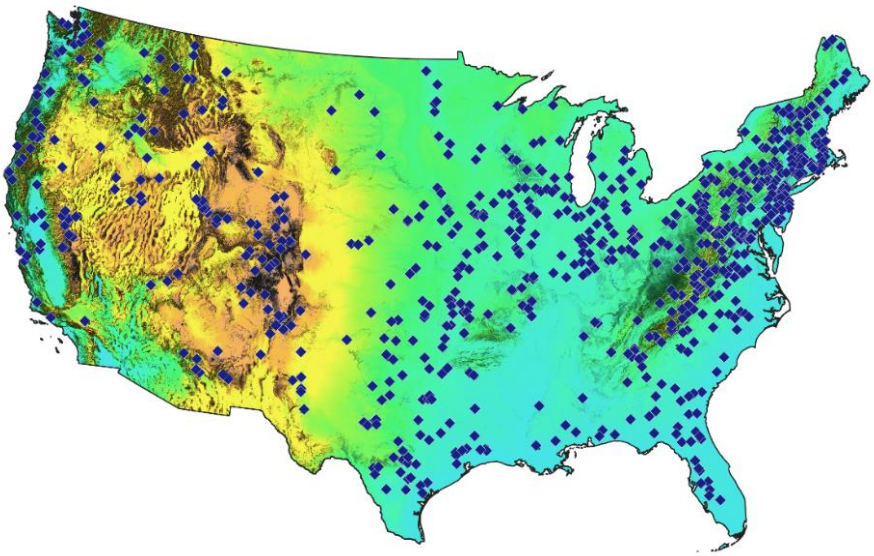
Chair of Hydrology

Albert-Ludwigs-University Freiburg i. Br

Lena Schemel

How do statistical flood thresholds relate to flood stages?

Comparison of recurrence intervals and impact-based thresholds for catchments in the conterminous United States



Master thesis supervised by Dr. Manuela Brunner

Freiburg i. Br., October 2021

Chair of Hydrology

Albert-Ludwigs-University Freiburg i. Br

Lena Schemel

How do practical flood thresholds relate to flood impacts?

Comparison of recurrence intervals and impact-based thresholds for catchments in the conterminous United States

Examiner: Dr. Manuela Brunner

Second examiner: Prof. Dr. Markus Weiler

Master thesis supervised by Dr. Manuela Brunner

Freiburg i. Br., October 2021

Acknowledgments

First and foremost, I would like to thank my thesis supervisor Dr. Manuela Brunner, who provided me with continued feedback and food for thought. Thank you for your encouragement and support and for granting me freedom while researching the topic you provided.

I would also like to thank Prof. Dr. Markus Weiler for agreeing to be the second examiner.

A special thanks to Dr. Bailey Anderson for elaborately answering all of my questions about her research. I would like to express my deepest gratitude to Dr. Jonathan Sheppard for his thorough proofreading and many valuable tips. I would also like to thank my friends, for their helpful comments and the occasional appreciated distraction from work. Last but not least I want to thank my parents for their never-ending support throughout my studies.

Table of Contents

List of Figures

List of Tables

List of Tables in the Appendix

Extended English summary

Obtaining reliable estimates of extreme flows is of increasing importance, the question of how to obtain them is a well-known problem in both applied and scientific hydrology (Okoli et al., 2019). Two popular methods of flood estimation are statistical thresholds and impact-based thresholds. Statistical thresholds are based on discharge time series, defining floods based on their probability of exceedance or recurrence interval (return period). Impact-based thresholds are based on actual observed flood impacts in the vicinity of a stream.

Statistical thresholds are frequently examined and applied in the literature, impact-based thresholds, however, have only been sparsely used in research. Moreover, it is unclear how these statistical thresholds, based on a certain probability of exceedance, relate to actual flood impacts, observed in the area around a stream. The goal of this thesis is, to evaluate, how well statistical thresholds identify impact-triggering events of a certain level.

The US implement impact-based thresholds, classifying the severity of flooding using flood categories divided into a minor stage, moderate stage, and major stage. Each stage is assigned a flow value, triggering the stage, derived from an impact-based threshold. Since observed impacts are not available for all catchments, especially in remote areas, flood stages are assigned return periods corresponding to expected impacts. The minor stage corresponds to a recurrence interval of 5 – 10 years, the moderate stage to a recurrence interval of 15 – 40 years, and the major stage to a recurrence interval of 50 – 100 years.

The relationship between statistical and impact-based thresholds has only been sparsely reported in the literature. This thesis aims to examine said relationship by comparing flood stages, derived from impact-based thresholds, with assigned recurrence intervals, based on statistical thresholds obtained from the discharge time series. Selected catchment characteristics are evaluated, trying to explain the spatial variability of the relationship. Further, it is evaluated if a prediction of the relationship using catchment characteristics is possible.

Stations are selected for the analysis based on the availability of gapless discharge data, flood stage data, and catchment characteristics. A flood frequency analysis using the peak over threshold approach is performed to obtain discharge values (Q_T) of certain return periods (T) and return periods (T_{Stage}) of the flood stage triggering flows (Q_{Stage}). Stations are classified based on the previously calculated discharge values of return periods (Q_T) and the flood stage triggering flows (Q_{Stage}). Due to the range of recurrence intervals assigned to each flood stage, two values of Q_T are calculated for each stage Q_{T_lower} and Q_{T_upper}. The classification represents the relationship between statistical and impact-based thresholds and is divided as follows: *below* = $Q_{Stage} < Q_{T\_lower}$, *in* = $Q_{T\_lower} < Q_{Stage} < Q_{T\_upper}$, *above* = $Q_{Stage} > Q_{T\_upper}$. Selected catchment characteristics are used in correlation and

regression analysis. To select regression models that best describe the relationship for each flood stage, stepwise model selection is performed.

The return period range (statistical threshold) assigned to the different flood stages (impact-based threshold) could not be confirmed, as the flood stages exhibited a wide range of return periods (T_{Stage}). Impact-based thresholds were within the assigned range of statistical thresholds for on average only 8% of stations over all three flood stages. A majority of stations were classified *below* for the minor (71%) and moderate stage (52%) and *above* for the major stage (62%). The Correlation analysis showed only a weak correlation at best between catchment characteristics and the classification, meaning no single catchment characteristic could sufficiently explain the spatial variability of the relationship. The regression models possessed sufficient prediction accuracy (61% – 73%), showing that the relationship can be modeled using catchment characteristics. However, the classification *in* was never correctly predicted by any model. Precipitation is the only parameter included in all final models.

The models tended to over- or underestimate depending towards which classification category the data was skewed. A more even distribution across all categories might improve the prediction accuracy of the models further. The distribution used for the flood frequency analysis showed a poor fit to the tails of the data distribution and with that, poorly estimated high return periods. Despite this, the classification of stations is valid. Additionally, daily mean discharge data was used, resulting in the highest flood peaks being missed out on and because of that, an underestimation of return period flows.

Based on the results, statistical thresholds are not a good alternative to impact-based flood stages, as they do not identify impact-triggering events well. Hydrologists applying statistical thresholds for flood warnings and protection measures must be aware of the discrepancy in the relationship, to avoid over or underestimating floods and their impacts.

Keywords: statistical threshold, impact-based threshold, flood categories, flood stage, flood frequency analysis, peak over threshold, catchment characteristics, ordinal logistic regression, CONUS

Zusammenfassung

Zuverlässige Schätzungen extremer Abflüsse gewinnen zunehmend an Bedeutung, die Frage, wie man diese erhält, ist ein bekanntes Problem, sowohl in der angewandten als auch in der wissenschaftlichen Hydrologie (Okoli et al., 2019). Zwei gängige Methoden der Hochwasserabschätzung sind statistische Schwellenwerte und auswirkungsbasierte Schwellenwerte. Statistische Schwellenwerte basieren auf Abflusszeitreihen, sie definieren Hochwasser anhand ihrer Überschreitungswahrscheinlichkeit bzw. ihres Wiederkehrintervalls (Jährlichkeit). Auswirkungsbasierte Schwellenwerte basieren auf den tatsächlich beobachteten Auswirkungen eines Hochwassers auf die Umgebung eines Flusses.

Statistische Schwellenwerte werden in der Literatur häufig untersucht und angewandt, wirkungsbezogene Schwellenwerte jedoch wurden in der Forschung nur selten verwendet. Darüber hinaus ist unklar, in welcher Beziehung die statistischen Schwellenwerte, die auf einer bestimmten Überschreitungswahrscheinlichkeit basieren, zu den tatsächlichen Hochwasserauswirkungen stehen, die in der Umgebung eines Flusses beobachtet wurden. Ziel dieser Arbeit ist es zu bewerten, wie gut statistische Schwellenwerte Auswirkungen von Hochwasserereignissen verschiedener Größen identifizieren.

In den USA werden auswirkungsbasierte Schwellenwerte angewendet, das Ausmaß von Überflutungen wird anhand von Hochwasserkategorien klassifiziert, welche in minor stage, moderate stage und major stage unterteilt werden. Jeder Hochwasserstufe (flood stage) wird ein Abflusswert zugewiesen, der die Stufe auslöst, ein auswirkungsbasierter Schwellenwert. Da nicht für alle Einzugsgebiete, insbesondere in abgelegenen Gebieten, beobachtete Auswirkungen vorhanden sind, werden den Hochwasserstufen Jährlichkeiten zugeordnet, die den erwarteten Auswirkungen entsprechen. Die minor stage entspricht einem Wiederkehrintervall von 5 - 10 Jahren, die moderate stage einem Wiederkehrintervall von 15 - 40 Jahren und die major stage einem Wiederkehrintervall von 50 - 100 Jahren.

Die Beziehung zwischen statistischen und auswirkungsbasierten Schwellenwerten ist in der Literatur nur spärlich beschrieben. In dieser Arbeit soll diese Beziehung untersucht werden, indem Überschwemmungskategorien, die von auswirkungsbasierten Schwellenwerten abgeleitet wurden, mit zugewiesenen Wiederkehrintervallen verglichen werden. Die Wiederkehrintervalle basieren auf statistischen Schwellenwerten, welche anhand von Abflusszeitreihen berechnet wurden. Anhand ausgewählte Einzugsgebietseigenschaften wird bewertet, ob diese verwendet werden können, um die räumliche Variabilität der Beziehung zu erklären. Des Weiteren wird geprüft, ob eine Vorhersage der Beziehung zwischen statistischen und auswirkungsbezogenen Schwellenwerten anhand von Einzugsgebietseigenschaften möglich ist.

Die Stationen wurden auf Grundlage von Verfügbarkeit lückenlosen Abflussdaten, Abflusswerten der Hochwasserkategorien und Einzugsgebietseigenschaften für die Analyse ausgewählt. Es wird eine Hochwasserhäufigkeitsanalyse unter Verwendung des Peak-over-Threshold-Ansatzes durchgeführt, um Abflusswerte (Q_T) bestimmter Wiederkehrperioden (T) und Wiederkehrperioden (T_{Stage}) der stufenauslösenden Abflusswerte (Q_{Stage}) zu erhalten. Die Stationen werden auf der Grundlage der zuvor berechneten Abflusswerte der Wiederkehrintervalle (Q_T) und der stufenauslösenden Abflusswerte (Q_{Stage}) klassifiziert. Aufgrund des Bereichs der Wiederkehrintervalle, die jeder Hochwasserstufe zugeordnet sind, werden für jede Stufe zwei Werte für Q_T berechnet: Q_{T_lower} und Q_{T_upper}. Die Klassifizierung stellt das Verhältnis zwischen statistischen und auswirkungsbezogenen Schwellenwerten dar und ist wie folgt unterteilt: *below* = $Q_{Stage} < Q_{T\_lower}$, *in* = $Q_{T\_lower} < Q_{Stage} < Q_{T\_upper}$, *above* = $Q_{Stage} > Q_{T\_upper}$. Ausgewählte Einzugsgebietseigenschaften werden in Korrelations- und Regressionsanalysen verwendet. Um diejenigen Regressionsmodelle auszuwählen, die die Beziehung für jede Hochwasserstufe am besten beschreiben, wird eine schrittweise Modellauswahl durchgeführt.

Der Bereich der Wiederkehrperioden (statistischer Schwellenwert), der den verschiedenen Hochwasserstufen (auswirkungsbezogener Schwellenwert) zugeordnet wurde, konnte nicht bestätigt werden, da die Hochwasserstufen einen große Wertebereich an Wiederkehrperioden (T_{Stage}) aufwiesen. Die auswirkungsbezogenen Schwellenwerte lagen nur bei durchschnittlich 8 % der Stationen innerhalb des zugeordneten Bereichs der statistischen Schwellenwerte. Die Mehrheit der Stationen wurde für die minor (71 %) und moderate stage (52 %) als *below* und für die major stage (62 %) als *above* klassifiziert. Die Korrelationsanalyse zeigte bestenfalls schwache Korrelation zwischen den Einzugsgebietseigenschaften und der Klassifizierung, was bedeutet, dass keine einzelne Einzugsgebietseigenschaft die räumliche Variabilität der Beziehung ausreichend erklären konnte. Die Regressionsmodelle wiesen eine ausreichende Vorhersagegenauigkeit auf (61 % - 73 %), was zeigt, dass die Beziehung anhand von Einzugsgebietseigenschaft modelliert werden kann. Die Klassifizierung *in* wurde jedoch von keinem Modell korrekt vorhergesagt. Der Niederschlag ist der einzige Parameter, der in allen finalen Modellen enthalten war.

Die Modelle neigten zu Über- oder Unterschätzungen, je nachdem, zu welcher Klassifizierungskategorie die Daten gewichtet waren. Eine gleichmäßigere Verteilung der Daten über alle Kategorien könnte die Vorhersagegenauigkeit der Modelle weiter verbessern. Die für die Analyse der Überschwemmungshäufigkeit verwendete Verteilung zeigte eine schlechte Anpassung am hinteren Ende der Daten und damit eine schlechte Schätzung hoher Wiederkehrperioden. Trotzdem ist die Klassifizierung der Stationen gültig. Darüber hinaus wurden tägliche Mittelwerte der Abflüsse verwendet, was dazu führte, dass die höchsten Hochwasserspitzen nicht berücksichtigt wurden und die Abflüsse der Wiederkehrperioden daher unterschätzt wurden.

Auf der Grundlage der Ergebnisse sind statistische Schwellenwerte keine gute Alternative zu auswirkungsbasierte Schwellenwerten (flood stages), da sie die Auswirkungen von Hochwasserereignissen verschiedener Größen nicht gut identifizieren. Hydrologinnen und Hydrologen, die statistische Schwellenwerte für Hochwasserwarnungen und Schutzmaßnahmen anwenden, müssen sich der Diskrepanz in der Beziehung bewusst sein, um eine Über- oder Unterschätzung von Hochwassern und deren Auswirkungen zu vermeiden.

Stichworte: statistische Schwellenwerte, auswirkungsbasierte Schwellenwerte, Hochwasserstufe, flood stage, Hochwasserhäufigkeitsanalyse, peak over threshold, Einzugsgebietseigenschaften, ordinale logistische Regression, CONUS

1 Introduction

The number of reported floods has increased by a factor of ten between 1950 and 2010, based on data from the international disaster database. Economic loss and insured damages caused by floods are increasing, with flood damage causing the highest losses in the US, compared to other natural hazards (Gall et al., 2011; Jha et al., 2012; Zhou et al., 2017).

On one hand, there is an increasing population density in urban settlements and heightened exposure to floods (Changnon et al., 2000; Slater and Villarini, 2016). On the other hand, the characteristics of floods are changing due to dynamic catchment characteristics like climate variables, land use, land cover, and anthropogenic modifications (O'Driscoll et al., 2010; Slater and Villarini, 2016; Saharia et al., 2017; Villarini and Slater, 2017; Hounkpè et al., 2019).

Flow records of catchments in the central US show an increase in flood frequency but not flood magnitude (Hirsch and Archfield, 2015; Mallakpour and Villarini, 2015). Regarding the entire conterminous US, trends in flood characteristics strongly vary spatially with both increases and decreases of flood frequencies, increases of all flood properties, and minimal changes to flood properties being observed (Archfield et al., 2016; Slater and Villarini, 2016).

Because of the variability of catchment characteristics, heightened exposure to flooding, and lack of generalization of flood changes, reliable flood estimations are increasingly important, however, obtaining them is a well-known challenge of scientific and applied hydrology (Okoli et al., 2019).

