the percentage of clay, there is a very strong agreement around 0.20 - 0.25. Around 25 – 30% clay, the observations show a small jump to 0.30. The model picks up this jump but responds to it with just a slightly exaggerated effect, after which

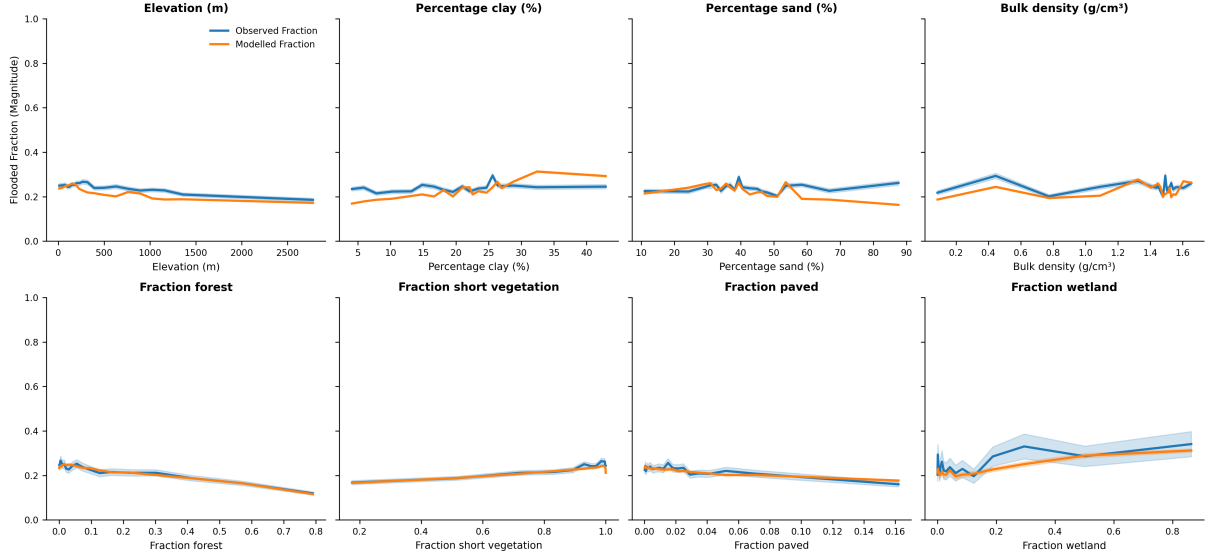


Figure 5.5: The flooded fraction evaluated across binned intervals of environmental variables for flooded pixel. The blue line depicts the observations and the orange line depicts the model output

it stabilizes. Also for the percentage sand, the overall trend is reflected very well. Only at sand percentages above 50%, the model has a slight underestimation. The curve is also very similar for bulk density. The observed data shows a peak at 0.4 g/cm³ and some fluctuations around 1.5 g/cm³. The model follows the average trend and avoids extreme outliers.

Regarding land cover, the forest, short vegetation, and paved fractions show excellent results, with the orange line falling almost entirely within the margin of uncertainty of the observations. As in figure 5.4, the forest fraction shows an increase in flooded fraction with the rising forest fraction and a decreasing trend with increasing short vegetation and an increasing paved fraction. However, these trends in the regression are much less pronounced than in the classification. The only panel with a clear deviation is that of the wetland fraction. With a low wetland fraction (< 0.15), the lines converge at approximately 0.20, but as soon as the wetland fraction goes above 0.2, the actual observed flood magnitude rises to 0.30 - 0.35, while the model rises more slowly towards 0.30.

Graph 5.4 shows that the LSTM model has a certain physical knowledge: it recognizes exactly where and under what conditions the risk is higher or lower (trends match perfectly). However, there is a constant over-prediction bias on the probability, resulting in high sensitivity. Figure 5.5, on the other hand, proves that the model is calibrated well for the magnitude. As soon as the flood is triggered, it predicts the actual surface fraction across virtually all environmental gradients rather good.

To check if there is a significant difference between the value of an environmental variable of the dry pixels and the flooded pixels, the Mann-Whitney U test results are summarised in table 5.4.

Table 5.4: Mann-Whitney U test results and Rank-Biserial Correlation (RBC) effect sizes comparing dry vs. flooded conditions for observed and predicted states.

Variable	Dataset	U-statistic	p-value	RBC
Elevation	Observed	$2.637 \cdot 10^9$	$< 10^{-16}$	0.1473
	Predicted	$2.379 \cdot 10^9$	$< 10^{-16}$	0.4623
Percentage clay	Observed	$2.621 \cdot 10^9$	$2.099 \cdot 10^{-10}$	0.0216
	Predicted	$3.447 \cdot 10^9$	$< 10^{-16}$	0.0706
Percentage sand	Observed	$2.755 \cdot 10^9$	$4.353 \cdot 10^{-4}$	0.0118
	Predicted	$4.642 \cdot 10^9$	$< 10^{-16}$	-0.1968
Bulk density	Observed	$1.793 \cdot 10^9$	$< 10^{-16}$	0.1378
	Predicted	$2.048 \cdot 10^9$	$< 10^{-16}$	0.2755
Fraction forest	Observed	$1.412 \cdot 10^9$	$< 10^{-16}$	-0.2621
	Predicted	$2.587 \cdot 10^9$	$< 10^{-16}$	-0.5543
Fraction short vegetation	Observed	$7.555 \cdot 10^8$	$< 10^{-16}$	0.3130
	Predicted	$2.982 \cdot 10^8$	$< 10^{-16}$	0.8172
Fraction paved	Observed	$1.118 \cdot 10^9$	0.4715	0.0027
	Predicted	$1.600 \cdot 10^9$	$2.803 \cdot 10^{-45}$	0.0440
Fraction wetland	Observed	$1.249 \cdot 10^9$	$< 10^{-16}$	-0.1003
	Predicted	$2.136 \cdot 10^9$	$< 10^{-16}$	-0.2639

For nearly all variables, the sign of the effect size (RBC) is equal for the observations and the predictions, indicating that the model recognizes the right directions. Both in reality and in the model, flooded areas have for example a lower elevation, less short vegetation, and more forest and wetland. Only for the percentage sand in the subsoil, the sign is opposite: the model predicts more floods on a sandy soil, while in reality, floods occur a slight bit less when the sand percentage increases. The model strongly amplifies the influence of the variables, as was already observed in the previous figures. The RBC of short vegetation for example increases from 0.3130 in the observations to 0.8172 in the model. The presence of low vegetation is a very strong indicator for the model that a cell must be dry. Elevation (RBC rises from 0.147 to 0.462) and forest cover (RBC rises from -0.262 to -0.554) are also weighted more than twice as heavily. These factors were in table 5.2 indicated as the factors with the highest (but limited) impact on the model error.

Table 5.5 contains again the test results of a Mann-Whitney U test, but now not to compare the dry and the flooded pixels on a certain variable. Instead, we compare the observed and the modelled pixels (dry vs dry and flooded vs flooded). The goal is to determine whether the environmental profile of a modeled location is comparable to that of the locations that actually are flooded.

Table 5.5: Mann-Whitney U test results and Rank-Biserial Correlation (RBC) effect sizes comparing observed vs. predicted conditions within the flooded and dry subsets.

Variable	Class Evaluated	U-statistic	p-value	RBC
Elevation	Flooded	$2.474 \cdot 10^9$	$1.159 \cdot 10^{-23}$	0.0336
	Dry	$6.850 \cdot 10^9$	$< 10^{-16}$	-0.2814
Percentage clay	Flooded	$2.271 \cdot 10^9$	0.3243	0.0034
	Dry	$4.555 \cdot 10^9$	$< 10^{-16}$	-0.0450
Percentage sand	Flooded	$2.533 \cdot 10^9$	$< 10^{-16}$	-0.0712
	Dry	$3.945 \cdot 10^9$	$< 10^{-16}$	0.1374
Bulk density	Flooded	$1.810 \cdot 10^9$	$1.656 \cdot 10^{-10}$	-0.0234
	Dry	$3.857 \cdot 10^9$	$< 10^{-16}$	-0.1605
Fraction forest	Flooded	$8.731 \cdot 10^8$	$6.159 \cdot 10^{-24}$	0.0405
	Dry	$1.380 \cdot 10^9$	$< 10^{-16}$	0.3256
Fraction short vegetation	Flooded	$9.008 \cdot 10^8$	0.5971	-0.0023
	Dry	$2.987 \cdot 10^9$	$< 10^{-16}$	-0.4963
Fraction paved	Flooded	$8.973 \cdot 10^8$	$5.503 \cdot 10^{-4}$	0.0136
	Dry	$2.120 \cdot 10^9$	$3.615 \cdot 10^{-21}$	-0.0283
Fraction wetland	Flooded	$9.288 \cdot 10^8$	0.0115	-0.0052
	Dry	$1.749 \cdot 10^9$	$< 10^{-16}$	0.1576

The Flooded class is relatively accurate. For almost all variables, the RBC is extremely close to 0. For clay ($p = 0.3243$) and short vegetation ($p = 0.5971$), the difference between observed and predicted flooded pixels is not even statistically significant, despite the large sample size. When the model indicates that a location is flooding, the physical profile of that location (in terms of soil, elevation, and vegetation) corresponds very well with places that flood in reality. The dry class, on the other hand, shows large deviations, with RCB values deviating far from 0. For the elevation, for example, the pixels designated as dry by the model are on average significantly higher than the observed dry pixels (RBC = -0.2814). In addition, the predicted dry pixels also contain much more short vegetation (RBC = -0.4963) and much less forest (RBC = 3256) than actually dry pixels. It is no coincidence that these three variables (elevation, forest, short vegetation) also appeared in table 5.4, where they were the factors with the greatest absolute increase in RBC. The model only classifies pixels as dry if they are located at extremely high altitudes and are dominated by low vegetation. As a result, lower-lying dry areas or dry forests are incorrectly identified as flooded (or vice versa), which explains why the observed and predicted dry zones differ so strongly in content.

Table 5.6 compares the distribution of the observed and the modelled flood fraction. It also shows the results of the Wilcoxon signed-rank test, which is used to test whether the distribution of the modelled flood fractions differs from the observed distribution.

Table 5.6: Comparison of observed and modelled flood fraction distributions ($N = 38, 225$).

Metric	Observed	Modelled	
Mean	0.2423	0.2243	
Standard deviation	0.2053	0.0620	
Test	U-statistic	p-value	RBC
Wilcoxon signed-rank test	$U = 5.814 \cdot 10^8$	$< 10^{-16}$	0.2041

The model is very close to reality in terms of the average (22.4% vs. 24.2%). The total flooded surface area is therefore estimated moderately to reasonably well. But the big difference lies in the spread: the standard deviation of the model (0.0620) is more than 3 times smaller than in reality (0.2053). The model is conservative: it does not dare to predict extreme flood fractions and safely hovers around the average of 0.22. The significance of the Wilcoxon signed-rank test points in the same direction: despite the correct mean, the predicted distribution differs systematically from the observations. This points again on the regression to the mean behaviour observed earlier. A similar behaviour was for example observed in figure 5.1.