Flooding requires multiple responses on different scales. The construction of buildings and bridges adapted to flooding on the local scale as well as damns and retention basins on the catchment scale. Land use management in terms of municipal planning to reduce flood exposure and landscape changes to reduce surface runoff and increase natural retention. Flood forecasts both short and long term, to warn people of coming floods and to implement building restrictions in inundation areas.

The aforementioned flood risk assessment and the consequent responses are based on the results of a flood frequency analysis (Kidson and Richards, 2005). For the analysis, a time series of flow values is used to relate discharge magnitudes to their probability of being equaled or exceeded in any year. That probability is then converted to a recurrence interval or a return period, which equals the estimated average time between events (Archer, 1998). Said probabilities of exceedance and within that return periods are called statistical thresholds. They are theoretical, meaning for example in any given 50-year period, an event with a return period of 50 years can be exceeded more than once or not at all.

These statistical thresholds are, for example, used for the construction of flood defense structures, as those have to provide protection against floods of a specific return period (Kidson and Richards, 2005). They are also used to restrict development areas, as seen in the German water law (WG) and

Water Management Act (WHG). For example, as given in §65 WG, where flood areas are designated, where statistically flooding is expected every 100 years and §78 restricts the construction of buildings in those areas.

In Baden-Wuerttemberg, the State Institute for the Environment (LUBW) provides an interactive map for many streams, showing areas of potential inundation for floods of different return periods, based on statistical flood height estimations and hydraulic calculations (LUBW, 2021). The Flood Forecast Centre of the LUBW classifies floods using return periods (HVZ, 2021), the same classification is used for example in Austria, France, Luxemburg, and the Netherlands (LfU, LUBW, 2018).

An alternative to statistical thresholds based on flow time series are impact-based thresholds, classifying flood magnitudes based on the impact of a flood on the area around the stream. All German states (except for Rhineland-Palatinate and Baden-Wuerttemberg) apply these impact-based flood classifications, as well as countries like Switzerland, Czechia, and the USA (LfU, LUBW, 2018).

In the US the impact-based flood categories of the National Weather Service (NWS) are divided into four flood stages, based on the severity of expected flood impacts in the stream reach. A stage is the water level above an arbitrary reference point, also known as gauge height. Flood stage is an established water level, above which rising water will cause inundation to the surrounding areas of a stream, impacting the population, property, and commerce. Reach refers to a section of the stream, for which the stage measured at a gauge is representative of the conditions NOAA, 2019).

The water level assigned to a flood stage is based on specific observed impacts. The severity of flooding at a given stage varies throughout the reach, due to differing characteristics of the surrounding areas. Because of this, the stage of a flood category is chosen depending on the most significant flood impact within the reach (APRFC, 2021).

The flood categories are divided into action stage, minor stage, moderate stage, and major stage. If the action stage is reached, the gauge is monitored closely and preparations for a possible flood are made. At the minor stage minimal to no property damage is caused, areas near the stream (e.g. roads, trails, yards, campgrounds) may become flooded possibly causing a public threat. Reaching the moderate stage some structures and roads near the stream may become flooded, evacuations and transfer of property to higher elevations might be necessary, disrupting daily life. If the major stage is reached, extensive flooding of structures, roads, and critical infrastructure (e.g. hospitals, schools, police, and fire stations) is to be expected, leading to significant evacuations and transfer of property to higher elevations (NOAA, 2019).

The NWS also provides an interactive map, showing inundation caused by different flood categories for several stations, predominately located in the eastern US. Both river gauge observations and forecasts are classified into the flood categories named above, data is available for 3823 gauges all over the US (NOAA, 2021).

Statistical thresholds and discharge records have been used for many studies, a small selection is named in the following. For analyzing trends in discharge over time, as elaborated above (Hirsch and Archfield, 2015; Mallakpour and Villarini, 2015; Archfield et al., 2016; Slater and Villarini, 2016). Or for examining of the influence of catchment characteristics on flood magnitude and recurrence: Hounkpè et al. (2019) reported an increase in flood characteristics with the expansion of agricultural land, O'Driscoll et al. (2010) found urbanization results in an increase in stream stage overall and an increase in peak flows and Davenport et al. (2020) named snowpack as a natural reservoir, that reduces winter floods, with an increase of rain fraction leading to larger streamflow peaks.

Flood stage data has been used by Slater and Villarini (2016) to assess trends in inundation frequencies. They found regional patterns of both increasing and decreasing flood risk, that are overall dependant on wetness and potential water storage of a catchment. Saharia et al. (2017) used flood stage exceedances of US catchments to examine how floods vary with climate, topography, and geomorphology. They found, that the seasonality of flooding varies greatly over the US. Precipitation is named as the primary driver of floods, with the magnitude of floods being highest in areas with the greatest precipitation.

Statistical thresholds are frequently examined and applied in the literature, as the short compilation of literature showed, impact-based thresholds, however, have only been sparsely used in research. More importantly, it is unclear, how statistical thresholds, derived from a discharge time series, relate to actual observed flood impacts.

For Alaskan remote areas with few specific impacts, flood stage values are determined using recurrence intervals, giving each stage a reference range of return periods assumed to cause the corresponding amount of flooding. The minor stage corresponds to a recurrence interval of 5 – 10 years, the moderate stage to a recurrence interval of 15 – 40 years, and the major stage to a recurrence interval of 50 – 100 years (APRFC, 2021). Figure 1.1 depicts the relationship between the water level (stage) and flood categories at a gauge. A specific flood stage refers to all flows between the stage triggering flow up until the next flood stage is triggered. While the minor and moderate stages eventually transition into the next higher stage, if the water level keeps rising, the major stage refers to all flows above the flood stage triggering flow and has no upper limit. The recurrence intervals named above correspond to the triggering flow of a flood stage.

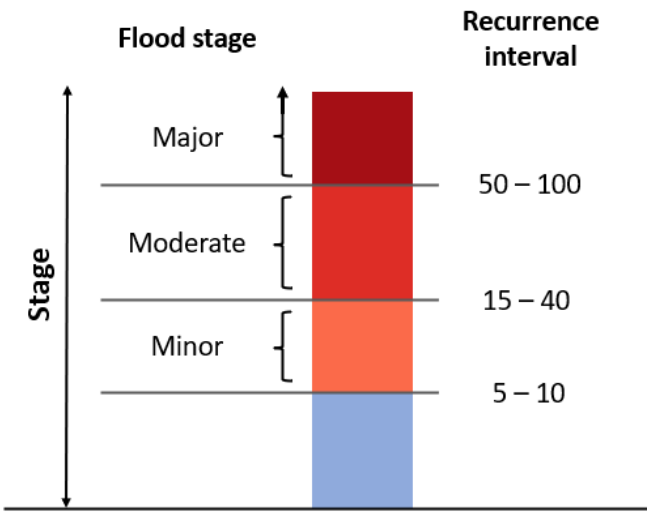


Figure 1.1: Meaning of the flood categories (flood stages, left) and assigned recurrence intervals(right). Adapted from NOAA (2019)

This assignation of statistical thresholds (recurrence intervals) to impact-based thresholds (flood stages) was used by Anderson (2016). She examined the relationship between flood stages and stages associated with the aforementioned recurrence intervals for catchments in Alaska. Her analysis showed a strong relationship between statistical and impact-based thresholds, results differing from the APRFC pairing especially on the upper end. The moderate stage best fits the 25-year recurrence interval and the major stage the 100 – 500-year recurrence interval. She linked the deviation for the major stage to the rarity of the events and the limited available data in Alaska. Catchment characteristics could not be used to explain the outliers found.

Anderson´s (2016) analysis was limited to a small number of around 40 catchments in Alaska and she only anecdotally analyzed the outliers found in the relationship. The goal of this thesis is to analyze the relationship between impact-based flood stages and statistical recurrence intervals on a larger scale using catchments of the conterminous United States and a more in-depth examination of the catchment characteristics influencing the relationship.

2 Aim and research questions

The relationship between statistical flood thresholds and flood stages used to classify inundation impacts has not been widely reported. This thesis aims to compare recurrence intervals and impact-based thresholds, to evaluate to what extend statistical thresholds can identify impact-triggering events of a certain level. Additionally, the relationship between statistical thresholds and flood stages is to be examined regarding variability in space and by hydro-climatic catchment characteristics.

Research Questions:

I What is the relationship between statistically calculated flood thresholds and implemented flood stages using gauge height as an indicator for flooding?

II How does said relationship vary across all catchments and which hydro-climatic parameters can be used to explain that variability?

III Can catchment characteristics be used to predict the relationship between recurrence intervals and impact-based thresholds?

IV To which extend can statistical thresholds be used to classify the impact of flooding? Are statistical thresholds a good alternative to impact-based flood stages?

3 Data and methods

The methodical approach of this thesis is as follows. First stations providing sufficient data are selected for the analysis. A flood frequency analysis is performed, to obtain discharge values for the later comparison. Stations are classified based on the relationship between statistical and impact-based thresholds, derived from values previously calculated. Relevant catchment criteria are selected and used in correlation and regression analysis. A stepwise model selection is performed, to choose regression models that best describe the relationship between statistical thresholds and flood stages.

3.1 Study area and available data

This study focused on gaging stations and catchments in the conterminous United States of America (CONUS). CONUS refers to the 48 States and the District of Columbia, excluding Alaska and Hawaii (USGS, 2021). They occupy an area of 8,081,866 km^2 (U.S. Census Bureau, 2018).

Annual average state-wide precipitation between 1990 and 2020 ranges from 10,16 inches (258.06 mm) in Nevada to 59.69 inches (1516.13 mm) in Louisiana, with the national average being 31.34 inches (796.04 mm). Annual average state-wide temperatures between 1990 and 2020 vary from 5°C in North Dakota to 21.94°C in Florida, the national average being 11.78°C (NOAA, 2021a, 2021b).

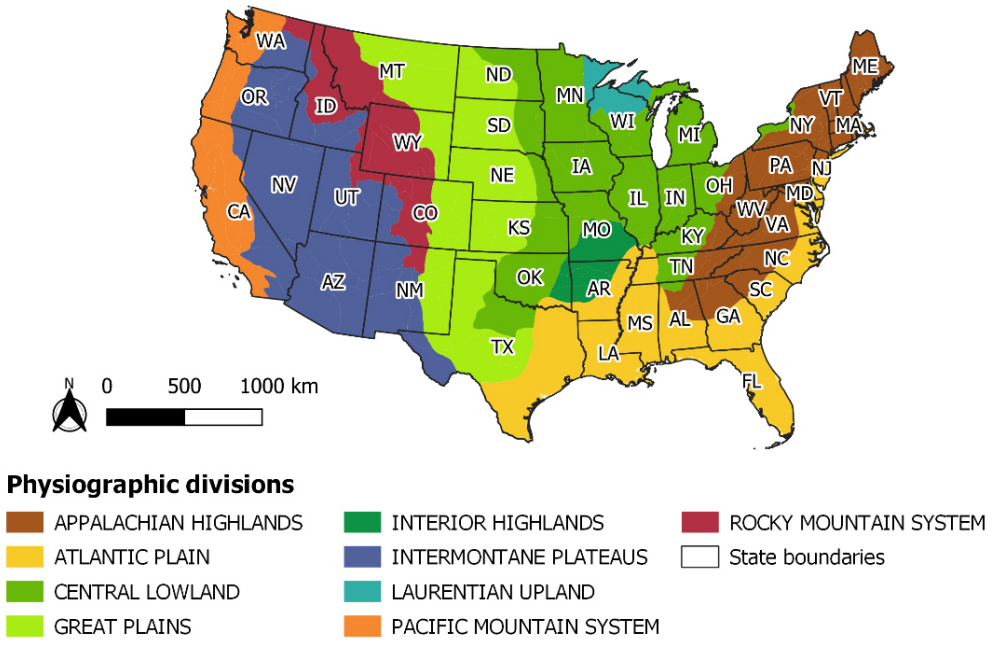


Figure 3.1: Map of the physiographic divisions of the conterminous US, showing the states. Created with QGIS, adapted from Fenneman and Johnson (1946).

Figure 3.1 depicts the physiographic divisions of CONUS, the Appalachian Highlands in the east, separating the Atlantic Plain from the Great Lakes and the Central Lowland. Further west the flat Great Plains transition to the mountain ranges of the Rocky Mountains, extending north to south. Between the Rocky Mountain System and the Pacific Mountain System in the west lies the

Intermontane Plateaus, a large, arid desert consisting of smaller mountain ranges and plateaus. The Cascades and Sierra Nevada ranges from the eastern border of the Pacific Mountain System and are followed by a series of valleys and low mountain ranges to the west.

With its large size and high physiographic variability, CONUS exhibits a wide range of climate types and consequent ecoregions which leads to greatly varying catchment properties. Figure 3.2 depicts a summary of that variability in the form of hydrologic landscape regions (HLR). US watersheds were grouped into regions based on similarities in climate variables, land-surface form, and geologic texture. The lands-surface form was divided into plains (1-8), plateaus (9-13), playas (14), and mountains (15-20). Geologic texture referred to the permeability of soil and bedrock, and climate similarities were grouped in very humid, humid, sub-humid, semi-arid, and arid (Wolock, 2003). The description of each region can be found in the appendix (Table A-1).

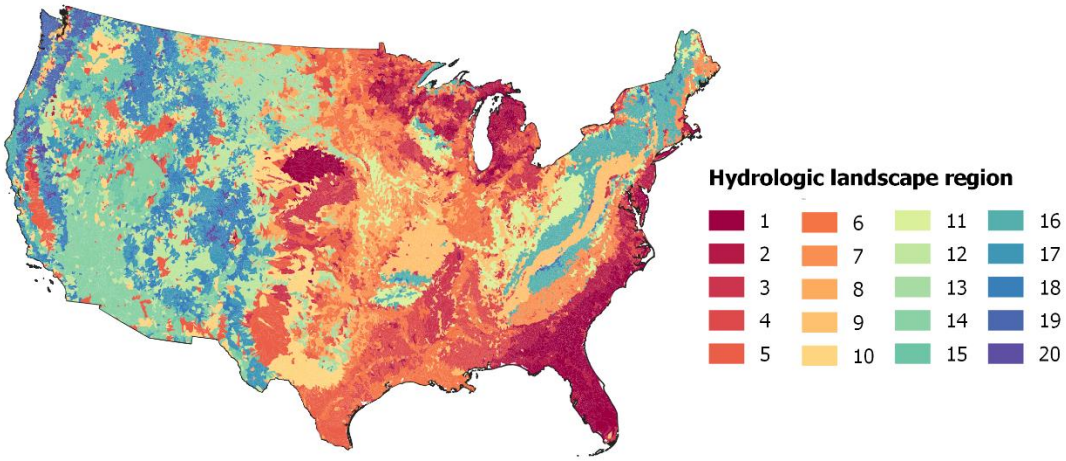


Figure 3.2: Map of the hydrologic landscape regions of the conterminous US. Created with QGIS, using data adapted from Wolock (2003)

The data used within this thesis consisted of

(1) Daily mean discharge data (ft^3/s) for catchments monitored by the United States Geological Survey (USGS)

(2) Rating tables for active USGS stream gauges obtained from the USGS

(3) Flood stage database from the National Weather Service of the US (NWS), giving flood stage corresponding gauge height in ft.

(4) Catchment characteristics for 9,322 stream gauges maintained by USGS from the GAGES II database

(5) Geodata for the conterminous US, catchment boundaries, gauge locations, and state boundaries

In (2) tables were given with gauge height in feet and corresponding streamflow in ft^3/s. In (3) data was given for 10,365 Stations. In (4) Data for 354 catchment characteristics was given including basin identification (basin ID, station name, coordinates for the gauge, drainage area, State at gauge location), environmental features (e.g. climate, geology, hydrology, soils, topography), and anthropogenic influences (e.g. land use, presence of damns/ canals, population density, impervious surfaces). This data was compiled only for gauges with 20+ years of complete discharge record since 1950 or currently active gauges with at least 50 years of discharge data in 2009 (Falcone et al., 2010; Falcone, 2017).

R Studio with R version 3.6.0 was used for the analysis of the data and the creation of graphics and maps unless otherwise specified (R Core Team, 2019).

3.2 Methods

3.2.1 Data preparation

3.2.1.1 Selection of stations

The USGS collects water-level data at more than 10,000 stream gauges, to determine which of those are suitable for this research, the following selection steps were taken.

First valid station IDs were extracted from the NWS database in (3). Stations without a USGS ID or no number IDs were excluded, leaving only IDs consisting of 1-digit to 15-digit numbers. Since USGS station IDs have 8-digits, a 0 was added to all IDs with less than 8-digits. Finally, those stations that now have an 8-digit ID code were extracted. This left 7,345 gaging stations, with existing flood stage levels and streamflow measurements.

Since the discharge is given in ft^3/s, water levels of the flood stages needed to be converted to discharge, using the rating curves given in (2). For this, rating curves were obtained for all stations extracted from (3) and checked for data availability, furthermore, using only those stations where rating data is provided. 5,457 of the selected stations had rating curves available.

Going back to (3), out of the 5,457 stations with rating curves, only stations with water levels for minor, moderate, and major flood stages were selected, resulting in 3,440 stations remaining.

For later inclusion of catchment characteristics in the analysis, it was checked which of the previously selected 3,440 stations were included in (4), leading to 2,840 stations with corresponding characteristics available.

In order to calculate statistical return periods of certain annularity, a long enough time series of streamflow was required. The rule of thumb is, to not estimate return periods that are more than 3 times the streamflow time series length (Meylan et al., 2012). With return periods of the major stage

being up to 100 years, at least 30 years of gapless streamflow data were needed. In addition, at least 500 stations were desired for the analysis, to ensure a sufficient sample size and covering the spatial variability mentioned earlier. The data availability information for each station includes record count in days of discharge, which was then converted to full years. The goal of selecting stations was to have as long a time series as possible, while still having a sample of at least 500 stations. The number of stations that had discharge data of a certain number of years of record can be seen in Table 3.1.

Table 3.1: Evaluation of available data per station: time series length and number of stations

Years of record	Number of stations
50	1986
60	1720
70	1440
80	1079
90	624
100	244

Table 3.1 shows, that the number of stations with 100 years of record is below the desired sample size, however, 90 years of record or less meet our desired number of stations. The discharge data of stations with 80 and 90 years of record was then examined more closely, checking for discharge below 0, the discharge being NA and, missing days in the time series.

Table 3.2 shows the number of stations removed due to missing values, resulting in 727 stations with 80 years and 448 stations with 90 years of gapless discharge data. The stations with 80 years of gapless data were chosen as the final stations for the analysis.

Table 3.2: Results of checking for gaps and missing values in time series of discharge

	80 years	90 years
Q < 0	0	0
NA flow values	14 Stations	7 Stations
Missing days	340 stations	175 stations

Before later converting flood stage level in ft to ft^3/s using the rating curve, it was verified that the logarithmic expansion was the right fit to model the rating data for all stations. This was not the case for one station, which was then removed, resulting in the final 727 stations with 80 years of gapless data and 448 stations with 90 years of gapless data. The location of the respective stations can be seen in Figure 3.3.

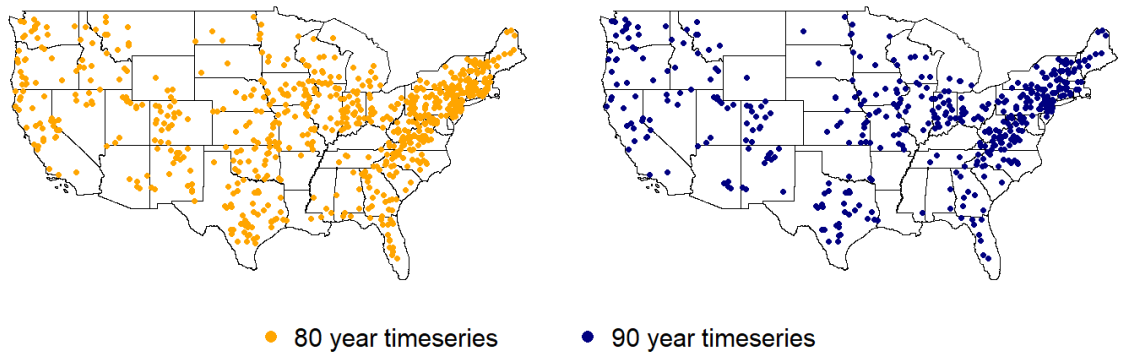


Figure 3.3: Map depicting the location of the selected USGS stations with 80 (left) and 90 (right) years of gapless data

The above figure shows, that the 80-year time series adds stations in states, where there had been only very few before (e.g. Nebraska, Wyoming, South Dakota, North Dakota). In addition, the station count is above 500, therefore, the 727 stations with an 80-year time series will be used for the analysis.

For said 727 stations discharge data was retrieved (01.11.1930 - 31.10.2020) and discharge values were converted from ft^3/s to mm/d using formula (3.1), where the catchment area A in km^2 is retrieved from the GAGES II database in (4).

$$Q\left[\frac{mm}{d}\right] = Q\left[\frac{\left(\frac{l}{m^2}\right)}{d}\right] = \frac{\left(Q\left[\frac{ft^3}{s}\right] * 0.02831685 * 1000 * 86400\right)}{A * 1000000} \qquad (3.1)$$

Climate, elevation, and drainage area of the selected 727 catchments are summarized in Table 3.3, the data is taken from the GAGESII database (Falcone et al., 2010; Falcone, 2017). Precipitation and Temperature are mean annual values for the entire watershed. The elevation refers to the median watershed elevation in meters above sea level.

Table 3.3: Summary of climate and elevation of the selected 727 catchments

	Precipitation [cm/a]	**Temperature [°C]**	**Elevation [m]**	**Drainage area [km^2]**
Minimum	32.2	0.16	22	10.6
Median	106.4	9.7	395	2252.7
Maximum	320.4	22.5	3299	49264.4

3.2.1.2 Calculating flood stage triggering flows

To convert the flood stage level from ft to ft^3/s a log-linear model was fitted (formula (3.2)) for each station's rating curve, where y was streamflow in ft^3/s and x was gauge height in feet. In R the *lm* function (formula (3.3)) from the *stats* package (R Core Team, 2019) was used.

$$\log(y_i) = \alpha + \beta \cdot x_i \tag{3.2}$$

$$model = lm(\log(y) \sim x) \tag{3.3}$$

For each station model, the *predict* function from the *stats* package was used with the model coefficients obtained from the *lm* function. Predict used formula (3.2) to calculate flood stage flow in ft^3/s for all flood stages (action, minor, moderate, major). For the transformation in mm/d formula (3.1) was used.

3.2.2 Flood frequency analysis

In previous steps, discharge data and flood stage discharge values for 727 stations were obtained. To examine the relationship between those flood stage flows and statistical return periods a flood frequency analysis was performed on the extreme values of the time series. Stationarity was assumed.

Flood records are regarded as random samples from a homogeneous population of flows with the assumption, that the record provides a reasonable approximation of the "true" probability distribution which generated the historic records. This means that the flood discharge magnitude Q is related to the exceedance probability of Q and its return period (T) (Archer, 1998). By fitting an appropriate distribution function to the record of flows from a gauge, we can estimate Q corresponding to a certain return period, but also estimate the return period of a particular Q.

For this estimation to be accurate, a sufficient number of peaks and a long enough measurement period must be included in the dataset, to adequately represent the flood frequency of different magnitude events. Since only the extreme values are to be examined, the entire flood record is not used. Instead, a number of extreme values have to be selected, using either the annual maximum flows (AMF) or peaks over a threshold (POT) (see Figure 3.4).

AMF (orange) refers to the maximum peak flow of each year and is the most widely used method in flood frequency analysis. However, several things have to be kept in mind when using this approach: Only a small number of flood peaks are considered, the number of which depends on the length of the record. Because of that, the return periods of smaller floods (<10 years) cannot accurately be estimated (Edwards et al., 2019). Low discharge values may be an annual maximum, in a year with overall smaller discharge values. The second-largest flood of a year can be greater, than another

year's maximum discharge, resulting in high peaks not being taken into consideration, because they are not the annual maximum.

To make up for some of the problems with using only AMF mentioned above an alternative is to use the POT (green) approach. Here a partial-duration series composed of peak flows that equal or exceed a specified threshold value, is used for the analysis. This approach requires continuous streamflow measurements, that allow the identification of individual storm peak flows. If two peaks occur too closely, meaning they are not independent (black), only the greater peak flow is included. The two major difficulties of this approach are assuring the independence of peaks chosen for the analysis and choosing an appropriate threshold value.

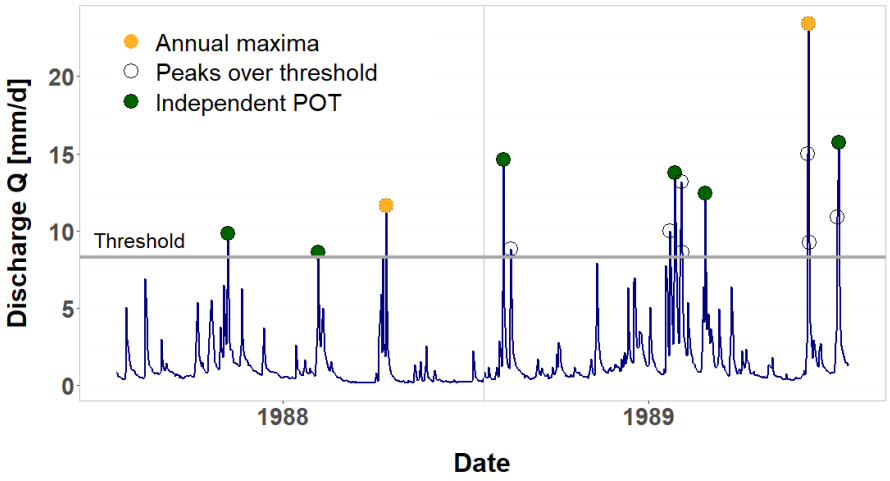


Figure 3.4: Possible peaks selected for the flood frequency analysis: Orange being the annual maxima (AMF), the black circles the peaks over the threshold, and green independent peaks over the threshold

The figure above shows an exemplary time series on which both the AMF and POT approach were performed. It illustrates the difficulties of both approaches: The AMF (orange) of 1988 is lower than the POT (green) in 1989, meaning 4 higher peaks in 1989 would be disregarded when using AMF´s only. Additionally using POT, there are 6 more peaks to use for the calculations, allowing a more accurate representation of lower frequency floods. Looking at the POT, it can be seen that several events have been excluded (black), to ensure independence. In regards to the threshold chosen, varying it up or down would change the number of POT events.

Several distributions are suitable for modeling the relationship between Q and T, based on the approximation of the asymptotic behavior of the observed extreme values. Extreme value theory states that the AMF follow a generalized extreme value distribution (GEV), while POT follow a generalized Pareto distribution (GPD) (Coles, 2001).

3.2.2.1 Peak over threshold approach

As our goal was to examine the recurrence of flood stage exceedance, using only one flow per year would neglect many relevant peaks, where stage triggering flows were surpassed. This would lead

to a probability distribution function that poorly estimates exceedance probability and return periods. Moreover, the analysis of return periods was not limited to recurrence intervals above ten years, therefore, the peak over threshold approach was used.

3.2.2.1.1 Independence criterium

Two peaks were considered independent from each other if they were separated by a certain time interval. Of the dependent peaks, only the larger one was used, since only the highest flow of an event was of interest. The purpose of the time interval was to ensure, that the flow had receded enough, for the subsequent flood peak not to occur on the recession curve of the previous flood peak. For the determination of the time interval or time lag, the catchment size was included. Saharia et al. (2017) showed for CONUS catchments, that the flooding rise time increases with catchment size. The following independence criterium was chosen and calculated for each catchment separately:

$$\theta < 5 + \log(A) \qquad (3.4)$$

Where A is the catchment area in square miles and θ is the number of days passed between two events for them to be considered independent. This formula was suggested by Beard (see Cunnane (1989)) and recommended by the Water Resources Council (US Interagency Advisory Committee on Water Data (USWRC), 1982), as well as used by Bezak et al. (2014). Svensson et al. (2005) also used the catchment size to determine how many days are used as an independence criterium, though here no formula was given.

3.2.2.1.2 Choice of threshold

Choosing a threshold means finding a balance between bias and variance. The threshold should be sufficiently high to ensure that the distribution of exceedances is approximated by our chosen probability value distribution. This reduces the bias, but increases the variance for the parameter estimators of the distribution, as there are fewer data to estimate them from. Lowering the threshold decreased the variance, as there is more data with which to estimate the parameters. The challenge lies in finding a threshold high enough to fulfill model assumptions and low enough to include sufficient data to get reliable parameter estimates (Scarrott and MacDonald, 2012).

One way of selecting a threshold is by using the minimum instantaneous peak flow value of the AMF series. The discharge record must be sufficiently long (>10 years) to ensure the robustness of the minimum annual peak flow estimate (Edwards et al., 2019). Another frequently used threshold is fixed quantiles of the discharge values, for example, the upper 10% rule. (DuMouchel, 1983), which is inappropriate from a theoretical viewpoint (Scarrott and MacDonald, 2012). Cunnane (1973) reported, that on average at least 1.65 events per year should be selected to achieve an advantage over the AMP approach. Tavares and Da Silva (1983) found that the POT approach had a lower variance compared to the AMF approach if at least two events per year were selected. Robson and

Reed (1999) defined their threshold value so that on average one, three, or five events per year were selected. Bačová-Mitková and Onderka (2010) selected four events per year on average for their analysis.

For this analysis, the goal was to have two events per year on average overall catchments. For each station selected above, the following quantiles of the discharge values were calculated: 0.8, 0.9, 0.95, 0.96, 0.97, 0.975, 0.98, 0.995. The average number of events per year for each station and from that the mean overall station was calculated. The quantile chosen as a threshold for all stations had an average number of events per year overall stations closest to two.

3.2.2.1.3 Fitting the generalized Pareto distribution

The selected POT are the upper tail of the underlying distribution function that generated the historical record of discharge values. Following Coles (2001), the generalized Pareto distribution (GDP) was used to approximate those asymptotically distributed values of flow. The GDP model has three continuous parameters and is expressed as:

$$F_X(x) = P_r(X \leq x) = \begin{cases} 1 - \left(1 + \xi \dfrac{(x - \mu_l)}{\sigma}\right)^{-\frac{1}{\xi}} & \xi \neq 0 \\ 1 - e^{\left(-\frac{x-\mu_l}{\sigma}\right)} & \xi = 0 \end{cases} \qquad (3.5)$$

Defined on $[x - \mu_l : \{x - \mu_l\} > 0 \; and \; \left\{1 + \xi\left(\frac{x-\mu_l}{\sigma}\right)\right\}$, where

F(x)	cumulative distribution function, non-exceedance probability of x
x	Flood peak in mm/d
X	Random variable
ξ	Shape parameter
μ_l	Location parameter
σ	Scale parameter

After selecting a distribution to fit to the POT, a method needs to be chosen for estimating the distribution parameters. Here the L-moments method was used, as introduced by Hosking (1990) and compared to other methods by Zea Bermudez and Kotz (2010a), who also stated the wide use of this method in hydrology. L-moments are linear combinations of order statistics, the elements of an ordered sample of values. Sankarasubramanian and Srinivasan (1999) found L-moments to be superior when estimating GPD parameters. According to Hosking (1990), they are less sensitive to outliers in the data, give better parameter estimates for small samples, and are less biased in their estimation.