Case studies

Three case studies are chosen to have a visual interpretation of the capabilities of the model. The case studies all occurred in 2023 and 2024, the two years that make up the test set. The three cases were selected in order to have a range of different climate zones. In figure 5.6, a case near the city of Ivanava in Belarus is depicted. This flood is observed at 22/09/2024, after the passage storm Boris. This storm caused many places in (Central-) Europe to flood (Riboldi et al., 2025). Figure 5.7 depicts a flood near the city of Uranch in Pakistan at 4/07/2024. Pakistan experiences regularly severe floods, caused by a combination of rapid snowmelt and monsoon rains (Waseem and Rana, 2023). Figure 5.8 depicts a case near Wanlaweyn in Somalia at 26/11/2023. Also Somalia experiences regularly floods, often caused by the Deyr (short rainy season in the Horn of Africa). The floods in 2023 affected 2.48 billion people of which 1.2 million people had to be displaced and 118 people died (Mabumbo et al., 2024).

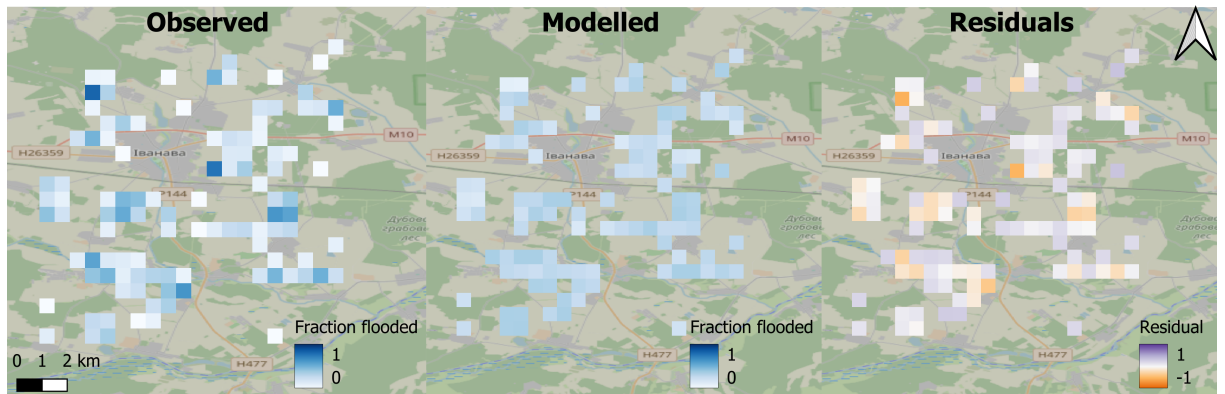


Figure 5.6: Case study around the city of Ivanava (Belarus) at 22/09/2024. The left panel represents the observations, the middle panel depicts the model results and the residuals (modelled-observed) are shown in the right panel.

In the first case, the LSTM model reproduces the general spatial patterns quite well: the main clusters are predicted at the correct locations. However, the highest observed fractions are smoothed out. Where observations show dark pixels locally, the model often predicts lighter values. This confirms the pattern earlier observed: the regression models towards the means instead of predicting extremes. The residual map shows predominantly light purple colors distributed across a large part of the area, along with some localised orange pixels where the model underpredicts flooding. No clear systematic spatial error pattern is visible.

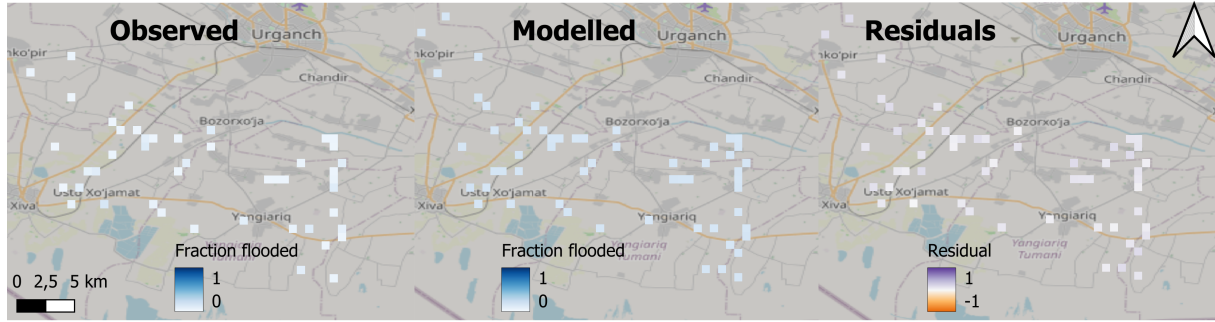


Figure 5.7: Case study around the city of Urganj (Pakistan) at 4/07/2024. The left panel represents the observations, the middle panel depicts the model results and the residuals (modelled-observed) are shown in the right panel.

In the second case, the flooding is much less dense. Nevertheless, the model's prediction is rather accurate, with only a few flooded pixels missed. The flood fractions are very low, and the model correctly estimates them as such. The residuals appear almost entirely white or very pale purple, indicating the absence of any clear systematic over- or underestimation. This example demonstrates that the model performs reliably, even in situations with limited inundation. However, this case requires cautious interpretation because both the observations and predictions are close to zero and therefore a small error automatically results in a favorable residual pattern. In other words, while the agreement is good, this is also a relatively simple case due to the absence of extreme inundation.

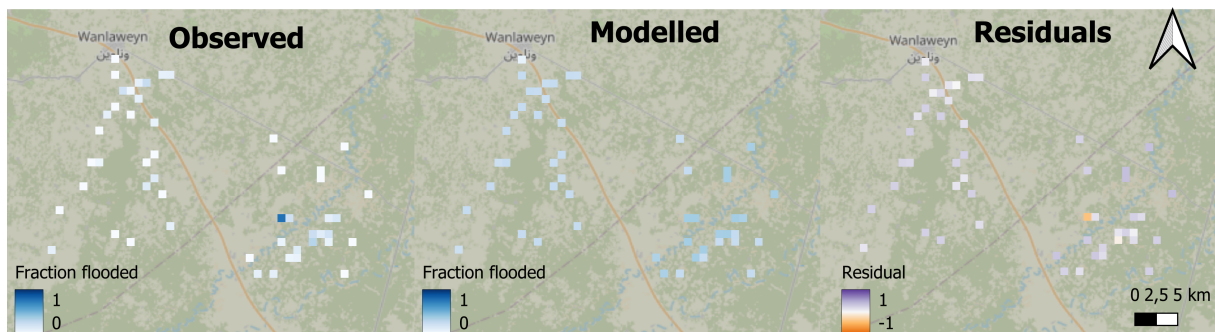


Figure 5.8: Case study around the city of Wanlaweyn (Somalia) at 26/11/2023. The left panel represents the observations, the middle panel depicts the model results and the residuals (modelled-observed) are shown in the right panel.

The third case exhibits a linear and more dispersed pattern. The observed inundation appears strongly linked to a network of low-lying areas or drainage structures. The model reproduces the general pattern remarkably well: the north-south oriented structure is

correctly captured, and the secondary cluster in the southeast is also predicted. The spatial distribution of the flooding is maintained. However, it is again noticeable that some local peaks disappear. The residuals consist almost exclusively of very light purple values, indicating a slight overestimation of the flood fraction. There are few locations showing significant underestimation. This case suggests that the model can reproduce not only compact inundation clusters but also more linear and spatially complex patterns. This is a substantial indication that the model is learning more than just an average flood probability.

Over the three cases, the key inundation areas are correctly identified. The model thus appears to effectively learn the spatial occurrence of pluvial flooding. In each case, the highest observed values appear to be attenuated in the predictions, as previously noted. This indicates a classic regression bias: high values are underestimated and low values slightly overestimated.

5.2 Climate attribution

Figure 5.9 compares for all dates the total flooded area predicted by the model under factual climatic conditions with the predicted area under counterfactual climatic conditions.

The spatial and temporal patterns of pluvial flooding remain largely preserved between both scenarios, but the actual climatic conditions systematically lead to larger flood magnitudes. That points at a clear pattern: more pluvial floods occur under the current climate than if there would not have been any climate change.

Figure 5.10 shows two histograms. The figure compares the distribution of daily flood magnitude for both climate scenarios. For each date, the total flooded area predicted by the model was calculated. Subsequently, the complete distribution was displayed for both climates. The histograms have been normalized so that they can be interpreted as probability densities.

The factual distribution is shifted towards larger flooded extents, indicating that larger daily flood impacts occur more frequently under factual climate conditions. An overview of the test results of the comparison of factual and counterfactual can be found in table 5.7

Table 5.7: Comparison of factual and counterfactual flood fraction distributions.

Metric	Factual	Counterfactual	
Mean	23957644.14	23849557.32	
Standard deviation	29613.64	26863.87	
Test	U-statistic	p-value	RBC
Wilcoxon signed-rank test	$U = 0.0$	$< 10^{-16}$	1

A relative increase of 0.5% is observed over the entire period and the entire globe. While extreme precipitation is generally expected to intensify under a warming climate (7% following the Clausius-Clapeyron equation), the response of pluvial flooding is more complex and additionally depends on factors such as antecedent soil moisture, land cover, drainage capacity and local topography. Consequently, changes in precipitation extremes are not expected to translate directly into proportional changes in flood extent. While

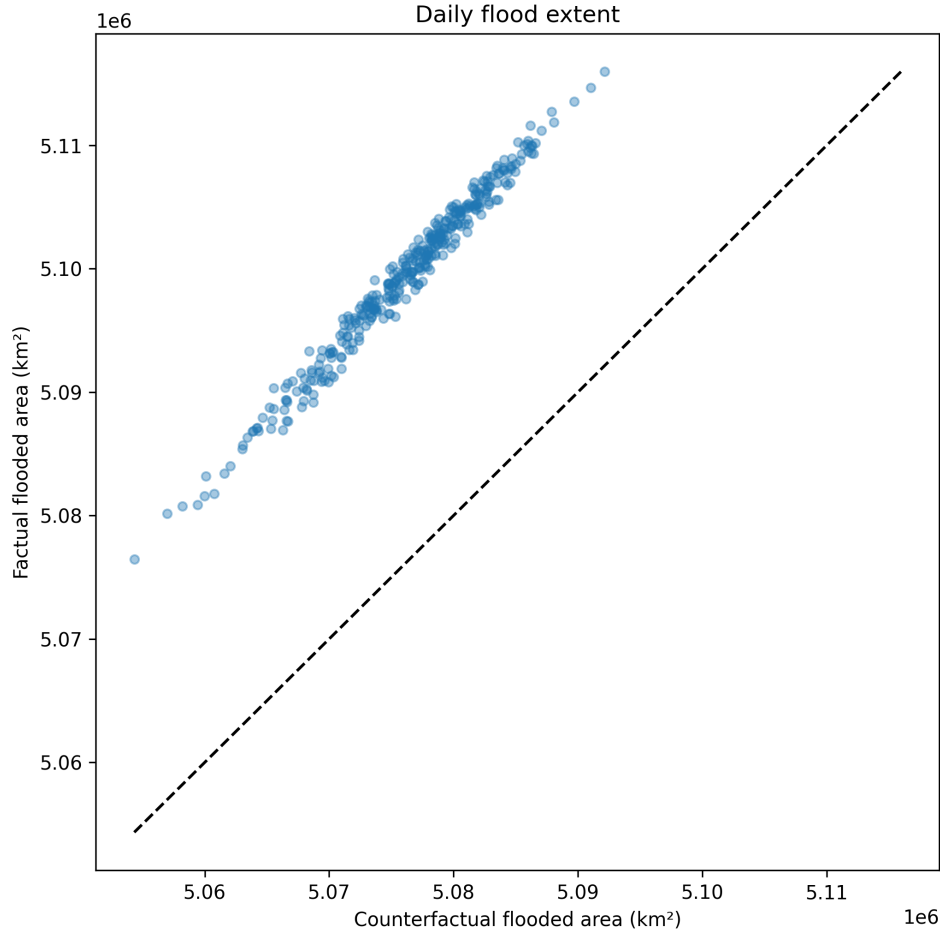


Figure 5.9: Daily flooded extent predicted under factual and counterfactual climate conditions. Each point represents one analysis date. The dashed line indicates equal flooded extent in both scenarios.

this increase is modest in magnitude, the signal is consistently detected across the attribution analyses and the test statistic is large. Since every day in the tested period experienced an increase, the U-statistic becomes 0 and the effect size is zero. Caution is however required, since the effect size is especially sensitive to how consistent the direction of the change is and less for the size of the change. In case the model would be thrust, another explanation could be the buffering effect of the soil. The relationship between precipitation extremes and pluvial flooding is inherently non-linear. Increased precipitation does not automatically translate into proportional increases in flood extent because runoff generation also depends on antecedent soil moisture, land cover, drainage capacity, and local topography.

For both the factual and counterfactual climate, all daily flood magnitudes were ranked from largest to smallest. The x-axis of figure 5.11 represents the ranking of events, while the y-axis represents the corresponding flood magnitude. The curve can be interpreted as an empirical exceedance curve: the most extreme events are located on the left, and the less exceptional events are located on the right.

Since the factual curve is located above the counterfactual curve over the entire range, the model points to an increase in pluvial flooding caused by climate change. The left

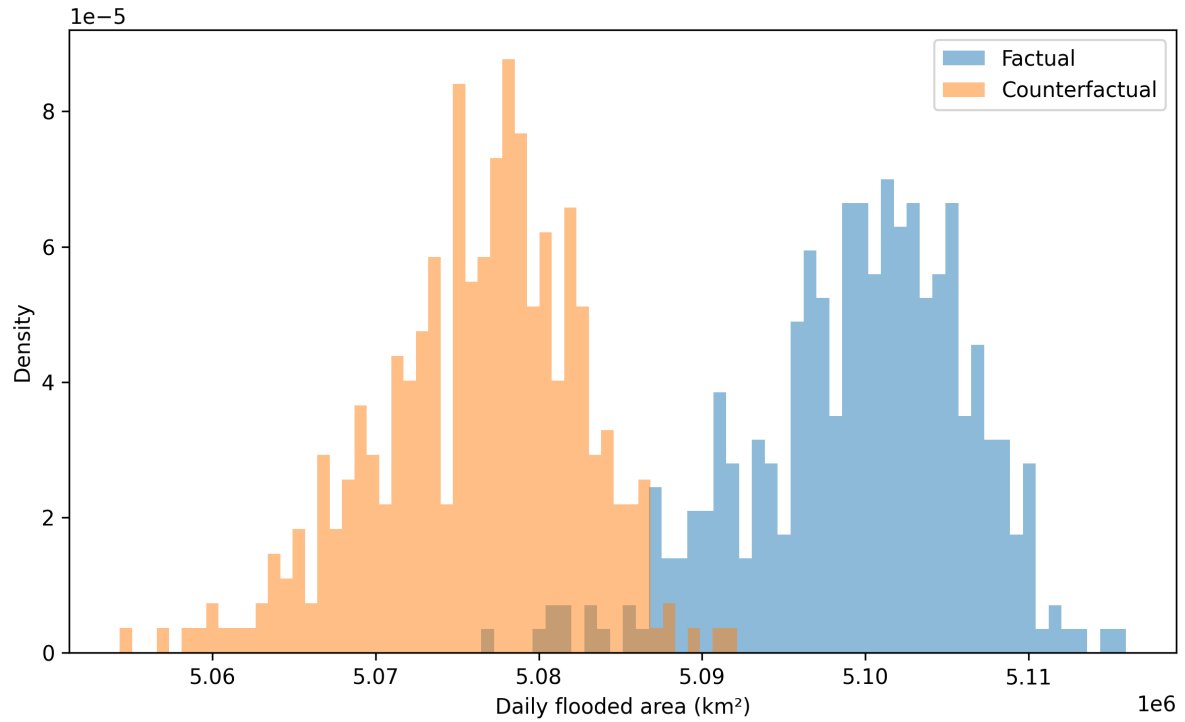


Figure 5.10: Distribution of daily flooded extent under factual and counterfactual climate conditions. Histograms are normalized to represent probability densities.

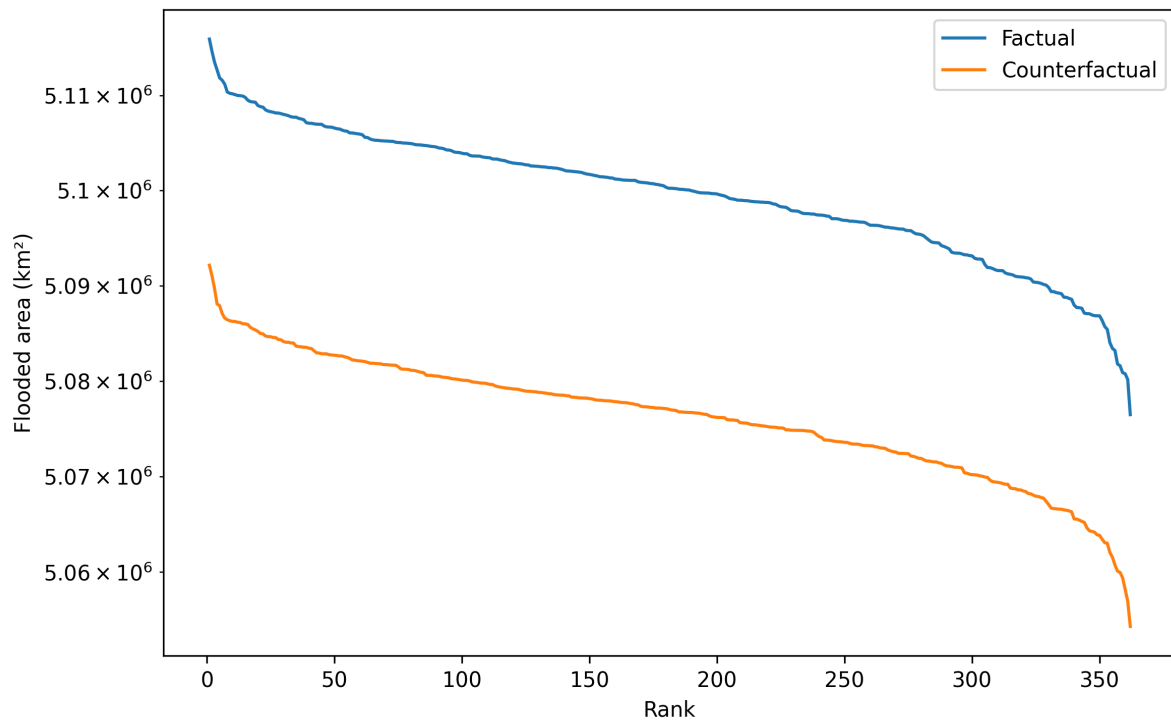


Figure 5.11: Exceedance curves of daily flooded extent under factual and counterfactual climate conditions. Events are ranked from largest to smallest flooded extent.

side of the figure, where the most exceptional events occur, is particularly relevant. Both curves follow exactly the same shape, indicating that the increase in pluvial floods occurs at the same intensity for all return periods. For many extreme events (like floods), the extremes increase more. This cannot be stated for pluvial floods. It has to be stated that we earlier concluded that the model tends to predict in the direction of the mean and underestimates extremes. This might be the primary cause of the observed pattern.

To get an idea about the spatial distribution of the changes, a map is made of the relative change in predicted pluvial flooding (figure 5.12). Red colours indicate an increase in pluvial flooding under actual climatic conditions, while blue colours indicate a decrease. The changes are calculated per tile of 1000 x 1000 km.

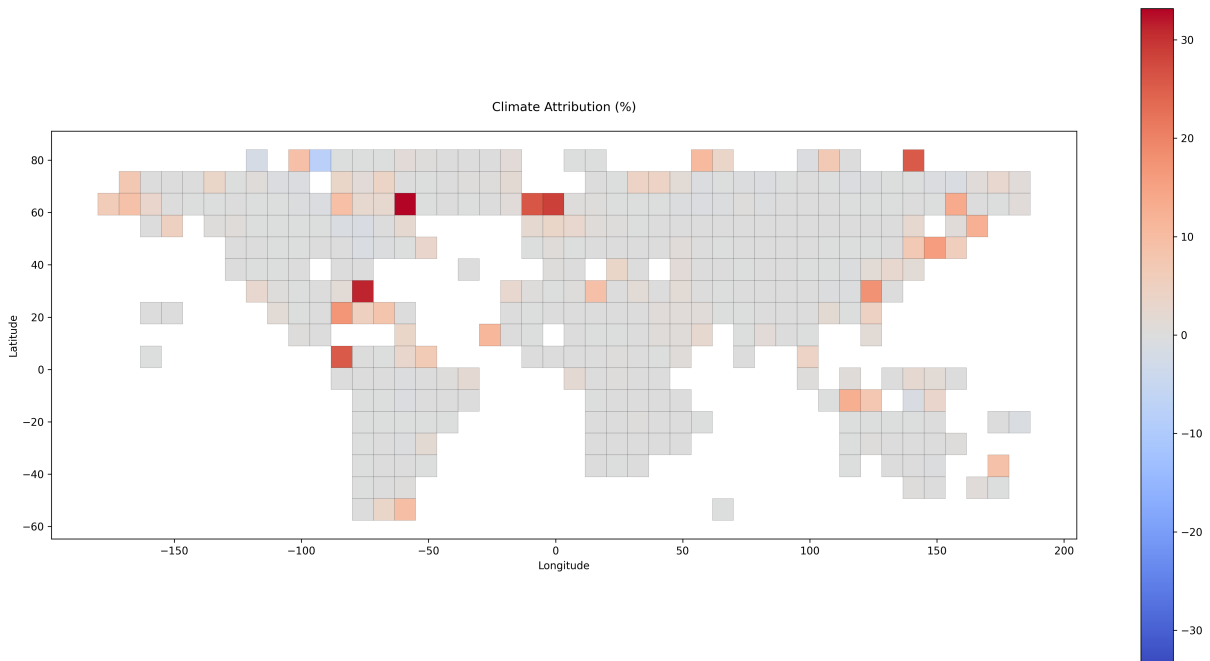


Figure 5.12: Spatial distribution of the relative change in modeled pluvial flood extent between factual and counterfactual climate conditions. Each tile represents the percentage change in total flooded fraction relative to the counterfactual simulation. Red shades indicate larger flooded extents under factual climate conditions, whereas blue shades indicate smaller flooded extents.