The following functions from the package *lmomco* (Asquith, 2021) were used: *lmom.ub* was used to calculate the following L-moments (λ) and L-moments ratios (τ) from a vector of POT values. In an ordered sample of size n, values ($X_{1:n} \leq X_{2:n} \leq \ldots \leq X_{n:n}$) are drawn from the distribution of X. The first four L-moments are shown below, with the nth L-moment λ_n being a linear combination of the expected value of the order statistics $E[X_{1:n}]$ (Hosking and Wallis, 1997).

$$\lambda_1 = E[X_{1:1}]; \quad \lambda_2 = \frac{1}{2}E[X_{2:2} - X_{1:2}];$$
$$\lambda_3 = \frac{1}{3}E[X_{3:3} - 2X_{2:3} + X_{1:3}]; \quad \lambda_2 = \frac{1}{4}E[X_{4:4} - X_{3:4} + X_{2:4} - X_{1:4}] \tag{3.6}$$

$$\tau_3 = \frac{\lambda_3}{\lambda_2};\ \tau_4 = \frac{\lambda_4}{\lambda_2} \tag{3.7}$$

The first L-moment expresses the mean, the second L-moment the L-scale, measuring the dispersion of the random variable X. The skewness of the distribution of X is measured by τ_3, with $\tau_3 > 0$ indicating a skewness to the right and $\tau_3 < 0$ indicating a skewness to the left. Kurtosis of the distribution is evaluated using τ_4, with $\tau_4 > 0$ indicating broader tails (Hosking and Wallis, 1997; Ulrych et al., 2000).

pargpa was used to estimate the parameters of the Generalized Pareto Distribution, given the previously calculated L-moments. *quagpa* (inverse cdf, formula (3.8)) was used to compute the quantiles corresponding to non-exceedance probabilities.

$$x(P_U) = \begin{cases} \mu_l + \frac{\sigma}{\xi}\left(1 - (1 - P_U)^{\xi}\right) & \xi \neq 0 \\ \mu_l - \sigma \log(1 - P_U) & \xi = 0 \end{cases} \tag{3.8}$$

With

$x(P_U)$	quantile function, giving the quantile for a non-exceedance probability P_u
x	Flood peak in mm/d
ξ	Shape parameter
μ_l	Location parameter
σ	Scale parameter

Figure 3.5 shows the POT data for one exemplary station, sorted by magnitude, and the fitted GPD. The cumulative non-exceedance probability P_U of a POT observation was calculated using formula (3.9). The observations were arranged in ascending order, R_i was the rank of an observation, and n

the total number of observations. The term $(n + 1)$ takes into account, that the maximum observed value can be exceeded.

$$P_U = \frac{R_i}{(n+1)} \tag{3.9}$$

Inserting the previously estimated parameters into the quantile function of the GPD, flow values Q corresponding to non-exceedance probabilities (P_u) were estimated. The vector of P_u values was calculated using formula (3.9) with n = 1000, to obtain enough values for a good fit.

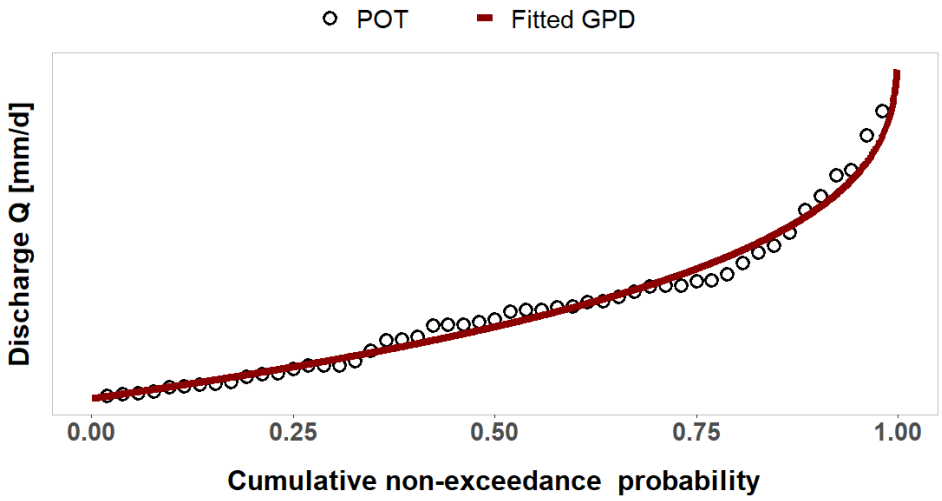


Figure 3.5: Exemplary plot of POT data (black) with fitted GPD (red)

3.2.2.1.3.1 Goodness-of-fit tests

To ensure that the GPD was a good fit to our POT data, three goodness-of-fit tests were performed: the Kolmogorov-Smirnow-test (KS test), the Cramér–von Mises test (CvM test), and the Anderson–Darling test (AD test). A one-sample test compares a sample with a reference distribution function, a two-sample test compares the distributions of two samples.

Using the Kolmogorov–Smirnov statistic (D) the distance between the empirical distribution function (edf) of two samples (two-sample) was calculated (Razali and Wah, 2011). A two-sample test was performed on the sorted POT data and calculated flow values for a vector of non-exceedance probabilities, using the quantile function with the estimated parameters of the GPD. D is defined in Formula (3.10), where in the two-sided case F_x and F_y are the edfs of the two samples. (R Core Team, 2019).

$$D = max\left|F_x(u) - F_y(u)\right| \tag{3.10}$$

The Cramér–von Mises test determines the goodness-of-fit of a cumulative distribution function (cdf) compared to a given empirical distribution function of a sample, using the Cramér–von Mises

criterion (ω^2). It is based on the square difference of edf and cdf (Anderson and Darling, 1954; Razali and Wah, 2011). The one-sample test was performed, using the sorted POT data, the cdf of the GPD, and the estimated GPD parameters. ω^2 is given in formula (3.11), with $F(x_i)$ being the cdf, n the sample size, and x_i the ordered POT data.

$$\omega^2 = \frac{1}{12n}\sum_{i=1}^{n}\left(F(x_i) - \frac{2i-1}{2n}\right)^2 \qquad (3.11)$$

The Anderson-Darling test is a modification of the CvM test, giving more weight to the tails of the distribution. The Anderson-Darling statistic (W_n^2) is also based on the square difference of edf and cdf (Anderson and Darling, 1954). A one-sample test was performed using the sorted POT data and the parameters of the GPD. The method of Braun (1980) was used to adjust for the effect of estimating the distribution parameters from the data. The computation (formula (3.12)) requires the sample size n, the specified cdf $F_0(x_i)$, and the ordered POT data x_i.

$$W_n^2 = -n - \frac{1}{n}\sum_{i=1}^{n}(2i-1)[\log(F(x_i)) + \log(1 - F(x_{n+1-i}))] \qquad (3.12)$$

In R the test was performed using *ks.test* from the *stats* package (R Core Team, 2019), *ad.tes*t from the *goftest* package (Faraway et al., 2019), and *cvm.test.lmomco* from the *lmomco* package (Asquith, 2021).

The null hypothesis for all tests was, that the sample is drawn from the reference distribution (cdf) (one-sample) or that both samples are drawn from the same distribution (two-sample). The p-value is the probability that the null hypothesis is correct. A good fit is indicated by a p-value above the significance level $\alpha = 0.05$, thereby accepting the null hypothesis.

3.2.2.1.3.2 Calculating return period of flood stages

After confirming the goodness-of-fit of the GPD to our POT data, the fitted distribution was used to estimate the return periods (T) of the flood stage triggering flows (Q_{Stage}) we calculated in 0. For this the non-exceedance probability P_u was calculated for Q_{Stage}, stage using the cdf function with the previously estimated GPD parameters. The mean time between two successive POT events (μ), was calculated for each station, using the mean number of flood events per year (formula (3.13)). The return period T of each Q_{Stage} was then calculated for each station, using formula (3.14), inserting P_u and μ. Both formulas were adapted from Brunner et al. (2016).

$$\mu = \frac{1}{number\ of\ flood\ occurences\ per\ year} \qquad (3.13)$$

$$T(Q_{Stage}) = \frac{\mu}{1 - P_U} \quad (3.14)$$

3.2.2.1.3.3 Calculating flow level of return periods

As mentioned before, the minor, moderate, and major flood stages are assigned theoretical statistical return periods. Table 3.4 shows the return period range corresponding to each flood stage. For later comparison of statistical and flood stage water levels, discharge (Q_T) for the following annual recurrence intervals (T) was calculated: 5, 10, 15, 40, 50, 100.

Table 3.4: Return periods assumed to cause flooding corresponding to the flood stages

Flood Stage	**Return period T [a]**	
	T_{lower}	T_{upper}
Minor	5	10
Moderate	15	40
Major	50	100

The non-exceedance probability P_U was computed, using formula (3.15) and the return periods (T) named above. P_U was then used in the quantile function of the GPD (formula (3.8)), along with previously calculated GPD parameters, resulting in values for Q_T.

$$P_U = 1 - \frac{\mu}{T} \quad (3.15)$$

3.2.3 Classifying stations

The previously calculated flow for each return period Q_T was in the next step compared to the flow triggering the flood stages (Q_{Stage}). For each flood stage (minor, moderate, major) the relationship between Q_T and Q_{Stage} was examined following formula (3.16), (3.17), and (3.18). Q_{T_lower} refers to the flow value of T_{lower} and Q_{T_upper} to flow value of T_{upper} for each flood stage. Exemplary for the minor stage, if Q_{Stage} was lower than $Q_{T=5}$, it would be classified as *Below*, if Q_{Stage} was higher than $Q_{T=10,}$ as *Above* and if Q_{Stage} was within the range of $Q_{T=5}$ to $Q_{T=10}$, it was classified as *In*. At the end of this step, we have a table, where each station is assigned an indicator for each flood stage, classifying if Q_{Stage} is *below*, *above*, or *within* the flow of the given recurrence interval range.

$$Below = Q_{stage} < Q_{T\_lower} \quad (3.16)$$

$$Above = Q_{stage} > Q_{T\_upper} \quad (3.17)$$

$$In = Q_{stage} > Q_{T\_lower} \ \& \ Q_{stage} < Q_{T\_upper} \quad (3.18)$$

Figure 3.6 illustrates the classification. The meaning of the categories (*above, in, below*) can be best explained using the return periods, where T_{Stage} is the return period of the flood stage triggering flow Q_{Stage}. If a station is classified *below*, it means that Q_{Stage} has a lower return period than T_{lower}. This results in the stage being triggered statistically more often, than we would expect given the return periods assigned to the flood stage. Is the station classified *above*, the stage is triggered statistically less often than expected, as T_{Stage} is higher than T_{upper}. Lastly, if the station is classified as *in*, the return period of Q_{Stage} is within the expected range between T_{lower} and T_{upper}, therefore, the statistical exceedance is as expected. The implication of the classification is also reflected by the chosen color code, as *above* is marked as green, *in* is colored blue, and *below* is red.

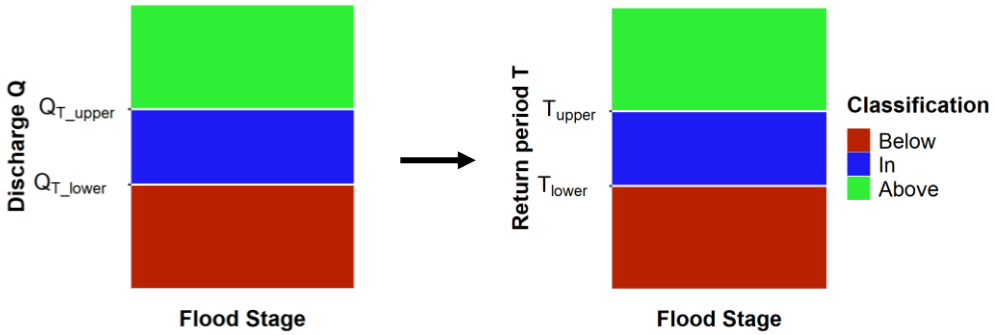


Figure 3.6: Illustration of the classification in Below/In/Above. On the left the used discharge values are depicted, Q_T being the discharge of the return periods. The right shows the meaning of the classifications for the return periods. Q_{Stage} and with that also T_{Stage} lies in one of the categories, depending on the relationship of the value of Q_{Stage} and Q_T.

To summarise: *below* means the given return period range (T_{lower} to T_{upper}) overestimates the actual return period of Q_{Stage} ($T_{Stage} < T_{lower}$), while *above* means T_{Stage} is underestimated ($T_{Stage} > T_{upper}$).

3.2.4 Selection of relevant catchment characteristics

To answer the question, whether catchment characteristics influence the classification made in 3.2.3, from the 354 available parameters, those relevant had to be chosen. The selection was based on parameters directly influencing hydrological processes, like climate and geological properties, but also anthropogenic influences. On one hand, a wide range of criteria was desired, to cover many different catchment characteristics. On the other hand, a redundancy of similar characteristics was to be avoided. As mentioned above the focus was on those characteristics directly influencing hydrological processes, in particular formation and occurrence of floods. As such the ability of a catchment to absorb precipitation before it gets to the stream and the ability to transport runoff out of the catchment quickly are important. Natural runoff formation processes were equally considered, as well as anthropogenic impacts on the catchments since the flood stages characterize damages to anthropogenic infrastructure and population.

For the previously selected 727 stations, the chosen parameters were extracted from the GAGES II database and saved in two files. One file contains the unmodified parameter data, the other contains the normalized parameter data x_{norm}, ranged 0 to 1 using formula (3.19).

$$x_{norm} = \frac{x - \min(x)}{\max(x) - \min(x)} \tag{3.19}$$

Table 3.5: Selected Catchment characteristics for later correlation and regression. Adapted from (Falcone et al., 2010; Falcone, 2017)

Characteristic	Abbreviation	Description	Unit
CLASS	CLASS	REF = reference (least-disturbed hydrologic condition); NON-REF = not reference.	[-]
DRAIN_SQKM	DRAIN	Watershed drainage area	[km^2]
RRMEDIAN	RRMEDIAN	Elevation - relief ratio, calculated as $(ELEV_{MEDIAN} - ELEV_{MIN})/(ELEV_{MAX} - ELEV_{MIN})$	[-]
RUNAVE7100	RUNAVE	Estimated watershed annual runoff, mean for the period 1971-2000	[mm/year]
STREAMS_KM_SQ_KM	STREAMS	Stream density	[km/km^2]
HIRES_LENTIC_PCT	LENTIC	Watershed surface area covered by "Lakes/Ponds" + "Reservoirs"	[%]
PPTAVG_BASIN	PPTAVG	Mean annual precipitation, calculated from 30 years period of record 1971-2000.	[cm]
PRECIP_SEAS_IND	PRECIP_SEAS	Precipitation seasonality index, based on monthly precipitation values from 30-year (1971-2000) PRISM: 0 (precipitation spread out exactly evenly in each month) to 1 (all precipitation falls in a single month).	[-]
SNOW_PCT_PRECIP	SNOW	Snow percent of total precipitation estimate, mean for period 1901-2000	[%]
DEVNLCD06	DEVLP	Watershed percent "developed" (urban), 2006	[%]
PLANTNLCD06	PLANT	Watershed percent "planted/cultivated" (agriculture), 200	[%]
FORESTNLCD06	FOREST	Watershed percent "forest", 2006	[%]
RIP800_DEV	RIP_DEV	Riparian 800m buffer "developed" (urban), 2006: area 800m each side of stream centreline, for all streams in the watershed	[%]
HLR_BAS_DOM_100M	HLR	Dominant Hydrologic Landscape Region within the watershed.	HLR region (1 - 20)

3.2.5 Relationship examination

3.2.5.1 Correlation

Correlation analysis evaluates if there exist mutual patterns between two random variables x_1 and x_2. Correlation only measures the similarity of variation, it does not imply a cause-effect relationship between the variables. Covariance s_{x1x2} is the totalized product of the variance of each variable value (x_{1i}, x_{2i}) from the variables mean ($\bar{x}_1, \bar{x}_2$). The higher the absolute value of the covariance (formula (3.20)), the stronger the correlation between x_1 and x_2. A positive covariance means both variables vary in the same direction, while a negative covariance means one variable increases the more the other variable decreases. If the covariance is 0, there is no correlation between the variables (Dormann, 2017).

$$cov(x_1 x_2) = s_{x1x2} = \frac{1}{n}\sum_{i=1}^{n}(x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2) \qquad (3.20)$$

The absolute value of the covariance is dependent on the absolute values of x_1 and x_2, so in order to be able to compare results, we need to normalize the value of the covariance. One way of normalization is to use Pearson's correlation coefficient ρ, ranging from $-1 \leq \rho \leq$:

$$cov(x_1 x_2) = \rho = \frac{\sum_{i=1}^{n}(x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2)}{\sqrt{\sum_{i=1}^{n}(x_{1i} - \bar{x}_1)^2 \sum_{i=1}^{n}(x_{2i} - \bar{x}_2)^2}} \qquad (3.21)$$

Values of ρ close to 1 (positive) or -1 (negative) show a strong correlation, values close to 0 show no correlation between the two variables. Pearson's ρ is based on the assumption that the two variables are normally distributed. For not normally distributed data, an alternative correlation coefficient is Spearman's ρ, calculated as Pearson's ρ of rank transformed data. For this, the actual values of the variables are replaced by their position in the sorted dataset, herby calculating the correlation between the ranks of data points in x1 and x2. This reduced the influence of outliers in the data on the resulting correlation (Dormann, 2017).