For many regions, the predicted pluvial flood magnitude changes hardly between factual and counterfactual. That may reflect that the global signal is primarily carried by a limited number of regions with a stronger response. There are a number of prominent red areas: northern South America, parts of North America, parts of East Asia, some regions in Oceania, and scattered zones in Africa. There are far less pronounced negative zones. There are some blue tiles near Alaska, northern Canada, and at a few locations in North Asia, but far fewer than the red ones. There appears to be no simple continental pattern. The signal appears spatially heterogeneous. This is typical of pluvial floods, which are highly dependent on local precipitation patterns and regional climate systems and are much less uniform than temperature attribution.

Chapter 6

Limitations

As for many data-driven methods, the quality of the model is highly impacted by the data quality. Data are limited by temporal and spatial resolution and data accuracy.

The CHELSA precipitation dataset has daily data on a 1 km resolution. Since pluvial floods are often related to thunderstorms, especially the temporal scale might be too coarse to capture the rainfall intensity properly. A volume of water falling on a certain pixel spread over a few hours could cause a flood, while the same volume spread over the entire day does not cause flooding on that pixel. The downscaling algorithm (CHELSA V2.1) corrects for orographic effects, but convection was not included. Therefore, the downscaled precipitation does not perform better for convective precipitation than the coarse W5E5 dataset. The fact that a model trained on a simple nearest neighbour downscaling of the ERA5 precipitation performs only slightly worse than the model using the CHELSA V2.1 algorithm also indicates that CHELSA might not be the optimal downscaling approach for this case. Given the importance of convective precipitation for pluvial floods, using a downscaling algorithm that explicitly includes convection is likely to increase the accuracy of the model.

The landcover data are taken from the situation in 2020. In reality, the landcover changes over time. Although the simplification of stable landcover is generally valid, on a local scale, changes can be significant (e.g. urban expansion, deforestation). Given the importance of landcover to predict pluvial floods, including a changing landcover map (e.g. annual landcover) could potentially increase the model performance. The original 11 landcover classes from table 4.1 are reduced to four new classes in order to reduce the degrees of freedom and therefore ease the learning process. This however makes the classes broader and therefore less precise. For example the class 'short vegetation' forms a relatively heterogeneous group. Large (hydrologic) differences within a class complicates the training process. The landcover data are available on a 10 m resolution. Since the data needed to be upscaled to the 1 km resolution of the precipitation data, for each of the new landcover classes, a percentage of the larger pixel that is taken by the original pixels is saved. This reduces the loss of information, but not entirely prevents that some of the initial information is lost.

The splitting of the observed floods was done by using the JRC inundation map for a 100 year return period. All pixels with an inundation depth of at least 10 cm on the T100-map are indicated as 'fluvial' pixels, meaning that any flood on these pixels will automatically be classified as fluvial. Floods on all other pixels are classified as pluvial floods. The major drawback of this method is that pluvial floods occurring in river floodplains are potentially not taken into account. Especially for rivers crossing urban

areas and cities, the risk of missing pluvial floods is severe. A 100-year return period was chosen as a middle ground between return periods of 10 and 500 years. The trade-off consists of having more pluvial floods indicated as fluvial when a longer return period would be used and taking too much fluvial floods into account if a shorter return period would be used.

Chapter 7

Future research

7.1 Improved analysis with the current setup

As described in the methodology section (Section 3), the model was first trained on high-resolution precipitation data, downsampled by using the CHELSA-algorithm. That reveals the potentials the model could have. However, since reanalysis data and precipitation projections are not downsampled by CHELSA yet, the model is retrained by using the lower resolution ERA5 data that are downsampled by a nearest neighbour algorithm, which lowers the accuracy. Once the data for the counterfactual and the future scenarios are available, the analysis will increase in accuracy. As stated in section 6, using another downscaling algorithm, which allows for convection processes, could potentially even perform better.

Currently, only a climate attribution has been performed by using the model. Another potential application of this model is to make projections for the future under different climate scenarios. Additionally, the model could also be used to make the step to impact analysis, when it is combined with population data, mortality rates, etc. Also comparing the amount of pluvial floods during the lifetime of different age-cohortes could be done with the model as it is.

7.2 Improvements to the model

As denoted earlier, the soil moisture content is an important factor to determine whether it will flood or not. The model with all fluxes included has a significantly higher accuracy than the model where only the precipitation is provided. To determine the amount of water stored in the soil, soil moisture evaporation is a potentially important process, that is amongst others determined by the temperature and the wind speed. Also the transpiration is determined by other meteorological factors except for the precipitation (temperature, solar irradiation, etc.). A potential change to the model could be to add other meteorological variables from the CHELSA database (like temperature, humidity, etc.). The counterpart of adding these meteorological datasets to the model, is increasing the model complexity and therefore its required training time. Additionally, the amount of data stored increases significantly.

Nearing et al. (2021) observed that for cases with good datasets, the model develops internal states that strongly correlate with physical properties like the snow pack or the soil moisture. Ideally, the model could learn the concepts of soil moisture and evapotranspiration (and snowpack) itself, but only if the data are informative enough. Therefore,

adding a feature that is more directly correlated with these fluxes helps the model to learn these processes and converge easier.

The Standardised Precipitation Index (SPI) is a commonly used indicator for meteorological droughts (McKee et al., 1993). It measures the deviation in precipitation (P) of a certain period from the average value ($\hat{P}$), divided by the standard deviation (σ_P), after the timeseries have been transformed to a gamma function (normal distribution) (7.1). A wet period has a positive SPI and a dry period has a negative SPI. SPI values could be calculated over different timescales (1 month, 3 months, 6 months, a year, etc.), making it relevant for different processes (soil moisture fluctuates over shorter periods than groundwater). As the SPI can be used as indicator for soil moisture, adding this index explicitly to the model potentially increases the model performance.

$$SPI = \frac{P - \hat{P}}{\sigma_P} \quad (7.1)$$

While SPI captures the effects of the soil moisture, it does not account for the evapotranspiration, which largely affects the hydrological cycle. An index that potentially captures both the effect of soil moisture and evapotranspiration is the Standardised Precipitation-Evapotranspiration Index (SPEI). This index calculates exactly the same as the SPI, but the potential evapotranspiration (PET) is subtracted from the precipitation (Vicente-Serrano et al., 2010). The PET can be calculated based on a parametric equation, as for example the Penman-Monteith equation that is used by the FAO. However, this requires information on the air temperature, the relative humidity, wind speed, solar irradiation and the air pressure and therefore the storage for all these variables (downscaled by CHELSA).

7.3 Landcover change attribution

In the current setup, the model is used to attribute the impact of climate change on the occurrence of pluvial floods. Apart from climate change, landcover changes are likely to impact pluvial flooding. Especially deforestation and urbanisation are suspected to have a significant impact. Attributing these impacts, besides of those of climate change, will provide a more complete image of the driving forces behind the mechanisms of pluvial floods. Since landcover interventions are often suggested as potential adaptation measures, a proper global assessment could give an idea about the effectiveness of these measures. However, as indicated in Section 6, to have a model that can capture the impacts of landcover types properly, the model should be trained on a time-dynamic landcover map. Future simulations could be done by adding one or multiple landcover change scenarios as well.

7.4 Regional application

The current model is trained on a global scale with a daily resolution and a spatial resolution of 1 km. The model architecture however can be applied to all spatial and temporal resolutions and scales. Therefore, the model could be applied to a specific region only. The advantage of training on one area only is that the model weights can be adapted specifically to model the situation in that particular region, instead of performing on all places on earth. Additionally, for a regional scale, often more detailed information

is available. Therefore the model could be trained on a subdaily and subkilometer scale, based on data that are likely to have a higher accuracy (e.g. a local soil map). This could significantly improve the model accuracy. It is likely that this kind of model could be used for forecasting as well.

Chapter 8

Conclusions

8.1 Model

In this thesis, we prove that a relatively simple deep learning architecture is able to learn global useful patterns of pluvial floods. Despite only a very limited amount of static and dynamic variables were used, the model achieves systematically better results than a random classification. That proves that pluvial floods on a global scale are sufficiently related to these variables to be predictable.

The comparison between ANN and LSTM demonstrates the relevance of temporal memory. The flux-based LSTM outperforms the flux-based ANN across virtually all classification metrics. This may reflect that the sequence of wet and dry days contains important information that cannot be fully derived from static variables. Including soil moisture and evapotranspiration improves almost all classification metrics. This supports the theory that pluvial floods are not only driven by precipitation on the day of the event itself. The model performs significantly better when information regarding antecedent hydrological conditions is available. This supports classical hydrological theory concerning saturation-excess runoff, antecedent soil moisture, and soil storage processes.

The sensitivity analyses show that the model understands the protective effect of forest, associates higher risks with low vegetation, learns the effects of soil texture, and picks up on topographical flood control. Even when the absolute probabilities are not perfect, the model generally follows the correct physical trends. This implies that the model learns not only statistical correlations but also processes that are hydrologically plausible.

The classification performs relatively well. Particularly in the case studies, the flooded pixels were rendered quite accurately. The regression head performs much worse. Evidence for this includes a systematic underestimation of major floods, an overestimation of minor floods, very low R^2 values, and a strong reduction in the standard deviation of the predicted flood fractions. Although the average flood fraction is reproduced well (0.242 observed, 0.224 modeled), the variance is less accurate (0.205 observed, 0.062 modeled).

8.2 Climate attribution

The climate attribution analysis indicates that climate change has increased global pluvial flood occurrence and extent. Across the analysed period, factual climate conditions resulted in larger flooded areas than their counterfactual counterparts. Although the estimated increase is relatively modest, approximately 0.5% at the global scale, the signal

is consistently present across the different attribution analyses.

Strong regional differences were observed, with several hotspots showing a more pronounced response to climate change than others. This spatial heterogeneity reflects the strong dependence of pluvial flooding on local precipitation regimes and environmental conditions. While the absolute magnitude of the change should be interpreted with caution due to model limitations and the tendency to underestimate extremes, the results provide evidence that climate change has already contributed to increased pluvial flooding worldwide.

As climate change continues to alter precipitation extremes worldwide, improving our ability to understand, predict and attribute pluvial flooding becomes increasingly important. This thesis demonstrates that deep learning can contribute to this challenge by extracting physically meaningful flood patterns from large-scale environmental datasets. While the current model still has clear limitations, it shows that global pluvial flood modelling and climate attribution using deep learning are feasible and hold considerable potential for further development. Ultimately, improving such modelling capabilities will be essential for quantifying the impacts of climate change and supporting adaptation efforts in flood-prone regions around the world.

Bibliography

- Abbot, J. and Marohasy, J. (2014). Input selection and optimisation for monthly rainfall forecasting in queensland, australia, using artificial neural networks. *Atmospheric Research*, 138:166–178.
- Aghenda, M., Labbaci, A., Bouchaou, L., Hssaisoune, M., and Ait Brahim, Y. (2025). Flood prediction using machine learning and deep learning models: a systematic review. *Mediterranean Geoscience Reviews*, pages 1–19.
- Alfieri, L., Cohen, S., Galantowicz, J., Schumann, G. J., Trigg, M. A., Zsoter, E., Prudhomme, C., Kruczkiewicz, A., de Perez, E. C., Flamig, Z., et al. (2018). A global network for operational flood risk reduction. *Environmental science & policy*, 84:149–158.
- Ali, H., Fowler, H. J., Lenderink, G., Lewis, E., and Pritchard, D. (2021). Consistent large-scale response of hourly extreme precipitation to temperature variation over land. *Geophysical Research Letters*, 48(4):e2020GL090317.
- An, H. and Ouyang, C. (2025). A hybrid framework for real-time flash flood forecasting in small ungauged catchments: integrating hydrodynamic simulations with lstm networks. *Journal of Hydrology*, 661:133688.
- Arnell, N. W. and Gosling, S. N. (2016). The impacts of climate change on river flood risk at the global scale. *Climatic Change*, 134(3):387–401.
- Barichivich, J., Gloor, E., Peylin, P., Brienens, R. J. W., Schöngart, J., Espinoza, J. C., and Pattanayak, K. C. (2018). Recent intensification of Amazon flooding extremes driven by strengthened Walker circulation. *Science Advances*, 4(9):eaat8785.
- Bauer-Marschallinger, B., Cao, S., Tupas, M. E., Roth, F., Navacchi, C., Melzer, T., Freeman, V., and Wagner, W. (2022). Satellite-based flood mapping through bayesian inference from a sentinel-1 sar datacube. *Remote Sensing*, 14(15):3673.
- Baugh, C., Colonese, J., D’Angelo, C., Dottori, F., Neal, J., Prudhomme, C., and Salamon, P. (2024). Global river flood hazard maps. European Commission, Joint Research Centre (JRC) [Dataset]. Consulted on 12 February 2025.
- Berghuijs, W. R., Woods, R. A., Hutton, C. J., and Sivapalan, M. (2016). Dominant flood generating mechanisms across the United States. *Geophysical Research Letters*, 43(9):4382–4390.
- Bharali, B. (2021). Rate of infiltration for different soil textures using rainfall simulator and green-ampt model. *ISH Journal of Hydraulic Engineering*, 27(sup1):70–76.

- Blöschl, G., Hall, J., Viglione, A., Perdigão, R. A., Parajka, J., Merz, B., Lun, D., Arheimer, B., Aronica, G. T., Bilibashi, A., et al. (2019). Changing climate both increases and decreases european river floods. *Nature*, 573(7772):108–111.
- Bulti, D. T. and Abebe, B. G. (2020). A review of flood modeling methods for urban pluvial flood application. *Modeling earth systems and environment*, 6(3):1293–1302.
- Cea, L., Sañudo, E., Montalvo, C., Farfán, J., Puertas, J., and Tamagnone, P. (2025). Recent advances and future challenges in urban pluvial flood modelling. *Urban Water Journal*, 22(2):149–173.
- Chen, W., Hong, H., Li, S., Shahabi, H., Wang, Y., Wang, X., and Ahmad, B. B. (2019). Flood susceptibility modelling using novel hybrid approach of reduced-error pruning trees with bagging and random subspace ensembles. *Journal of Hydrology*, 575:864–873.
- Chini, M., Hostache, R., Giustarini, L., and Matgen, P. (2017). A hierarchical split-based approach for parametric thresholding of sar images: Flood inundation as a test case. *IEEE Transactions on Geoscience and Remote Sensing*, 55(12):6975–6988.
- Cho, C., Li, R., Wang, S.-Y., Yoon, J.-H., and Gillies, R. R. (2016). Anthropogenic footprint of climate change in the June 2013 northern India flood. *Climate Dynamics*, 46(3–4):797–805.
- Copernicus Emergency Management Service (2025). Glofas global flood monitoring (gfm). <https://global-flood.emergency.copernicus.eu/technical-information/glofas-gfm/>. Consulted on 2 December 2025.
- Costabile, P. and Macchione, F. (2015). Enhancing river model set-up for 2-d dynamic flood modelling. *Environmental Modelling & Software*, 67:89–107.
- CRED and UNISDR (2015). The human cost of natural disasters: A global perspective. Technical report, UNISDR, Geneva, Switzerland.
- Cui, Z., Zhou, Y., Guo, S., Wang, J., Ba, H., and He, S. (2021). A novel hybrid xaj-lstm model for multi-step-ahead flood forecasting. *Hydrology Research*, 52(6):1436–1454.
- Dankers, R., Arnell, N. W., Clark, D. B., Falloon, P. D., Fekete, B. M., Gosling, S. N., Heinke, J., Kim, H., Masaki, Y., Satoh, Y., Stacke, T., Tang, Q., and Wada, Y. (2014). First look at changes in flood hazard in the Inter-Sectoral Impact Model Intercomparison Project ensemble. *Proceedings of the National Academy of Sciences*, 111(9):3257–3261.
- Deng, H., Chen, W., and Huang, G. (2022). Deep insight into daily runoff forecasting based on a cnn-lstm model. *Natural Hazards*, 113(3):1675–1696.
- Ding, Y., Zhu, Y., Feng, J., Zhang, P., and Cheng, Z. (2020). Interpretable spatio-temporal attention lstm model for flood forecasting. *Neurocomputing*, 403:348–359.
- Dookie, N., Chadee, X. T., and Clarke, R. M. (2019). Trends in extreme temperature and precipitation indices for the Caribbean small islands: Trinidad and Tobago. *Theoretical and Applied Climatology*, 136(1):31–44.

- Dudley, R. W., Hodgkins, G. A., McHale, M. R., Kolian, M. J., and Renard, B. (2017). Trends in snowmelt-related streamflow timing in the conterminous United States. *Journal of Hydrology*, 547:208–221.
- Dunn, R. J. H., Alexander, L. V., Donat, M. G., Zhang, X., Bador, M., Herold, N., Lippmann, T., Allan, R., Aguilar, E., Brunet, M., Caron, M., de Lima, R., Frich, P., García-Asensio, A., Jacob, O., Kiktev, D., Kruger, A., Kubota, H., Lal, M., Lashkari, A., Lim, G., McCabe, G., Mendes, M., Milewska, E., Nandintsetseg, B., Raymond, C., Rusticucci, M., Srivastava, A., Tamararatne, S., Timmerman, S., Toreti, A., Twist, G., van der Schrier, G., Villaroel, C., Vincent, L., Vischel, T., Witten, D., and Zong, C. (2020). Development of an updated global land in situ-based data set of temperature and precipitation extremes: HadEX3. *Journal of Geophysical Research: Atmospheres*, 125(16):e2019JD032263.
- European Space Agency (2025). Copernicus dem dataspace. <https://dataspace.copernicus.eu/explore-data/data-collections/copernicus-contributing-missions/collections-description/COP-DEM>. Consulted on 2 December 2025.
- European Union’s Copernicus Land Monitoring Service (CLMS) (2026). Product validation report: Global land cover map, 10 m (lcm-10), version 1. Technical Report Issue 1.0, European Commission, Joint Research Centre (JRC). Copernicus Global Land Cover and Tropical Forest Mapping and Monitoring service (LCFM).
- FAO, IIASA, ISRIC, ISS-CAS, and JRC (2008). Harmonized world soil database (version 1.0). Technical report, FAO and IIASA, Rome, Italy and Laxenburg, Austria. Version 1.0.
- FAO, IIASA, ISRIC, ISS-CAS, and JRC (2012). Harmonized world soil database v1.2. <https://www.fao.org/soils-portal/data-hub/soil-maps-and-databases/harmonized-world-soil-database-v12/en/>. Accessed on: 22 februari 2026.
- Fischer, A. M. et al. (2015). Projected changes in precipitation intensity and frequency in switzerland: a multi-model perspective. *International Journal of Climatology*, 35(11):3204–3219.
- Fowler, H. J. et al. (2021). Anthropogenic intensification of short-duration rainfall extremes. *Nature Reviews Earth & Environment*, 2(2):107–122.
- Fronk, C. and Petzold, L. (2025). The vanishing gradient problem for stiff neural differential equations. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 35(11).
- Furquim, G., Pessin, G., Faical, B. S., Mendiondo, E. M., and Ueyama, J. (2016). Improving the accuracy of a flood forecasting model by means of machine learning and chaos theory: A case study involving a real wireless sensor network deployment in brazil. *Neural computing and applications*, 27(5):1129–1141.
- Global Land Evaporation Amsterdam Model (2025). Global land evaporation amsterdam model. <https://www.gleam.eu/>. Consulted on 2 December 2025.
- Guo, J. (2024). Explicit solution of horton’s equation for infiltration capacity. *Journal of Irrigation and Drainage Engineering*, 150(6):06024002.

- Hallegatte, S., Green, C., Nicholls, R. J., and Corfee-Morlot, J. (2013). Future flood losses in major coastal cities. *Nature climate change*, 3(9):802–806.
- Helsen, S. et al. (2020). Consistent scale-dependency of future increases in hourly extreme precipitation in two convection-permitting climate models. *Climate Dynamics*, 54(3):1267–1280.
- Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., et al. (2020). The era5 global reanalysis. *Quarterly journal of the royal meteorological society*, 146(730):1999–2049.
- Hirabayashi, Y., Mahendran, R., Koirala, S., Konoshima, L., Yamazaki, D., Watanabe, S., Kim, H., and Kanae, S. (2013). Global flood risk under climate change. *Nature climate change*, 3(9):816–821.
- Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8):1735–1780.
- Horton, R. E. (1940). An approach toward a physical interpretation of infiltration capacity. In *Soil science Society of America proceedings*, volume 5, page 24. Madison.
- Jonkman, S. N. (2005). Global perspectives on loss of human life caused by floods. *Natural hazards*, 34(2):151–175.
- Kabir, S., Patidar, S., Xia, X., Liang, Q., Neal, J., and Pender, G. (2020). A deep convolutional neural network model for rapid prediction of fluvial flood inundation. *Journal of Hydrology*, 590:125481.
- Karger, D. N., Lange, S., Hari, C., Reyer, C. P., Conrad, O., Zimmermann, N. E., and Frieler, K. (2023). Chelsa-w5e5: daily 1 km meteorological forcing data for climate impact studies. *Earth System Science Data*, 15(6):2445–2464.
- Karger, D. N. and Zimmermann, N. E. (2021). Climatologies at high resolution for the earth land surface areas chelsa v1. 2: technical specification. *Swiss Federal Research Institute WSL*. Retrieved January, 7:2025.
- Karim, F., Armin, M. A., Ahmedt-Aristizabal, D., Tychsen-Smith, L., and Petersson, L. (2023). A review of hydrodynamic and machine learning approaches for flood inundation modeling. *Water*, 15(3):566.
- Kharin, V. et al. (2018). Risks from climate extremes change differently from 1.5°C to 2.0°C depending on rarity. *Earth’s Future*, 6(5):704–715.
- Kim, B., Sanders, B. F., Famiglietti, J. S., and Guinot, V. (2015). Urban flood modeling with porous shallow-water equations: A case study of model errors in the presence of anisotropic porosity. *Journal of Hydrology*, 523:680–692.
- Kratzert, F., Klotz, D., Brenner, C., Schulz, K., and Herrnegger, M. (2018a). Rainfall–runoff modelling using long short-term memory (lstm) networks. *Hydrology and Earth System Sciences*, 22(11):6005–6022.
- Kratzert, F., Klotz, D., Herrnegger, M., and Hochreiter, S. (2018b). A glimpse into the unobserved: Runoff simulation for ungauged catchments with lstms. In *Workshop on Modeling and Decision-Making in the Spatiotemporal Domain, 32nd Conference on Neural Information Processing Systems (NeurIPS 2018)*, Montreal, Canada.