Depending on the variable type, different correlation coefficients needed to be calculated. A distinction is made between continuous and discrete variables, discrete variables being separated in dichotomous (two possible values) and polytomous (more than two possible values) (Dormann, 2017).

- Biserial correlation: between a continuous and a dichotomous variable
- Polyserial correlation: between a continuous and polytomous variables
- Polychoric correlation: between two polytomous variables
- Tetrachoric correlation: between two dichotomous variables

mixedCor from the package *psych* was used on the not-normalized parameter data, to determine the correlation in R (Revelle, 2020). This function was able to compute correlations between different types of variables, by manually specifying continuous, polytomous, and dichotomous parameters (as seen in Table 3.6). All parameters needed to be in a numeric format, for this, non-numeric variables were first transformed to characters and then to numeric values. The HLR parameter had too many categories and was, therefore, considered a continuous variable. For continuous variables, the spearman's ρ was calculated, the correlation for the other variable combinations was calculated as listed above. By rank-transforming, the continuous variables when calculating correlation with discrete variables, the spearman-version of the correlation coefficients was calculated (Dormann, 2017).

Table 3.6: variable type of the variables selected in 3.2.4, used for examination of the correlation

Variable type	**Variable name**
Continuous	DRAIN, RRMEDIAN, RUNAVE, STREAMS, LENTIC, PPTAVG, PRECIP_SEAS, SNOW, DEVNLP, PLANT, FOREST, RIP800_DEV
Dichotomous	CLASS
Polytomous	indicator

To determine the significance of the correlation found, a p-value was calculated. The null hypothesis being, that there was no linear correlation between x_1 and x_2 ($\rho = 0$). The correlation values (ρ) found were significant if $p < \alpha = 0.05$, thereby rejecting the null hypothesis. In R the p-values of the continuous and dichotomous variables were calculated using a modified version of *cor.ci* from the *psych* package (Revelle, 2020) and verified using *cor.test* from the *stats* package (R Core Team, 2019). The p-values of the polychoric and polyserial correlation were calculated using *cor_to_p* from the *correlation* package (Makowski et al., 2020).

There was no uniform classification of the correlation coefficients, outside of 1 being a perfect correlation and 0 showing no correlation. Therefore, the coefficient was evaluated following Table 3.7, which was adapted from Akoglu (2018) and Yan et al. (2019).

Table 3.7: Interpretation of the spearman correlation coefficient (ρ): grading table.

Coefficient ρ	Strength of correlation
ρ = 0	None
0 < |ρ| < 0.2	Very weak
0.2 ≤ |ρ| < 0.4	Weak
0.4 ≤ |ρ| < 0.6	Moderate
0.6 ≤ |ρ| < .0.8	Strong
0.8 ≤ |ρ| ≤ 1	Very strong
1	Monotonic/ perfect

3.2.5.2 Regression

Regression analysis examines and models the relation between variables. Regression implies a directional dependence between one dependent variable (y = response) and one or multiple independent variables (x = predictors). This means that y is a function of x, denoted as y~x or y = f(x), so that x influences y, but not the other way around (Dormann, 2017). Formula (3.22) expresses a simple linear regression model, while formula (3.23) shows a multiple regression model. x_1, …, x_n being the predictors, β_0 the intercept, and β_1, …, β_n the regression coefficients. In a linear model, the regression coefficients express the change in y for a unit change in x_i, when the other x_n are kept constant (Naghettini, 2017).

$$y = \beta_o + \beta_1 \cdot x_1 \tag{3.22}$$

$$y = \beta_o + \beta_1 \cdot x_1 + \cdots + \beta_n \cdot x_n \tag{3.23}$$

Before fitting a model, we have to decide, which distribution function f(x) the response is drawn from. The link function g(y) is used to ensure that the response values predicted with the calculated linear model still conform to the selected distribution, meaning the values fall within the possible range of the distribution.

For later reference, the binomial regression model or logistic regression is examined more closely. Here the response (y) is binary data, meaning it can be in one of two categories (0, 1). The link function of the binomial distribution is the logit (formula (3.24)), used to keep the predictions between 0 and 1 (Dormann, 2017). The model follows formula (3.25) and estimates the probability P(y = 1|x), of y to be in one of the two categories, depending on the values of x. To get back to the simple structure of a linear model with a possible value range from -∞ to +∞, formula (3.25) is transposed to formula (3.26). The term on the left-hand side is called log-odds or logit. It is important to remember, that here a unit change of x changes the log-odds by β_1 and not $P(y = 1|x)$ (James et al., 2021).

$$y' = \ln\left(\frac{y}{1-y}\right) \tag{3.24}$$

$$F(x) = \mathrm{P}(y = 1|x) = \frac{e^{\beta_o + \beta_1 \cdot x_1}}{1 + e^{\beta_o + \beta_1 \cdot x_1}} \tag{3.25}$$

$$\log\left(\frac{\mathrm{P}(y = 1|x)}{1 - \mathrm{P}(y = 1|x)}\right) = \beta_o + \beta_1 \cdot x_1 \tag{3.26}$$

For estimating the regression coefficients, the maximum likelihood method is preferred. The likelihood measures, how plausible the distribution parameters are, given the data. In the case of the regression, we seek β_n estimates so that the predicted probability for the categories is closest to the observed probability (James et al., 2021). The likelihood is defined in formula (3.27) and quantifies the total probability density of the dataset. X is a vector of observations, n is the number of observations in X and θ the parameters of the distribution, so that P(X|θ) is the product of the likelihood of the single data points in X. Since the likelihood values of the singular observations are very small, the log-likelihood is calculated using formula (3.28). A high log-likelihood means, that the probability $P(X|\theta)$ to obtain the observations in X with given parameters θ is high. The parameter combination with the maximized log-likelihood is chosen for the regression model (Dormann, 2017).

$$L = P(X|\theta) = \prod_{i=1}^{n} P(X_i|\theta) \tag{3.27}$$

$$l = \log(L) = \ln\left(\prod_{i=1}^{n} P(X_i|\theta)\right) = \sum_{i=1}^{n} \ln\big(P(X_i|\theta)\big) \tag{3.28}$$

Knowing the correlation of the parameters calculated before is important for the regression. Correlation between two predictors is called collinearity and can cause problems for finding the optimal regression model. If two predictors are very similar, deciding on which one is important can be difficult. Additionally, when trying to find the optimal regression model similar predictors cause difficulties in finding the optimum, as there is an infinite number of parameter variations of those correlated predictors. Hence, the goal was to have little correlation between the parameters used as predictors in the regression model.

In the case of this analysis, the catchment parameters were the predictors (x), while the indicator categorized in 3.2.3 was the response (y), therefore a multinomial logistic regression model needed to be fit. With the response variable being a categorical variable of 3 ordered categories (k = *below*, *in*, *above*), an ordered categorical regression (ordinal) logistic regression model was used.

Ordinal logistic regression is very similar to the previously explained logistic regression, with the difference that the response has more than two categories that have a set order. Here that order is *below* < *in* < *above*. The analysis was done using *polr* from the *MASS* package (Venables and Ripley, 2002), by means of a cumulative link model. Cumulative models (CM) are based on the assumption of an underlying latent variable y* with a continuous distribution function F(.), of which the categorial version is observed (formula (3.29)). The proportional odds model is the most widely used CM, using the logistic distribution function, as seen in formula (3.30) (Agresti, 2007; Tutz, 2021). In R, *polr* follows formula (3.31), where r is the categories, ζ_k are the intercepts for the class boundaries and η coefficients of the linear predictors (Venables and Ripley, 2002).

$$P(y \leq r|x) = F(\beta_o + \beta_1 \cdot x_1), \quad r = 1, \dots, k \tag{3.29}$$

$$\log\left(\frac{P(y \leq r)}{\mathrm{P}\,(y > r)}\right) = logit(\mathrm{P}(y \leq r|x)) = \beta_o + \beta_1 \cdot x_1 \tag{3.30}$$

$$logit(\mathrm{P}(y \leq r|x)) = \zeta_k - \eta_1 \cdot x_1 - \dots - \eta_n \cdot x_n \tag{3.31}$$

For a proportional odds model to be applicable, the proportional odds assumption needed to be validated. To uphold this assumption, the slope estimates β or η have to be the same across all outcomes, meaning that a variable has an identical effect at each cumulative split of the response. In R we checked the proportional odds assumption using the brant test (Brant, 1990) from the *brant* package (Schlegel and Steenbergen, 2020). With a p-value > 0.05 for every predictor, the proportional odds assumption would be satisfied.

To obtain the outcome probabilities of the different response values for a parameter combination, the odds have to be calculated (formula (3.32)) and from that the probability for each possible outcome (formulas (3.33)). Since the probability is described as P(y≤r|x), the calculated probability will always be that y is in category r or less. Given the cumulative model, the probability P_r of each category can be calculated according to formula (3.34).

$$odds = e^{logit(\mathrm{P}(y \leq r|x))} \tag{3.32}$$

$$\mathrm{P}(y \leq r|x) = \frac{odds}{1 + odds} \tag{3.33}$$

$$\begin{aligned} P_{below} &= \mathrm{P}(y \leq below|x) \\ P_{in} &= \mathrm{P}(y \leq in|x) - P_{below} \\ P_{above}(x) &= 1 - P_{in} \end{aligned} \tag{3.34}$$

In R the *predict* function computed the probabilities of each response category (*below*, *in*, *above*) and from that the predicted class for each parameter combination. In order to be able to compare the η values of the predictors, the normalized parameter data, created in 3.2.4, was used. Standardizing allows using the coefficients to evaluate the importance of each predictor to the outcome of the model, without having to keep the original units in mind.

Looking at formula (3.31), the obtained parameters of the *polr* model will have an inverse effect on the outcome probability. In logistic regression (formula (3.26)) positive coefficients (β) will have an increasing effect on the logit of the probability. In the ordinal logistic regression applied here, positive coefficients will have a decreasing effect on the logit, caused by the minus-sign in front of the coefficients η. For interpretation purposes $-\eta_i = \beta_i$.

Using *polr*, the response was coded as an ordered factor, the continuous predictors were numeric, the discrete parameters were transformed from character strings to factors. For categorical predictors with more than 2 categories, *polr* creates dummy variables, turning each category into a new predictor, coded 0 or 1. So instead of 14 predictors, there were 31, with each of the 19 HLR classes being their own predictor. However, to avoid multicollinearity, one of the dummy variables has to be dropped, which resulted in 30 predictors. *Polr* does this automatically. For the latter interpretation of the coefficients of the dummy variables, their structure was important. If dummy variable A was 1, then all the other dummy variables were 0.

3.2.5.3 Stepwise model selection

In 3.2.4 14 catchment parameters were selected under the assumption, that they influence the classification of the indicator. The goal of the last step of the analysis was to determine which of those catchment parameters can be used to predict the category of the indicator for any catchment. Only the most suitable predicts were to be included in the regression model, keeping the number of predictors as low as possible. Prior a regression model was built with all 14 selected parameters, which was then used in the stepwise model selection.

In each step of a stepwise model selection, one of the predictors is iteratively be added or subtracted from the model, depending on the selected measure of performance (MOP). There are three main approaches: forward selection, backward selection, and bidirectional selection. In forward selection, the initial model has no predictors (null model), they are added one at a time until all predictors are included. At each step, the predictor with the greatest improvement of the MOP is added to the model until the MOP no longer improves. Once a predictor is added to the model it cannot be removed. The backwards selection starts with the full model, including all predictors and one at a time removes one predictor, that least improves the MOP. The Bidirectional selection is a combination of backwards and forward selection and can start with the full or null model. At each step, predictors can be added

or removed to the model depending on the MOP. Previously selected/removed predictors can be removed/selected in later steps (James et al., 2021).

The measures of performance used, were the Akaike's Information Criterion (AIC) and the Bayesian Information Criterion (BIC), where p is the number of fitted parameters, n the number of observations, and l the log-likelihood (Dormann, 2017). The lower the value of AIC and BIC, the better the fit.

$$AIC = -2l + 2p \tag{3.35}$$

$$BIC = -2l + p \cdot log(n) \tag{3.36}$$

Two MOPs were chosen because they each select the best model differently. With the AIC the left term (-2l) decreases, as more parameters are added, while the right term (2p) increases. Despite this trade-off between under and overfitting, the AIC is more likely to overfit, creating models with many parameters. BIC on the other hand punishes additional parameters more (log(727) >2), leading to models with fewer parameters that might be underfitted (Burnham and Anderson, 2002).

In R *stepAIC* from the *MASS* package (Venables and Ripley, 2002) was used to perform the stepwise selection. The bidirectional selection approach was chosen as it can both add and remove parameters (*direction = "both"*). The default MOP was the AIC (k=2), modifying *k = log(nrow(data))* computed the BIC instead.

First, a full model with all parameters was fit using the *polr* function. This model was input into *stepAIC*, selecting the model parameters based on the AIC or BIC, depending on the specifications of k. To assess the goodness-of-fit of the model, a k-fold cross-validation was applied. For this, the 727 observations were randomly divided into k groups (folds) of approximately equal size. The first fold was used to validate the model, that had been fit on the remaining $k - 1$ folds. The root-mean-square error (RMSE, formula (3.37)) was calculated on the observations and model predictions from the first fold, with n_{fold} being the number of observations. This is repeated k times, using a different fold for validation each time. Here k = 10 is used.

$$RMSE = \sqrt{\frac{1}{n_{fold}} \sum_{i=1}^{n} (observed - predicted)^2} \tag{3.37}$$

The k-fold cross-validation was repeated ten times, resulting in 100 computed best models for each MOP. The chosen parameters of those 100 models were extracted, together with the calculated RMSE. The final selection of the model was done by visually assessing the obtained AIC and BIC values and model parameters, keeping the calculated RMSE in mind. This process was done for each

flood stage, resulting in six final models in total, 2 per flood stage (one AIC and one BIC-based best model).

3.2.5.4 Model evaluation

To evaluate the six models created, they were used to predict the response (indicator), in another k-fold cross-validation procedure, similar to 3.2.5.3. For this, the models were fitted using the previously selected parameters of the best AIC and BIC model for each flood stage. Unlike in 3.2.5.3, data from all 727 stations was used to determine the model coefficients, only for determining the prediction accuracy, the data was divided into k = 10 subsets.

The prediction results for each of the k = 10 folds were compared to the classification of the indicator done in 3.2.3. Results of the comparison were given in percent of stations correctly estimated, underestimated, and overestimated. Additionally, it was examined how often which category was underestimated, overestimated, and correctly guessed, using the *count* function (Wickham, 2011).

For each of the six models mean results over all ten folds were calculated and from that, the best model for each flood stage was selected. A balance between model accuracy and number of parameters was crucial to the selection. In case the prediction accuracy between models of different MOPs was small, the model with less parameters was preferred. Smaller models are both easier to understand and have less estimation uncertainty in their reduced number of parameters.

Of the three selected models, the coefficients of the predictors were compared in their size, both within each model and between the three final models fitted. This allowed an assessment of the importance of the different predictors for modeling the indicator. The p-values of the parameter coefficients were calculated using *coeftest* from the package *lmtest* (Zeileis and Hothorn, 2002). The null hypothesis was $\beta = 0$, meaning there was no relationship between the response and the predictor.

4 Results

4.1 Data preparation

In 3.2.1, a selection of gauges was made for further analysis, and the flood stage triggering flow of each flood stage was calculated for every gauge. The results were evaluated using the flow time series, in regards to the exceedance of the different flood stages, this is shown in Figure 4.1. The hydrological year (October 1st to September 31st) was used, with the following seasonal division: spring (March, April, May), summer (June, July, August), fall (September, October, November), winter (December, January, February). This summary was calculated from all 727 stations, it does not reflect the varying seasonality depending on climate and catchment characteristics. It is merely the sum of the exceedances of all events per season over all catchments, showing the median seasonality over CONUS.