- Krieger, G., Moreira, A., Fiedler, H., Hajnsek, I., Werner, M., Younis, M., and Zink, M. (2007). Tandem-x: A satellite formation for high-resolution sar interferometry. *IEEE transactions on geoscience and remote sensing*, 45(11):3317–3341.
- Kroll, C. N. and Vogel, R. M. (2002). Probability distribution of low streamflow series in the united states. *Journal of hydrologic Engineering*, 7(2):137–146.
- Li, C. et al. (2021). Changes in annual extremes of daily temperature and precipitation in cmip6 models. *Journal of Climate*, 34(9):3441–3460.
- Li, X. and Willems, P. (2020). A hybrid model for fast and probabilistic urban pluvial flood prediction. *Water Resources Research*, 56(6):e2019WR025128.
- Liao, Y., Wang, Z., Chen, X., and Lai, C. (2023). Fast simulation and prediction of urban pluvial floods using a deep convolutional neural network model. *Journal of Hydrology*, 624:129945.
- Lin, M., Lucas, H. C., and Shmueli, G. (2011). Too big to fail: larger samples and false discoveries. *Robert H. Smith School Research Paper No. RHS*, pages 06–068.
- Mabumbo, D., Nohayi, N., Sanchez Ramirez, J. C., Dessalegn, M., and Ruckstuhl, S. (2024). Navigating climate resilience in fragile settings: A retrospective analysis of the 2023 flood impacts, early warning, and response in ethiopia’s somali region. Technical report, International Water Management Institute (IWMI), Colombo, Sri Lanka.
- Martinis, S., Kuenzer, C., Wendleder, A., Huth, J., Twele, A., Roth, A., and Dech, S. (2015). Comparing four operational sar-based water and flood detection approaches. *International Journal of Remote Sensing*, 36(13):3519–3543.
- McKee, T. B., Doesken, N. J., Kleist, J., et al. (1993). The relationship of drought frequency and duration to time scales. In *Proceedings of the 8th Conference on Applied Climatology*, volume 17, pages 179–183. Boston.
- Mengel, M., Treu, S., Lange, S., and Frieler, K. (2021). Attrici v1. 1–counterfactual climate for impact attribution. *Geoscientific Model Development*, 14(8):5269–5284.
- Mosavi, A., Ozturk, P., and Chau, K.-w. (2018). Flood prediction using machine learning models: Literature review. *Water*, 10(11):1536.
- Mosavi, A., Rabczuk, T., and Varkonyi-Koczy, A. R. (2017). Reviewing the novel machine learning tools for materials design. In *International Conference on Global Research and Education*, pages 50–58. Springer.
- Nachtergaele, F., Van Velthuisen, H., Verelst, L., Batjes, N., Dijkshoorn, K., van Engelen, V., Fischer, G., Jones, A., Montanarella, L., Petri, M., et al. (2010). Harmonized world soil database. In *Proceedings of the 19th world congress of soil science, soil solutions for a changing world, Brisbane, Australia*, volume 2010, pages 34–37. International Union of Soil Sciences.
- Nearing, G., Cohen, D., Dube, V., Gauch, M., Gilon, O., Harrigan, S., Hassidim, A., Klotz, D., Kratzert, F., Metzger, A., et al. (2024). Global prediction of extreme floods in ungauged watersheds. *Nature*, 627(8004):559–563.

- Nearing, G. S., Kratzert, F., Sampson, A. K., Pelissier, C. S., Klotz, D., Frame, J. M., Prieto, C., and Gupta, H. V. (2021). What role does hydrological science play in the age of machine learning? *Water Resources Research*, 57(3):e2020WR028091.
- Neri, A., Villarini, G., Slater, L. J., and Napolitano, F. (2019). On the statistical attribution of the frequency of flood events across the U.S. Midwest. *Advances in Water Resources*, 127:225–236.
- Nevo, S., Morin, E., Gerzi Rosenthal, A., Metzger, A., Barshai, C., Weitzner, D., Voloshin, D., Kratzert, F., Elidan, G., Dror, G., et al. (2022). Flood forecasting with machine learning models in an operational framework. *Hydrology and Earth System Sciences*, 26(15):4013–4032.
- Nie, J., Sobel, A. H., Shaevitz, D. A., and Wang, S. (2018). Dynamic amplification of extreme precipitation sensitivity. *Proceedings of the National Academy of Sciences*, 115(38):9467–9472.
- Nix, D. and Weigend, A. (1994). Estimating the mean and variance of the target probability distribution. In *Proceedings of 1994 IEEE International Conference on Neural Networks (ICNN'94)*, volume 1, pages 55–60 vol.1.
- NOAA’s National Weather Service (2024). Glossary. <https://forecast.weather.gov/>. Consulted on 30 October 2025.
- Office of Public Works (2026). FloodInfo.ie - National Flood Information Portal. <https://www.floodinfo.ie/>. Accessed on 4 May 2026.
- Paik, S. et al. (2020). Determining the anthropogenic greenhouse gas contribution to the observed intensification of extreme precipitation. *Geophysical Research Letters*, 47(12):e2019GL086875.
- Pall, P., Tallaksen, L. M., and Stordal, F. (2019). A climatology of rain-on-snow events for Norway. *Journal of Climate*, 32(20):6995–7016.
- Pendergrass, A. G., Lehner, F., Sanderson, B. M., and Xu, Y. (2015). Does extreme precipitation intensity depend on the emissions scenario? *Geophysical Research Letters*, 42(20):8767–8774.
- Perera, M. D. D., Angeloudis, A., Siripatana, A., and Beevers, L. (2026). Hybrid methods in flood inundation modeling: a systematic review. *Natural Hazards*, 122(8):357.
- Philip, S., Kew, S. F., van Oldenborgh, G. J., Otto, F. E. L., O’Keefe, S., Hauser, M., King, A. D., Zehe, E., Jadav, G., and Vautard, R. (2018). Validation of a Rapid Attribution of the May/June 2016 Flood-Inducing Precipitation in France to climate change. *Journal of Hydrometeorology*, 19(11):1881–1898.
- Rehmer, A. and Kroll, A. (2020). On the vanishing and exploding gradient problem in gated recurrent units. *IFAC-PapersOnLine*, 53(2):1243–1248.
- Rentschler, J., Salhab, M., and Jafino, B. A. (2022). Flood exposure and poverty in 188 countries. *Nature communications*, 13(1):3527.

- Riboldi, J., Noyelle, R., Agayar, E., Binder, H., Federer, M., Hartmuth, K., Sprenger, M., Thurnherr, I., and Vishnupriya, S. (2025). Storm boris (2024) in the current and future climate: a dynamics-centered contextualization, and some lessons learnt. *EGUsphere*, 2025:1–33.
- Rosenzweig, C., Arnell, N. W., Ebi, K. L., Lotze-Campen, H., Raes, F., Rapley, C., Smith, M. S., Cramer, W., Frieler, K., Reyer, C. P., et al. (2017). Assessing inter-sectoral climate change risks: the role of isimp.
- Samadi, A., Jafarzadegan, K., and Moradkhani, H. (2025). Dem-based pluvial flood inundation modeling at a metropolitan scale. *Environmental Modelling & Software*, 183:106226.
- Samset, B. H., Sand, M., Smith, C. J., Bauer, S. E., Forster, P. M., Fuglestad, J. S., Osprey, S., and Schleussner, C.-F. (2018). Climate impacts from a removal of anthropogenic aerosol emissions. *Geophysical Research Letters*, 45(2):1020–1029.
- Schumann, G. J.-P. and Moller, D. K. (2015). Microwave remote sensing of flood inundation. *Physics and Chemistry of the Earth, Parts a/b/c*, 83:84–95.
- Seneviratne, S. I. and Hauser, M. (2020). Regional climate sensitivity of climate extremes in cmip6 versus cmip5 multimodel ensembles. *Earth’s Future*, 8(9):e2019EF001474.
- Seneviratne, S. I., Nicholls, N., Easterling, D., Goodess, C. M., Kanae, S., Kossin, J., Luo, Y., Marengo, J., McInnes, K., Rahimi, M., Reichstein, M., Sorteberg, A., Vera, C., and Zhang, X. (2012). Changes in climate extremes and their impacts on the natural physical environment. In Field, C. B., Barros, V., Stocker, T. F., and Dahe, Q., editors, *Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation*, pages 109–230. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA. A Special Report of Working Groups I and II of the Intergovernmental Panel on Climate Change (IPCC).
- Seneviratne, S. I., Zhang, X., Adnan, M., Badi, W., Dereczynski, C., Di Luca, A., Ghosh, S., Iskandar, I., Kossin, J., Lewis, S., Otto, F., Pinto, I., Satoh, M., Vicente-Serrano, S. M., Wehner, M., and Zhou, B. (2021). Weather and climate extreme events in a changing climate. In Masson-Delmotte, V., Zhai, P., Pirani, A., Connors, S. L., Péan, C., Berger, S., Caud, N., Chen, Y., Goldfarb, L., Gomis, M. I., Huang, M., Leitzell, K., Lonnoy, E., Matthews, J. B. R. t., Maycock, T. K., Waterfield, T., Yelekçi, Ö., Yu, R., and Zhou, B., editors, *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*, pages 1513–1766. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.
- Sillmann, J., Stjern, C. W., Myhre, G., Samset, B. H., Hodnebrog, O., Andrews, T., Boucher, O., Faluvegi, G., Forster, P., Kassoar, M., Kharin, V. V., Kirkevåg, A., Lamarque, J.-F., Olivie, D. J. L., Richardson, T. B., Shindell, D., Takemura, T., Voulgarakis, A., and Zwiers, F. W. (2019). Extreme wet and dry conditions affected differently by greenhouse gases and aerosols. *npj Climate and Atmospheric Science*, 2(1):24.
- Solomatine, D. P. and Ostfeld, A. (2008). Data-driven modelling: some past experiences and new approaches. *Journal of hydroinformatics*, 10(1):3–22.