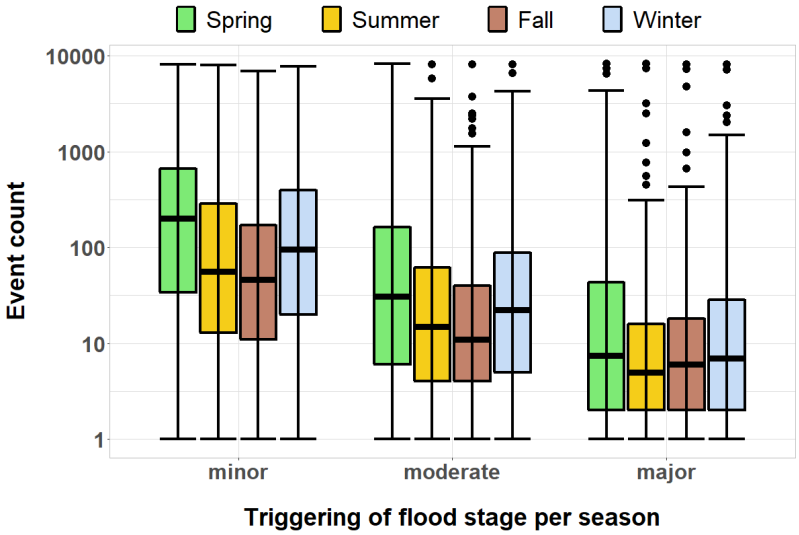


Figure 4.1: Seasonality of the triggering of the minor, moderate, and major flood stage viewed over the selected 727 gauges

Looking at the seasonality of flood stage triggering it was shown, that over all three stages, most flood stage triggering flows (events) occur in spring, followed by winter. While the difference between median events in spring and winter is relatively high (107 events) for the minor stage, it decreases with increasing flood stage to 8.5 for the moderate and 0.5 for the major stage. Summer and fall have a similar mean number of flood stage activations, differing only in ten (minor), four (moderate), and one (major) events. Figure 4.1 also shows a decrease in triggering overall with increasing flood stage. Table 4.1 shows the number of stations, where a flood stage is never triggered, during the given time series.

Table 4.1: Number of stations where the discharge of respective flood stage is not reached or exceeded

Minor stage	Moderate stage	Major stage
100	233	426

If the flood stage flow of a stage is not reached or exceeded, neither is the flow of any higher flood stage. For more than half of the 727 gauges analyzed, the major flood stage flow was never reached or exceeded. The moderate stage and above was not reached for 233 gauges, but for 133 stations out of the 233, the minor stage flow was reached. For 100 gauges neither flood stage flow was reached.

4.2 Flood frequency analysis

4.2.1 Independence criteria

Table 4.2 shows the mean, minimum and maximum days passed between peaks, for them to be considered independent. The larger the catchment area, the higher the time interval θ between flood peaks included in the POT time series.

Table 4.2: Summary of days between flood peaks for them to be assumed independent events

Mean	Min	Max
11.8	6.4	14.9

4.2.2 Choice of threshold

The table below shows the results of the discharge quantile calculation, that the choice of threshold was based on. With the goal of having around two events per year in mind, the 0.98 and 0.995 quantile were excluded, due to a too low number of events. The 0.975 quantile came closest to our desired mean of 2´two years, while the rest was higher. Despite not resulting in exactly two POT events per year, 0.975 was chosen as the threshold for the extreme value analysis.

Table 4.3: Summary of different quantiles examined to choose threshold from: mean, median, minimum, and maximum events per year over all stations

	0.8	0.9	0.95	0.96	0.97	0.975	0.98	0.995
Mean	4.40	3.95	3.12	2.82	2.44	2.21	1.94	0.75
Median	4.47	3.96	3.10	2.81	2.41	2.18	1.92	0.74
Min	0.81	0.61	0.44	0.34	0.23	0.20	0.19	0.07
Max	12.92	11.88	9.11	7.92	6.43	5.72	4.88	1.55

4.2.3 Goodness of fit

The results of the goodness-of-fit tests are shown in Figure 4.2. Results of the KS-tests, confirm matching sample distributions (two-sided test) for 711 stations. The CvM-test matches the previous findings, rating 720 stations as matching the reference distribution (GPD). The results of the AD-test, however, show, that giving more weight to the tails results in only 487 stations with a p-value above 0.05. Looking at the stations with p-values ≤ 0.05, there were no matching stations between the AD-test and the KS-test or CvM-test. However, for 7 stations both the KS-test and the CvM-test

rejected the null hypothesis. For the AD-test 220 stations had a p-value below α and an infinite value for W_n^2.

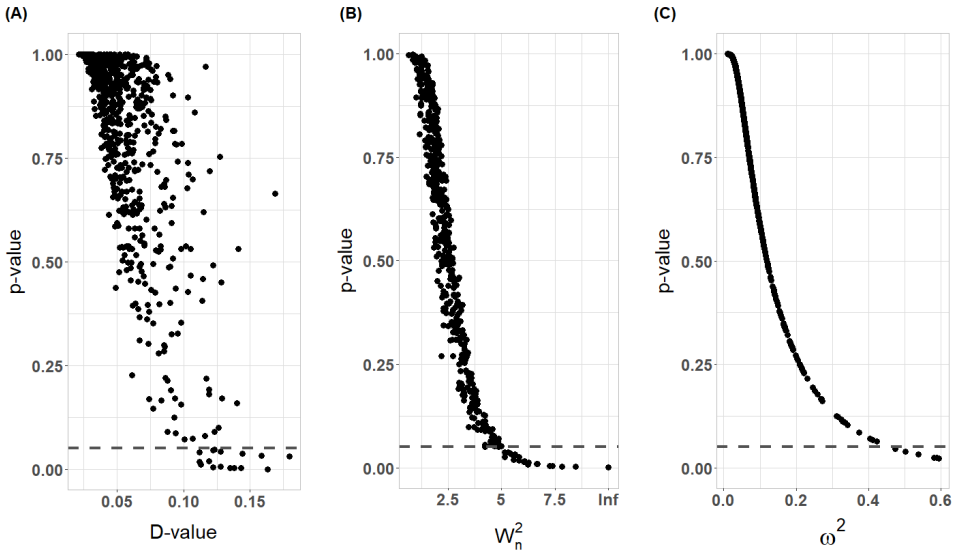


Figure 4.2: Results of the goodness-of-fit tests: KS-test (A), AD-test (B), CvM-test (C). the significance level α = 0.05 is marked by the dotted line in each plot

While there were stations, where the null hypothesis was rejected, for the majority of stations the GPD distribution was considered a good fit for the POT data. This confirmed the selection of the GPD to estimate the return periods of flood stage triggering flows.

4.2.4 Return periods of flood stage triggering flows

The figure below shows a summary of the calculated return periods of flood stage triggering discharge for all flood stages. A violin plot, representing the density of the values, was combined with a boxplot, to depict the quartiles. While the violin plot shows a clear high density at a return period of just below one year for the minor stage, it is much more drawn out for the moderate stage and major stage. Looking at the boxplots, an increase in the median return period can be seen with increasing flood stage. Likewise, the interquartile range increases with increasing flood stage.

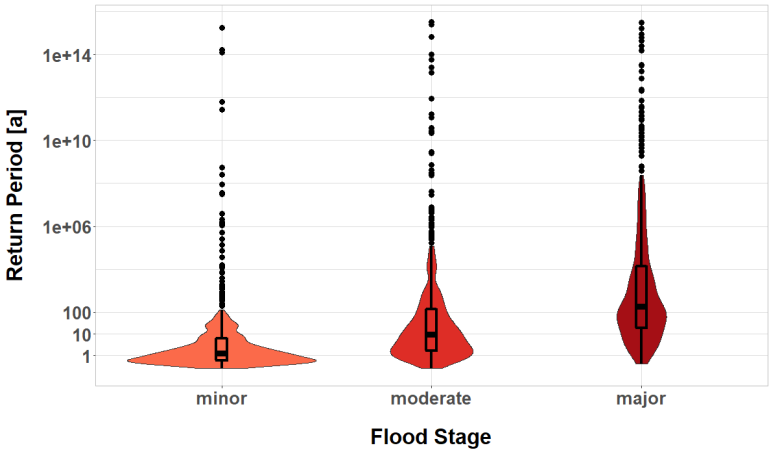


Figure 4.3: Summary of the recurrence intervals for all 4 flood stages: Plots showing the density of the calculated return periods, excluding infinite values

While the median of return periods ranges from 1.21 years (minor stage) up to 183.85 years (major stage), the large outliers lead to mean return periods of more than 10^{12} years (Table 4.4). Both the minimum return periods and the maximum return periods of all flood stages are within the same respective order of magnitude.

Table 4.4: Summary of the recurrence intervals for all 4 flood stages, excluding infinite values

	Minor	**Moderate**	**Major**
Mean	$2.9 \cdot 10^{12}$	$9.9 \cdot 10^{12}$	$1.2 \cdot 10^{13}$
Median	1.21	9.25	184
Min	0.24	0.25	0.4
Max	$1.7 \cdot 10^{15}$	$3.2 \cdot 10^{15}$	$3 \cdot 10^{15}$

For some stations, calculating the return periods resulted in infinite values (Inf), the location of those gauges is shown in Figure 4.4. For 35 stations, the calculated return period of all flood stages was Inf. For 30 stations, the return periods of the moderate and major flood stage were Inf and for 48 stations the calculated return period for just the major stage was Inf.

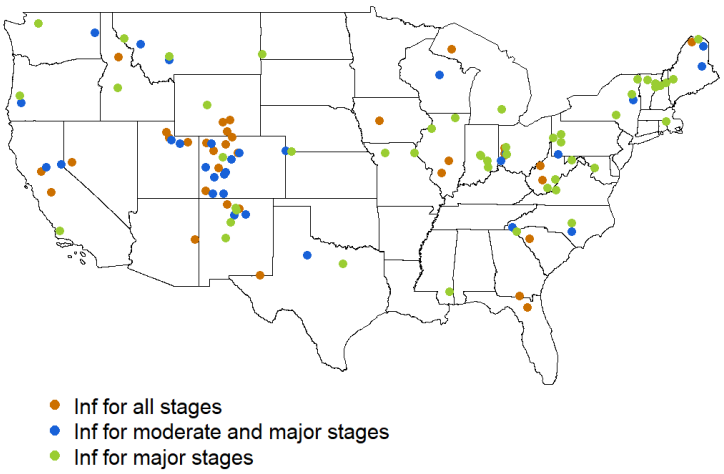


Figure 4.4: Gages with calculated infinite (Inf) return periods of flood stages. Differentiated between stations where the return periods of all flood stages were infinite, those where they were infinite for the moderate and major stage, and infinite return periods for the major stage.

4.2.5 Flow level of given return periods

Figure 4.5 shows the summary of the flow level corresponding to our selected return periods. Since the flow was normalized to mm/d using the catchment size, the obtained values can be compared. For a return period T of 5, 10, and 15 years, two peaks can be seen in the density of the flow values, the first one below the interquartile range, the second one within. While the peaks are well-formed for T = 5, they get less prominent with increasing T. For T = 40, 50, and 100, one peak can be seen at around 8 mm/d, after which the density decreases. For T = 40 a second small peak can be seen at around 68 mm/d and for and T = 50 at 75 mm/d. Over all the higher the return period, the higher the interquartile range (IQR) and the higher the range of values outside the IQR.

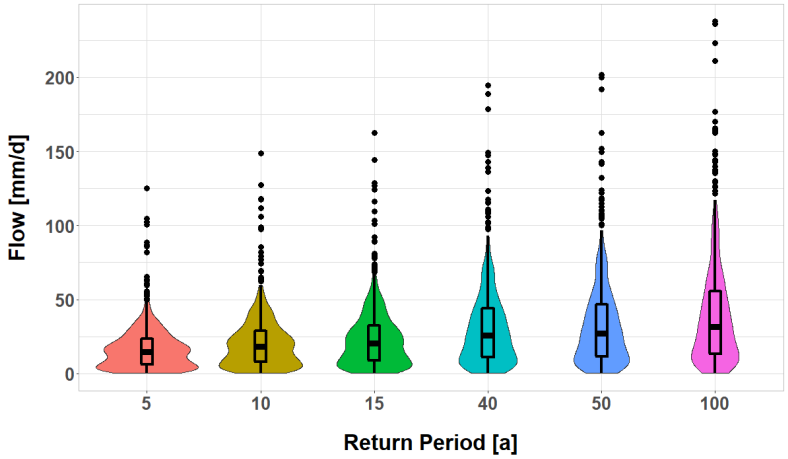


Figure 4.5: Summary of the calculated discharge of return periods 5, 10, 15, 40, 50, and 100 years. The plot shows the density and quartiles of the calculated flow

Table 4.5 contains the summary statistics of the calculated flow values for each return period. It shows a great overall difference between the minimum and maximum flow computed for each return period. This min-max difference increases with an increasing return period, from 124.94 mm/d for T = 5 to 237.03 mm/d for T = 100. The median flow of T = 100 is about twice as high as the flow of T = 5. The difference in the median between T = 40 and T= 50, is relatively small with only 1.27 mm/d. Unlike the results of the return periods of the flood stages shown in 4.2.4, here mean and median are quite similar, differing only in a range of 2.33 mm/d (T = 5) to 8.82 mm/d (T = 100).

Table 4.5: Summary of flow level [mm/d] for the recurrence intervals used in the later comparison (return periods of 5, 10, 15, 40, 50, and 100 years)

	5	**10**	**15**	**40**	**50**	**100**
Mean	17.0	21.2	23.9	31.5	33.5	40.3
Median	14.7	18.1	20.3	25.6	26.9	31.4
Min	0.22	0.28	0.32	0.45	0.48	0.60
Max	125.2	148.8	162.4	194.3	201.5	237.6

4.3 Classifying stations

The previously calculated flow corresponding to flood stages and selected return periods were then compared as described in 3.2.3. Figure 4.6 shows the results in percent, Table 4.6 in number of stations. For all flood stages, the number of stations classified as *in* was the lowest, compared to the other classifications. While for the minor stage there were more than three times as many stations *below* as *above*, for the moderate stage it is only 1.3 times as many. For the major stage, the *below*/*above* ratio is the opposite, with there being around twice as many stations *above* as *below*.

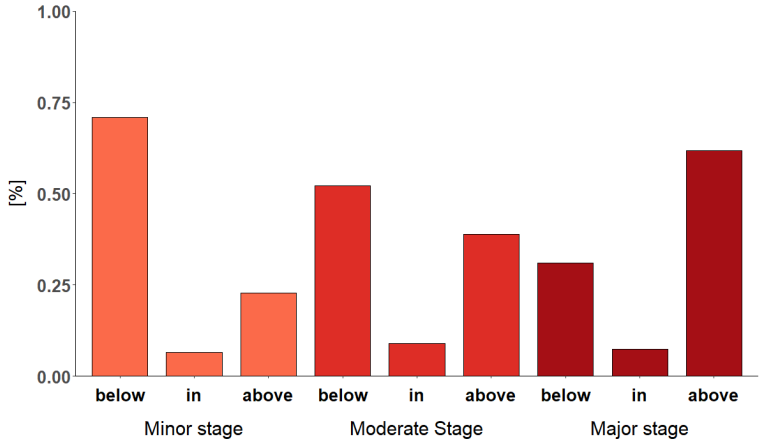


Figure 4.6: Classification of stations: placement in categories below, in, above based on the relationship between flood stage flows and return period flows

Table 4.6: Count of stations per indicator for every flood stage

Indicator	Minor	Moderate	Major
below	515	379	225
in	47	65	53
above	165	283	449

The most common indicator combinations over all flood stages were *below/below/below* with 220 stations and *above/above/above* with 163 stations. Only two stations were classified as *in* for all flood stages. The counts of all combinations can be found in Table A-2, in the appendix. Table A-3 and Table A-4 in the appendix show the number and percent of stations of each combination sequence, starting at the minor and major stage. The tables show, that of the 515 stations classified as *below* for the minor stage, 73% (377 stations) were also classified as *below* for the moderate stage and of that another 58% (220 stations) for the major stage. When the minor stage was classified as *above* (165 stations), 99% (163 stations) of those stations would also be classified *above* for the moderate stage, and of those 100% would be *above* for the major stage as well. Of the 449 stations classified as *above* for the major stage, 62% (282) were also *above* for the moderate stage.

Figure 4.7 shows the spatial distribution of the classifications for all stages, Figure 4.8 shows the location of stations that had the same classification over all flood stages. For the minor stage (Figure 4.7 (A)) a majority of stations in the Appalachian Highlands, along the Atlantic Plain, and in the Central Lowlands are classified as *below*. Additionally at the north-western end of CONUS, in the states of Washington, Oregon, Idaho, and Montana, stations are mainly classified as *below* or *in*. Stations located where the southern end of the Rocky Mountain System transitions into the Great Plains to the east and the Intermontane Plateaus to the west are mostly classified as *above* (Colorado, Utah, New Mexico, Wyoming).

Looking at the moderate stage in Figure 4.7 (B), the number of stations *above* increased. We still see the same patterns of *below* classifications in the East and a diagonal zone from Illinois to Texas,

though now less prominent. In the southwest, most stations are, like for the minor stage, also classified *above* for the moderate stage. Figure 4.7 (C) depicts the classification of the major stage and a clear increase in stations classified *above*. Areas, where stations were predominantly *below* before, are now marked as *above*. The diagonal zone from Illinois to Texas is barely visible, only in the north, where Illinois, Indiana, and Wisconsin meet, are a majority of stations marked *below* or *in*. Examining the West, only where the borders of New York, Pennsylvania, and New Jersey meet, is an area primarily classified as *below* or *in* visible.

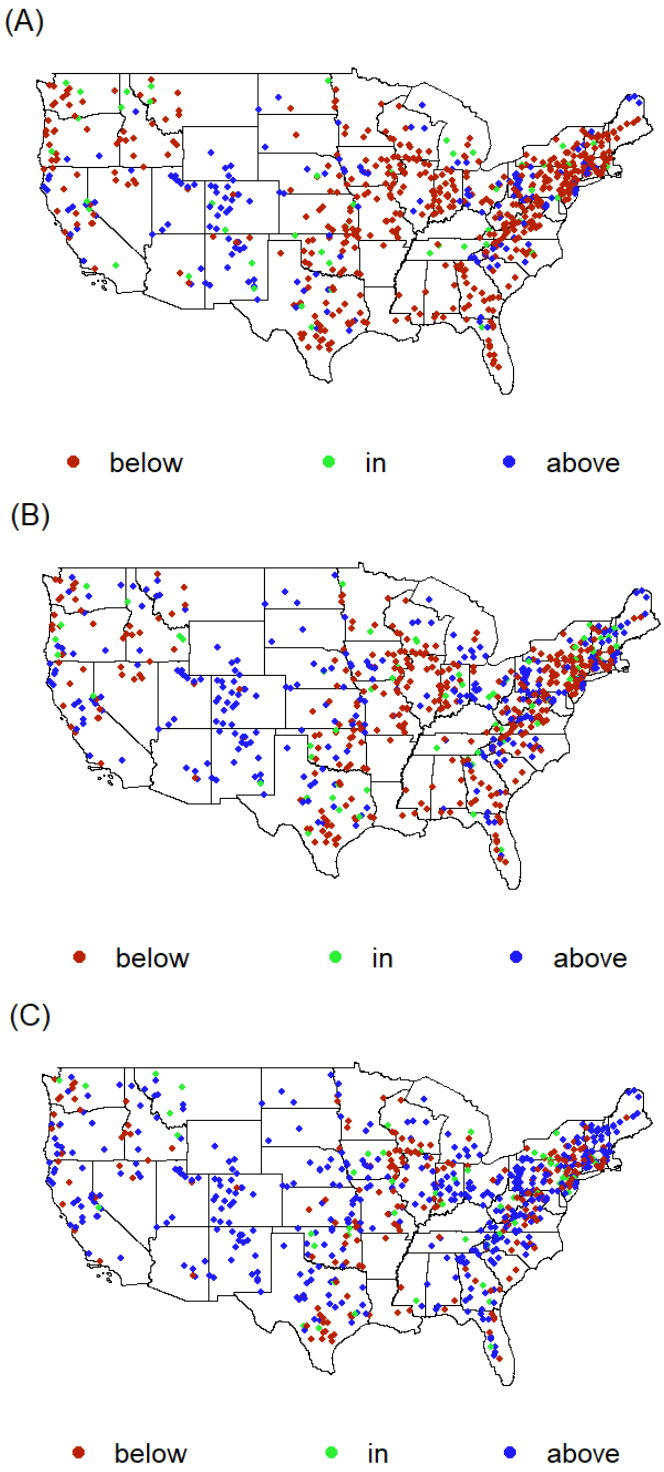


Figure 4.7: Maps of the classification of gauges made in 3.4: (A) minor stage, (B) moderate stage (C) major stage

Examining Figure 4.8, we see that the patterns described above for the minor stage (Figure 4.7 (A)) can also be seen here. There is an area of stations predominantly classified as *above* at the southern half of the Rocky Mountain System and transitioning to the Great Plain and the southern Intermontane Plateaus. The diagonal zone from Illinois to Texas and the areas in the north and on the eastern side of the Appalachian Highlands are mostly classified *below* over all stations.

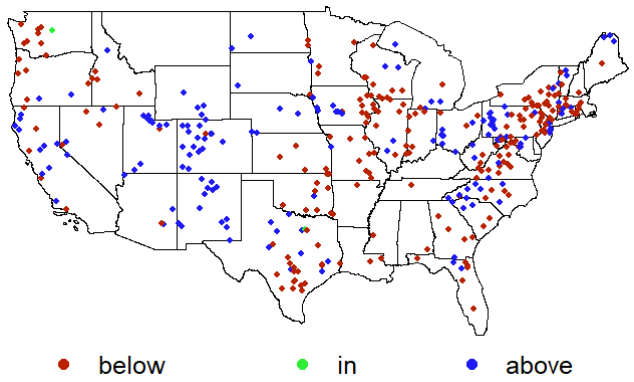


Figure 4.8: Stations where the relationship between flood stages and statistical thresholds was the same over all flood stages

4.4 Relationship examination

4.4.1 Correlation

Figure 4.9 and Figure 4.10 show the correlation heatmaps calculated using the selected catchment criteria and the classification of the indicator for all stages. Exact values of the correlation coefficients are only given for significant correlations.

As shown in Figure 4.9, the coefficients of 18 correlations were rated as moderate or above. Very strong correlations were found between DEVLP – RIP_DEP (0.99) and RUNAVE – PPTAVG (0.82). With $\rho = 0.99$, DEVLP – RIP_DEP had the overall highest correlation, almost reaching a perfect Spearman correlation. A strong correlation was found between FOREST – RUNAVE (0.68) and PLANT – HLR (-0.64). In addition to the strong and very strong correlations mentioned above, of the 14 moderate correlations found, the following were discussed later: DEVLP – PLANT, HLR – FOREST, FOREST – PLANT, HLR – SNOW.

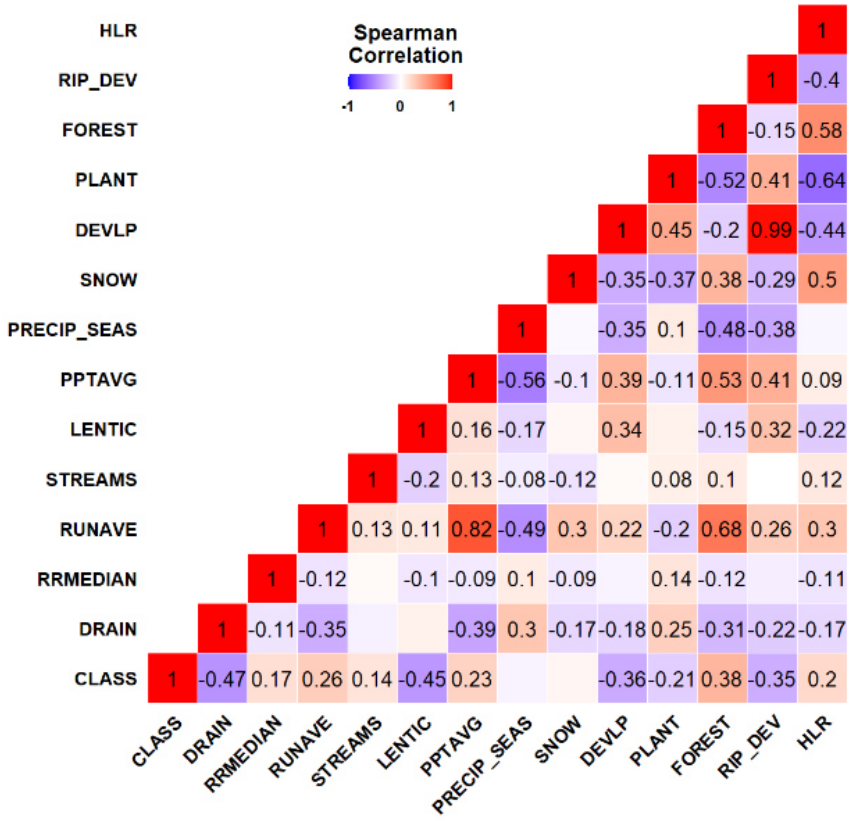


Figure 4.9: Heatmap of the calculated Spearman correlations between the selected catchment characteristics. The color indicates the size of correlation, red being above 0, blue being below 0. For significant correlations ($p < \alpha = 0.05$) the values of the correlation coefficients were printed in the corresponding box.

Looking at the correlation in Figure 4.10, a very strong correlation was found between the indicator values of all of the flood stages. The correlation coefficients between the minor and moderate stage and the moderate and major stage were almost equal, while it was smaller for the minor and major stage. The correlation between each of the flood stages and the catchment characteristics was found to be weak at most, with values of $|\rho|$ not exceeding 0.3.

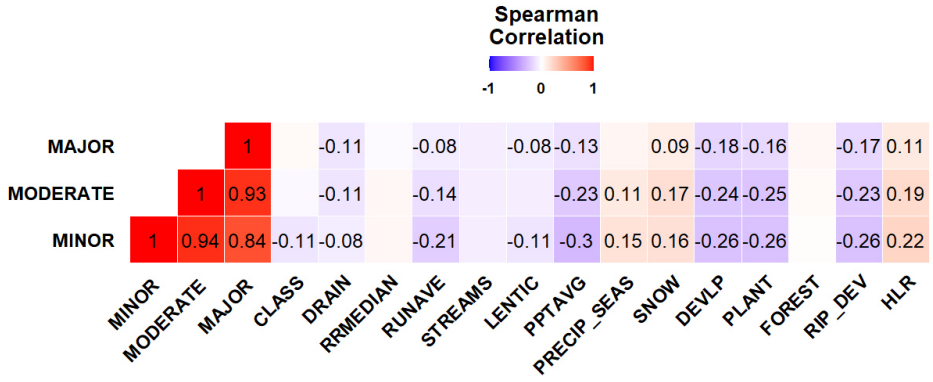


Figure 4.10: Heatmap of the calculated Spearman correlations between the selected catchment characteristics and the assigned indicators for the minor, moderate, and major flood stage. The color indicates the size of correlation, red being above 0, blue being below 0. For significant correlations ($p < \alpha = 0.05$) the values of the correlation coefficients were printed in the corresponding box.

Despite the strong correlation found between RIP_DEV and DEVLP, both parameters were included in the regression. This was done to include both the anthropogenic influences on the entire watershed (DEVLP), as well as the urbanization and anthropogenic land use around the streams (RIP_DEV). As the flood stages are a measure of damage to the area around a stream, the anthropogenic influences will, therefore, determine which flood stage is reached.

4.4.2 Stepwise model selection

The stepwise model selection resulted in 100 models, for each measure of performance (AIC, BIC), resulting in 200 models per flood stage. As said above the selection of the best model was done visually, using a heatmap of the measure of performance on the y-axis and the model parameters on the x-axis. The models with the ten lowest RMSE were kept in mind during the parameter selection. The selected final model parameters are listed in Table 4.7. It shows, that all models contain the parameter PPTAVG and all AIC models contain DRAIN and HLR. The following explains the reasoning behind their selection.

Table 4.7: Table of parameters of chosen six models for each flood stage and criterion

	MOP	**Parameters**				
Minor stage	AIC	PPTAVG	DRAIN	STREAMS	HLR	
	BIC	PPTAVG	SNOW	PLANT		
Moderate stage	AIC	PPTAVG	DRAIN	FOREST	RIP_DEV	HLR
	BIC	PPTAVG	DRAIN	PLANT		
Major stage	AIC	PPTAVG	DRAIN	FOREST	HLR	
	BIC	PPTAVG	FOREST			

4.4.2.1 Minor stage

Figure 4.11 shows, that for the models fitted using the AIC, the parameters PPTAVG and HLR were included in all best models. For the other parameters, there was great variability in the best models, as well as in the best RMSE models, which made a selection of parameters difficult. DRAIN and STREAMS were chosen since they were in the model with the lowest AIC and had also been included in several models with low RMSE values.

The selection of parameters for the BIC model was easier, as the best BIC parameter combination also had the lowest RMSE. PPTAVG, SNOW, and PLANT were, therefore, chosen as the final BIC model for the minor stage.

4.4.2.2 Moderate stage

For the moderate stage AIC model, DRAIN, PPTAVG, FOREST, and HLR were included in all best models. RIP_DEV was included in the model with the lowest AIC and lowest RMSE. DEVLP was

not selected, as the parameter was only included in one of the ten lowest RMSE models. The selected parameter combination had four out of the ten lowest RMSE, including the overall lowest.

For the BIC model of the moderate stage, PPTAVG was selected, as it was in all of the best models (Figure 4.12). PLANT was in the model with the lowest BIC and in all of the ten lowest RMSE models. As a third parameter DRAIN was chosen, as this combination had the lowest RMSE and was in the top ten lowest RMSE six times.

4.4.2.3 Major stage

For the major stage, DRAIN and PPTAVG were in all of the best AIC models, as seen in Figure 4.13. Outside of the two already selected parameters, there was a great variability of parameter combinations again, which made the selection difficult. FOREST and HLR were chosen, as they were included in the model with the lowest AIC. Additionally, that parameter combination had two of the ten lowest RMSE values.

Parameters chosen for the BIC model of the major stage were PPTAVG and FOREST. PPTAVG was included in every model, in combination with FOREST it resulted in the lowest BIC value. This combination also had seven of the ten smallest RMSE values, including the overall lowest.