- Sun, Q., Zhang, X., Zwiers, F., Westra, S., and Alexander, L. (2021). A global, continental, and regional analysis of changes in extreme precipitation. *Journal of Climate*, 34(1):243–258.
- Teufel, B., Diro, G. T., Whan, K., Milrad, S. M., Jeong, D. I., Ganji, A., Huziy, O., Matthews, H. D., and de Elia, R. (2019). Investigation of the mechanisms leading to the 2017 Montreal flood. *Climate Dynamics*, 52(7–8):4193–4206.
- The Inter-Sectoral Impact Model Intercomparison Project (2025). Isimip repository. <https://www.isimip.org>. Consulted on 11 November 2025.
- Thiery, W., Lange, S., Rogelj, J., Schleussner, C.-F., Gudmundsson, L., Seneviratne, S. I., Andrijevic, M., Frieler, K., Emanuel, K., Geiger, T., et al. (2021). Intergenerational inequities in exposure to climate extremes. *Science*, 374(6564):158–160.
- Torres, R., Snoeij, P., Geudtner, D., Bibby, D., Davidson, M., Attema, E., Potin, P., Rommen, B., Floury, N., Brown, M., et al. (2012). Gmes sentinel-1 mission. *Remote sensing of environment*, 120:9–24.
- Tramblay, Y., Villarini, G., and Zhang, W. (2020). Observed changes in flood hazard in Africa. *Environmental Research Letters*, 15(10):1040b5.
- Venkatesan, E. and Mahindrakar, A. B. (2019). Forecasting floods using extreme gradient boosting—a new approach. *International journal of civil engineering and technology*, 10(2):1336–1346.
- Vicente-Serrano, S. M., Beguería, S., and López-Moreno, J. I. (2010). A multiscalar drought index sensitive to global warming: the standardized precipitation evapotranspiration index. *Journal of climate*, 23(7):1696–1718.
- Ward, R. C. and Robinson, M. (2000). *Principles of Hydrology*. McGraw-Hill, London, 4th edition.
- Waseem, H. B. and Rana, I. A. (2023). Floods in pakistan: A state-of-the-art review. *Natural Hazards Research*, 3(3):359–373.
- Wasko, C. and Nathan, R. (2019). Influence of changes in rainfall and soil moisture on trends in flooding. *Journal of Hydrology*, 575:432–441.
- Wessel, B., Huber, M., Wohlfart, C., Marschalk, U., Kosmann, D., and Roth, A. (2018). Accuracy assessment of the global tandem-x digital elevation model with gps data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 139:171–182.
- Woldemeskel, F. and Sharma, A. (2016). Should flood regimes change in a warming climate? The role of antecedent moisture conditions. *Geophysical Research Letters*, 43(14):7556–7563.
- Yan, S. (2016). Understanding LSTM and its diagrams. <https://blog.mlreview.com/understanding-lstm-and-its-diagrams-37e2f46f1714>. Accessed on 11 August 2026.
- Zhang, X., Wan, H., Zwiers, F. W., Hegerl, G. C., and Min, S.-K. (2013). Attributing intensification of precipitation extremes to human influence. *Geophysical Research Letters*, 40(19):5252–5257.

Appendices

Appendix A

Data overview

Table A.1: Overview of data sources

Variable	Source	Available period	Temporal resolution	Spatial resolution
Floods	Global Flood Monitoring	2015-01-01 → 2024-12-31 (active)	Global coverage: 6 days	20 m
Precipitation	CHELSA	1941-01-01 → 2024-12-31 (active)	Daily	30 arcsec
Soil Moisture	Global Land Evaporation Amsterdam Model 4.3a	1980-01-01 → 2024-12-31	Daily	30 arcsec
Total evapotranspiration	Global Land Evaporation Amsterdam Model 4.3a	1980-01-01 → 2024-12-31	Daily	30 arcsec
Elevation	Copernicus DEM	2011 → 2015	Static	90 m
Percentage sand	Harmonised World Soil Database	Static	Static	1 km
Percentage clay	Harmonised World Soil Database	Static	Static	1 km
Bulk density	Harmonised World Soil Database	Static	Static	1 km
Paved	Copernicus Land Monitoring Service	2020	Static	100 m
Short vegetation	Copernicus Land Monitoring Service	2020	Static	100 m
Forest	Copernicus Land Monitoring Service	2020	Static	100 m
Wetlands	Copernicus Land Monitoring Service	2020	Static	100 m

Table A.2: Overview of the data after pre-processing

Variable	Number of files	Number of days	Period	Storage
Floods	899 961	3484	2015-01-01 → 2024-12-31	3.3 TB
Precipitation	1 715 593	3325	2015-01-01 → 2024-12-31	5.9 TB
Soil Moisture	1 884 948	3653	2015-01-01 → 2024-12-31	6.9 TB
Total evapotranspiration	1 884 831	3653	2015-01-01 → 2024-12-31	7.2 TB
Elevation	516	1	Static	2.0 GB
Percentage sand	516	1	Static	39 MB
Percentage clay	516	1	Static	39 MB
Bulk density	516	1	Static	38 MB
Paved	516	1	Static	221 MB
Short vegetation	516	1	Static	437 MB
Forest	516	1	Static	378 MB
Wetlands	516	1	Static	107 MB

Appendix B

Flood dependency on environmental conditions

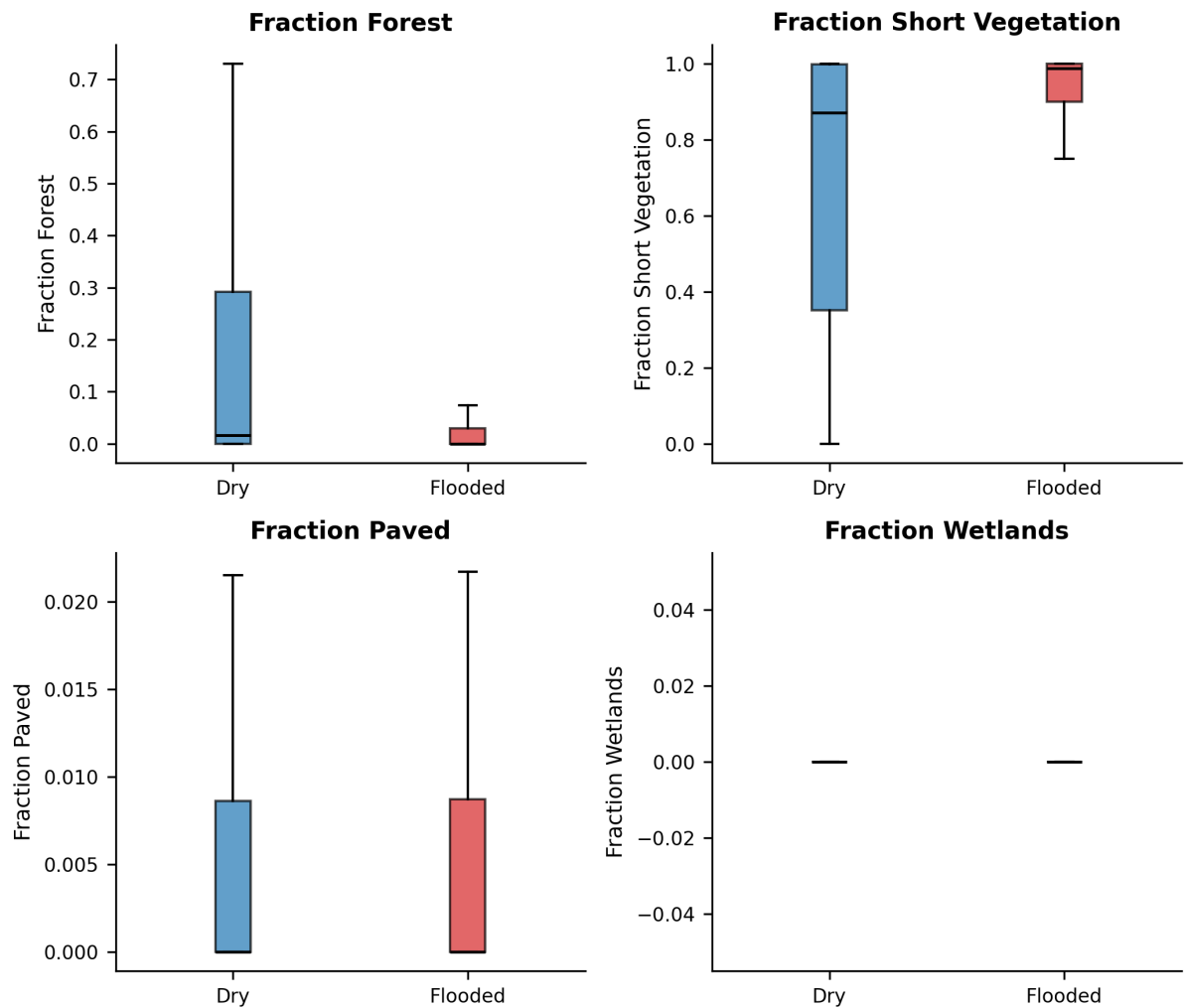


Figure B.1: Boxplots for the dry and the flooded pixels per fraction of the landcover class forest, short vegetation, paved and wetlands.

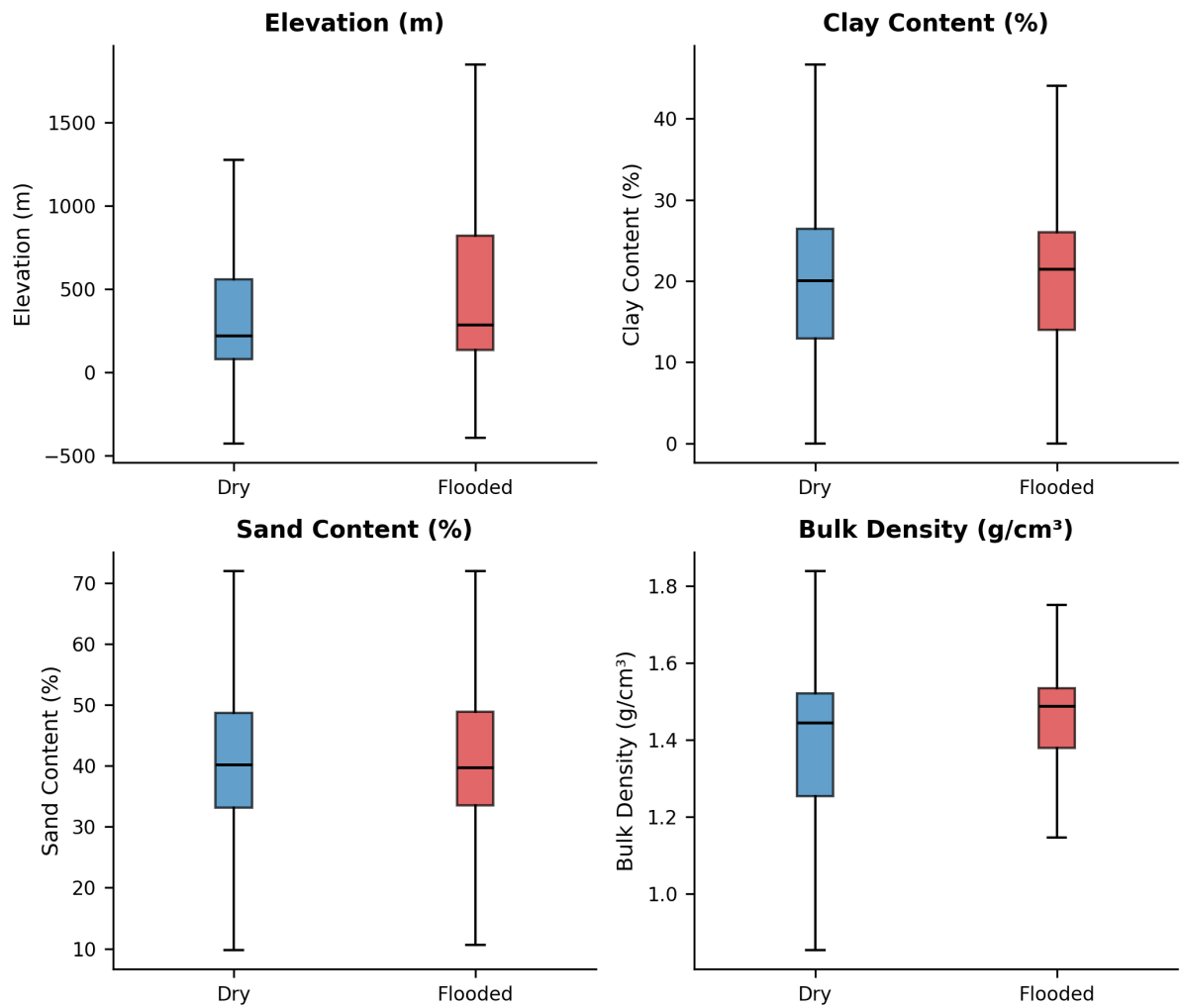


Figure B.2: Boxplots for the dry and the flooded pixels per the environmental variables elevation, clay content, sand content and bulk density.

Appendix C

Model performance

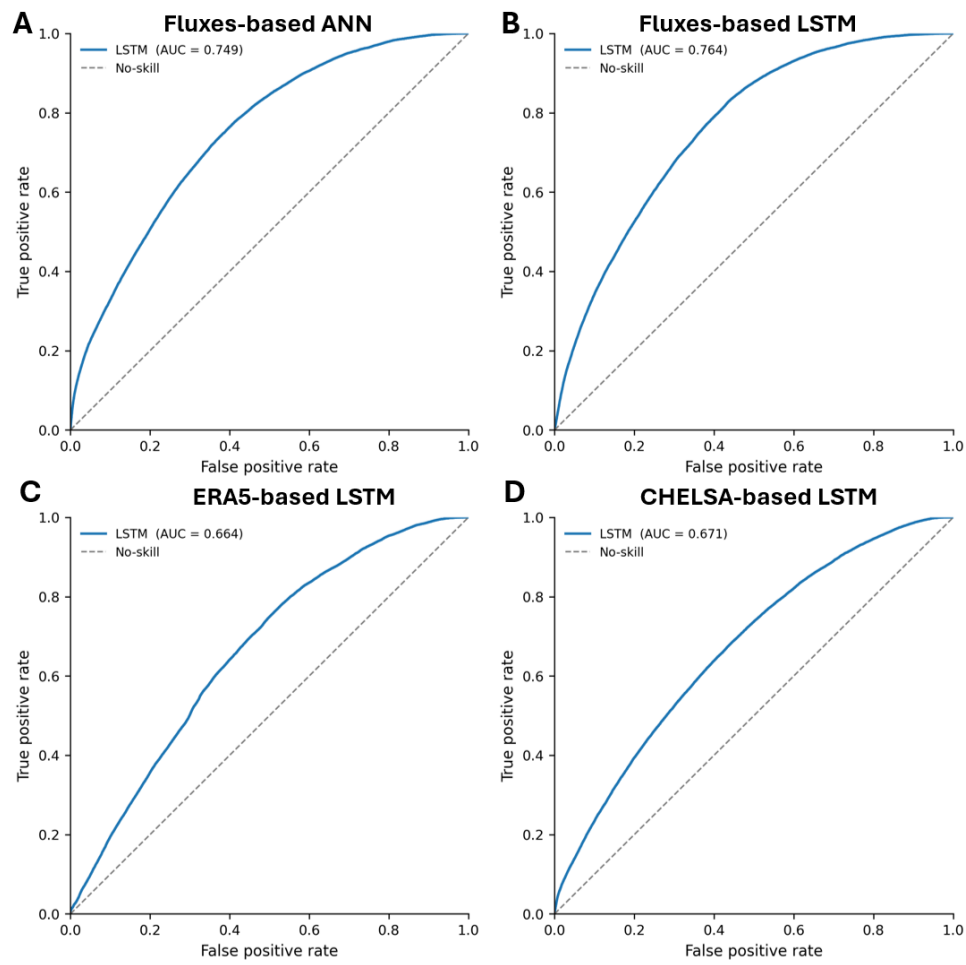


Figure C.1: Comparison of the ROC-curves for the fluxes-based ANN (panel A), the fluxes-based LSTM (panel B), the ERA5-based precipitation-only LSTM (panel C) and the CHELSA-based precipitation-only LSTM (panel D). The dashed grey line represents the random no-skill baseline

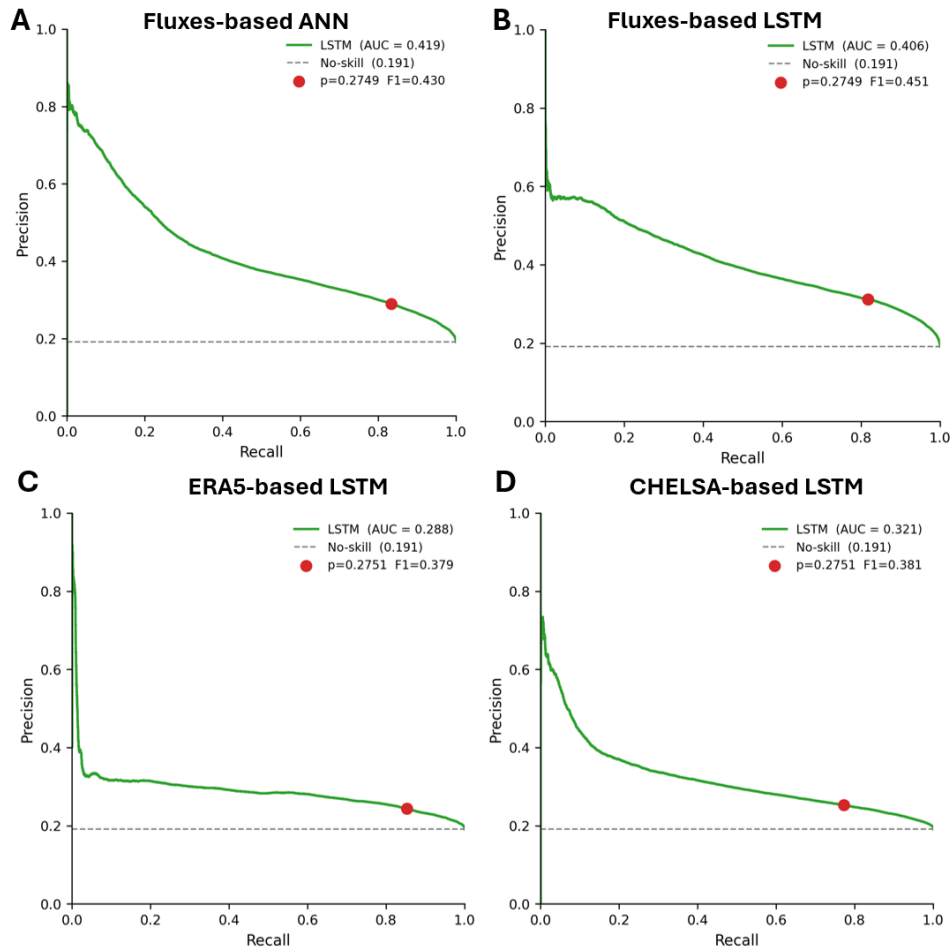


Figure C.2: Comparison of the PR-curves for the fluxes-based ANN (panel A), the fluxes-based LSTM (panel B), the ERA5-based precipitation-only LSTM (panel C) and the CHELSA-based precipitation-only LSTM (panel D). The dashed grey line represents the random no-skill baseline. The red marker highlights the selected decision threshold

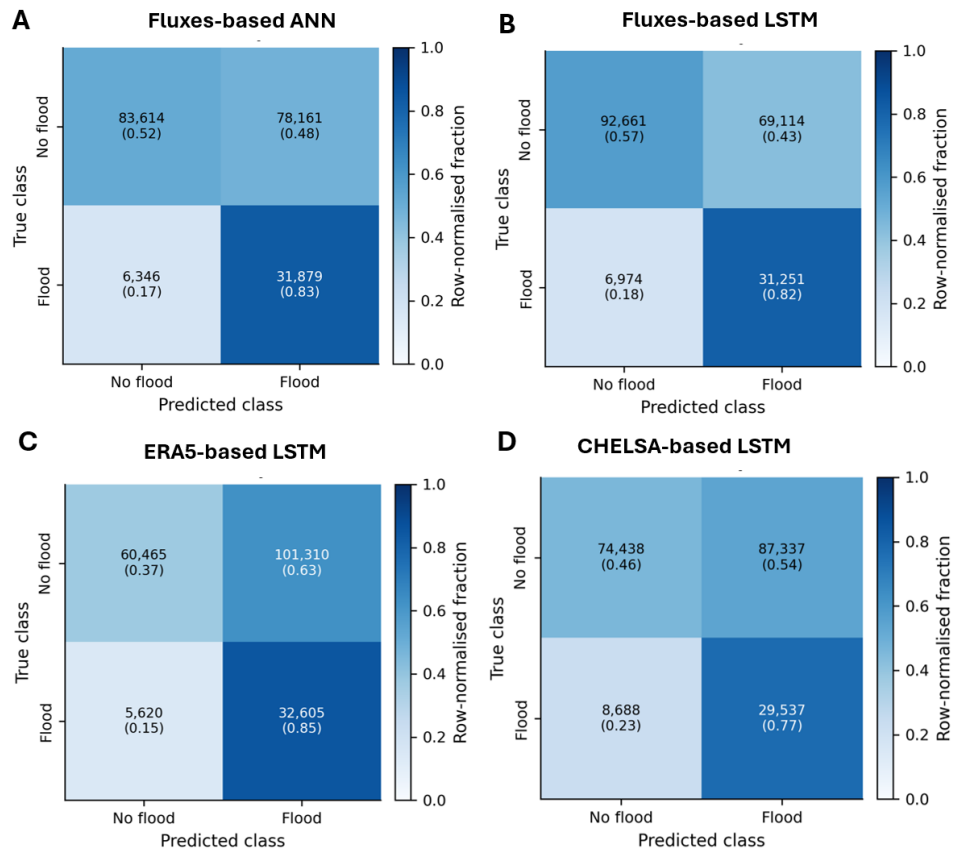


Figure C.3: Comparison of the confusion matrices for the fluxes-based ANN (panel A), the fluxes-based LSTM (panel B), the ERA5-based precipitation-only LSTM (panel C) and the CHELSA-based precipitation-only LSTM (panel D). It is important to note that classification is based on an optimal threshold that is determined for each model separately and therefore differs between the four confusion matrices.

WATER AND CLIMATE
Pleinlaan 2
1050 BRUSSEL, BELGIË
tel. +32 26 29 30 21
www.hydr.vub.be