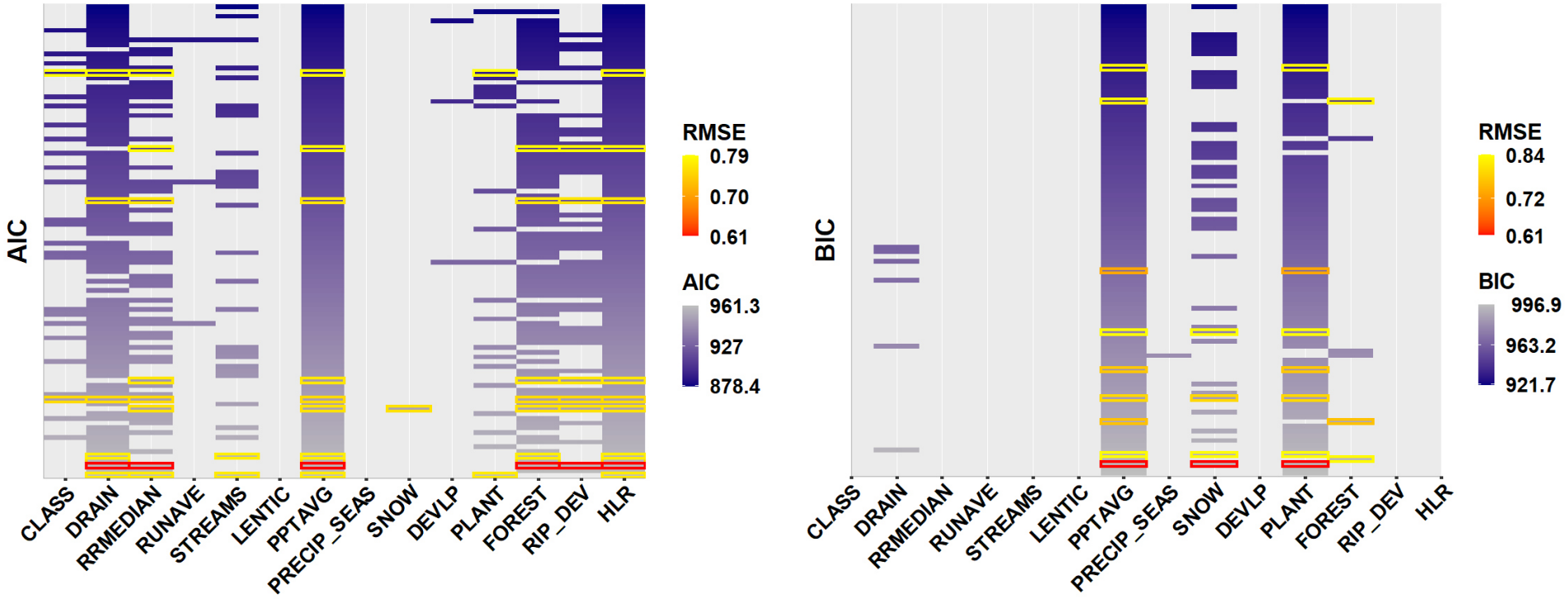


*Figure 4.11: Heatmap of the 100 AIC (left) and 100 BIC (right) selected models for the **minor** stage. The k-fold cross-validation was used, darker color indicated a better MOP value. The ten models with the lowest RMSE are marked as well, with darker colors indicating a lower RMSE.*

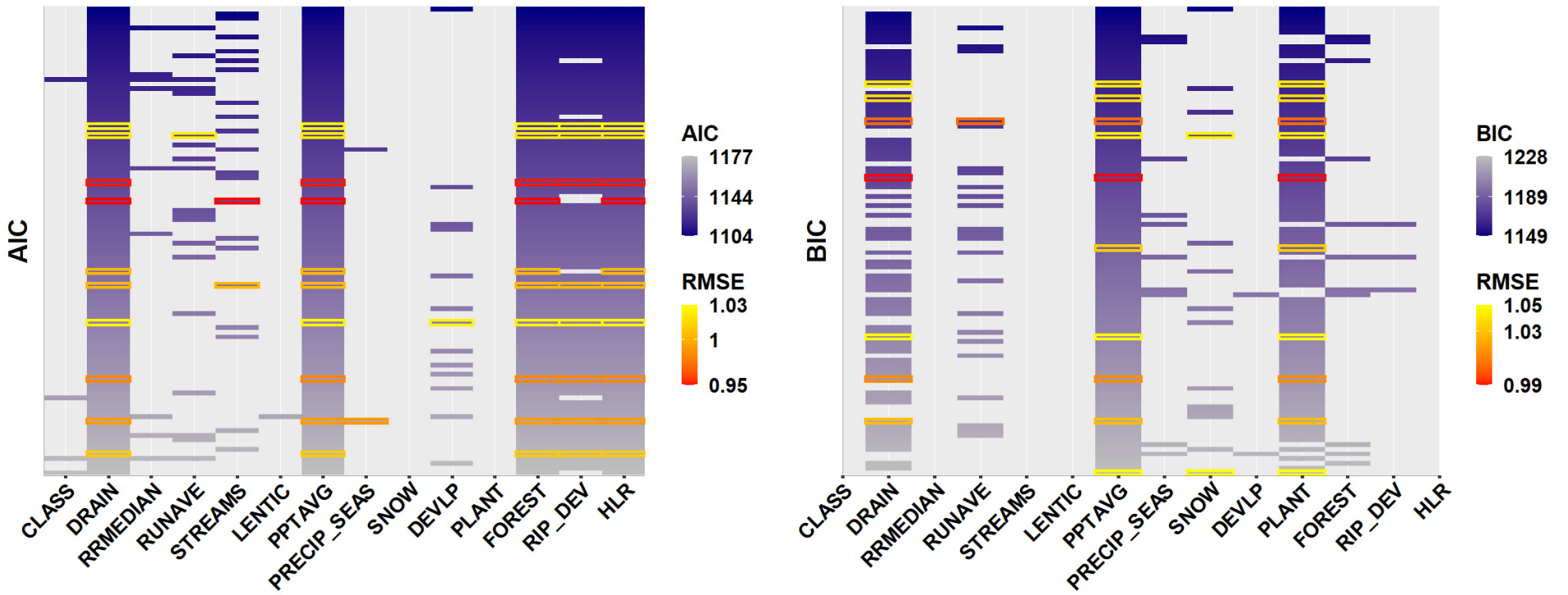


*Figure 4.12: Heatmap of the 100 AIC (left) and 100 BIC (right) selected models for the **moderate** stage. The k-fold cross-validation was used, darker color indicated a better MOP value. The ten models with the lowest RMSE are marked as well, with darker color indicating a lower RMSE.*

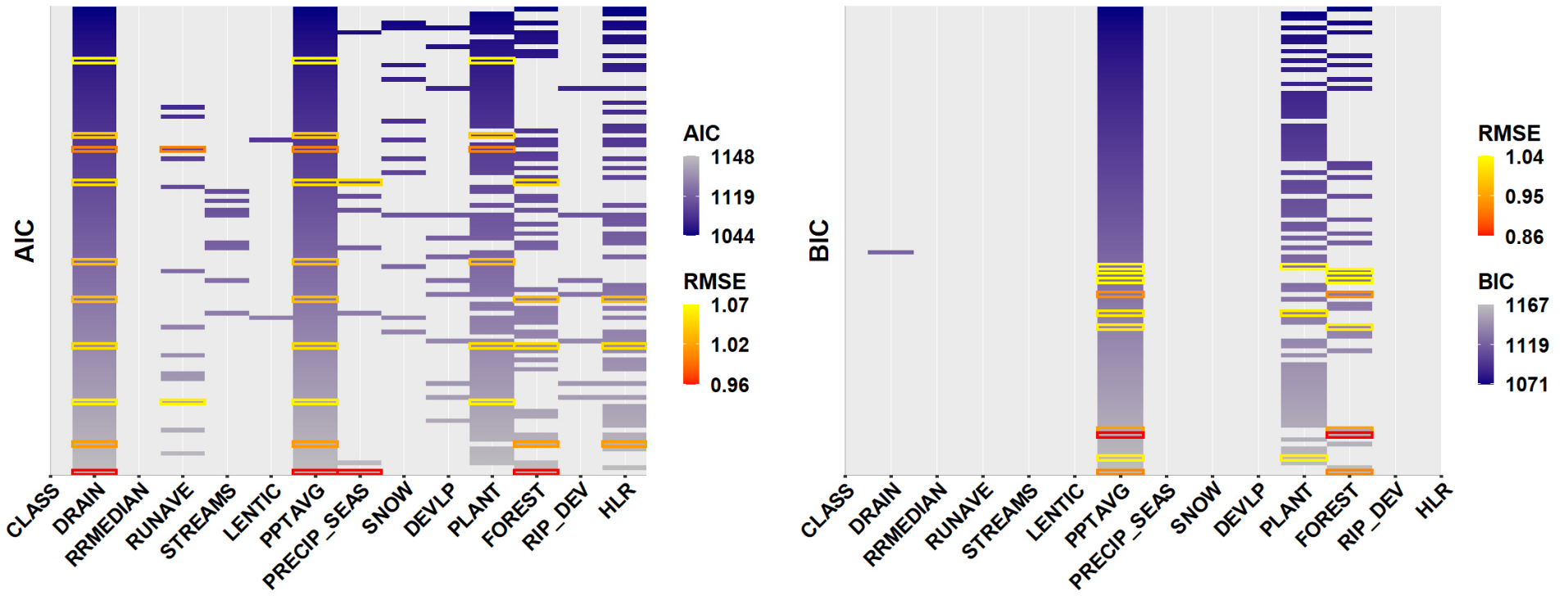


*Figure 4.13: Heatmap of the 100 AIC (left) and 100 BIC (right) selected models for the **moderate** stage. The k-fold cross-validation was used, darker color indicated a better MOP value. The ten models with the lowest RMSE are marked as well, with darker color indicating a lower RMSE.*

4.4.3 Model evaluation

The results of the k-fold cross-validation using the six models selected above can be seen in the following two tables below. Table 4.8 shows, that for the minor and moderate stage both the AIC and BIC model on average tend to overestimate the indicator significantly more than they underestimate. For the major stage, however, both models are underestimating much more than they overestimate. Comparing the models within each flood stage, the AIC model has a higher percentage of correctly guessed indicator values for all three flood stages. Though it only differs from the BIC models by a mean of 1,8 percent.

Table 4.8: Results of the prediction accuracy estimated using the k-fold cross-validation: mean percentage of stations underestimated, overestimated, and correctly estimated.

	MOP	Underestimated [%]	Overestimated [%]	Correct [%]
Minor stage	AIC	4.46	20.42	75.12
	BIC	3.67	23.20	73.13
Moderate stage	AIC	9.07	27.65	63.28
	BIC	8.67	30.17	61.16
Major stage	AIC	31.62	4.96	63.41
	BIC	36.05	1.71	62.23

In Table 4.9 the six models are compared using different model goodness measures. The comparison of the AIC shows a lower value for those models, that were chosen using the AIC as a MOP. The difference in AIC values matches the difference in the percentage correctly predicted in that sense, that it is largest between the models of the moderate and smallest between the models of the major flood stage. The same is true for the Loglikelihood. The values of the RMSE match the findings of the prediction accuracy. The minor stage has the lowest RMSE values and the highest correctly guessed indicator values, while the major stage has the highest RMSE values and the lowest percentage of correct predictions.

Table 4.9: Results of the evaluation of the models: comparison of the measure of performance (AIC, BIC), the goodness of fit to the data (loglikelihood), and the prediction quality (RMSE)

	MOP	AIC	Loglikelihood	RMSE
Minor stage	AIC	1029.78	-491.89	0.89
	BIC	1045.71	-517.85	0.93
Moderate stage	AIC	1268.37	-610.18	1.09
	BIC	1296.15	-643.08	1.13
Major stage	AIC	1220.67	-587.33	1.11
	BIC	1224.56	-608.28	1.13

Overall, the difference in prediction accuracy and RMSE of the AIC and BIC models is rather small. It does not justify using a model with more parameters (AIC) over a model with less (BIC). Therefore, the BIC models of all flood stages were chosen for a closer examination of the model parameter coefficients. Figure 4.14 shows a heatmap of the absolute values of the model parameter coefficients |β|. The algebraic sign of the coefficients is depicted as + or - within each tile. The model coefficients given by the output of *polr* were transformed into β-values as described in 3.2.5.2. It is important to remember, that the coefficients don't describe a direct influence on the category outcome of the response. Instead, they show the influence on the logit of the probability $P(y \leq r|x)$.

PPTAVG has the highest absolute β in all models, with around 4.5 in the minor and moderate and 3.3 in the major model. For the minor model SNOW and PLANT have similar absolute β values, however, PLANT has an increasing effect on the logit of the probability (1.16), while SNOW has a decreasing effect (-1.02). In the moderate model DRAIN and PLANT both have increasing effects, with betas of 1.33 and 1.19 respectively. FOREST in the major model has a decreasing effect on $logit(P(y \leq r|x))$, the β-value being -0.91. Regarding the β-values of all parameters, only SNOW and FOREST have a decreasing effect on the predictor.

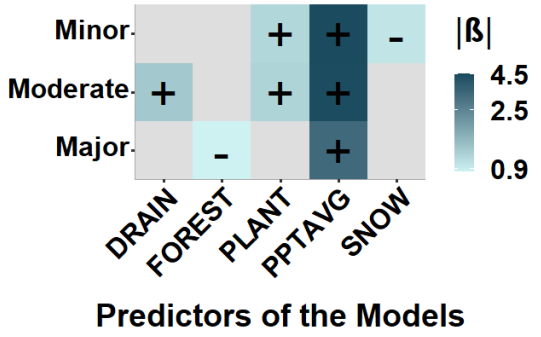


Figure 4.14: Heatmap of the absolute coefficient values |β| of the three chosen BIC models. The algebraic sign of each value is depicted with "+" for positive values and "-" for negative values.

The figure below shows, that all model parameters were statistically significant. The β-values of PPTAVG and PLANT were highly significant ($p < 0.001$), the other parameters had p-values well below the significance level of $\alpha = 0.05$ aswell ($p < 0.01$).

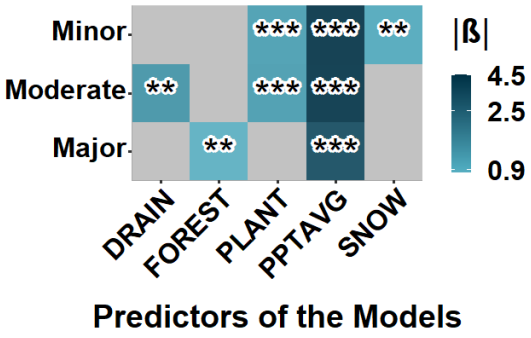


*Figure 4.15: Heatmap showing the absolute values of β and the significance of the coefficients of the BIC models. Significance coded as follows: 0 "***" 0.001 "**" 0.01 "*" 0.05 "·" 0.1 " " 1*

Table 4.10 shows, that the classification *in* was never correctly predicted by any of the three models. For the minor and moderate stage, *in* was more often underestimated than overestimated. For the major stage, the opposite was true with the majority of stations classified as *in* being overestimated. The tendency of the models to under- and overestimate is also shown in the count of classifications correctly predicted.

Table 4.10: Evaluation of the count of which classification category was underestimated, overestimated, and correctly estimated in the k-fold cross-validation for all three models

	Category	Under-estimated	Category	Over-estimated	Category	Correct
Minor stage	in	43	below	21	below	494
	above	126	in	4	above	39
Moderate stage	in	49	below	49	below	330
	above	168	in	16	above	115
Major stage	in	1	below	214	below	11
	above	11	in	52	above	438

5 Discussion

5.1 Methods

5.1.1 Flood frequency analysis

5.1.1.1 Independence criteria

When using the peak over threshold approach, assuring the independence of events is crucial. For this analysis the independence criteria included the catchment area, resulting in a varying time lag between independent events for each catchment. Another approach for assuring independence was introduced by Cunnane (1979), where events were required to be separated by three times the time to peak, calculated as an average over five hydrographs. He also suggested that the intermediate discharge value between two independent peaks must be less than ⅔ of the first peak. Both of those criteria were also used by Bayliss and Jones (1993) and by Bačová-Mitková and Onderka (2010). The US Interagency Advisory Committee on Water Data (USWRC) (1982) suggested that intermediate flows had to drop below ¾ of the lower peak. Ye et al. (2018) used both a 7 and 15-day interval between independent peaks.

As shown many different independence criteria have been proposed in literature, with none of them having a clear advantage. The here chosen criterium including catchment area is, thus, deemed a good choice as it was officially proposed for US catchments, by the US Interagency Advisory Committee on Water Data (USWRC) (1982).

5.1.1.2 GPD goodness of fit

Evaluating the GPD fitted plots showed, that the discharge of high return periods was often poorly estimated. The AD-test confirmed this, by indicating that for 240 stations when giving more weight to the tails of the distribution, the GPD was not a good fit to the data. This resulted in a wide range of calculated return periods and discharge values. Underestimation of the tail resulted in calculated return periods of a given discharge value being estimated too high, and discharge values assigned to a return period being too low. Overestimation of the tail led to the opposite, with discharge values of return periods being too high and return periods of discharge values too low.

Most of the return periods estimated for the minor stage were in a reasonable range between zero and 100 as seen in the violin plot (Figure 4.3). For the moderate and major stage, however, calculated return periods exceeded both annualities that were reasonably estimated without too much uncertainty (more than three times the time series length) and reasonable time periods. This further proves, that the tails of the distribution and with that upper return periods, were poorly estimated by the GPD.

In the literature a variety of possible distribution functions are introduced and used for the flood frequency analysis. Most notably the log-Pearson type III distribution, which has been the official model used for all US catchments since 1967 (US Interagency Advisory Committee on Water Data (USWRC), 1982). According to Meylan et al. (2012), the 2-parameter exponential distribution, along with the GPD, is often used to model the POT time series. Seeing how poorly the tails of the POT data were fitted for some stations, the selection of a different distribution might be advisable, especially since the validity of the further analysis heavily relies on accurate results of the flood frequency analysis. The method of parameter estimation could of course also be changed, but seeing as L-moments are considered superior for estimating the GPD, the change would most likely not improve the fit (Hosking, 1990; Sankarasubramanian and Srinivasan, 1999; Zea Bermudez and Kotz, 2010a).

The chosen threshold also influenced the calculated return periods, as the number of events per year is included in the μ parameter in formula (3.14). A lower threshold will lead to a higher number of events per year, consequently decreasing μ and the obtained return periods. However, this influence is not relevant to the goodness of fit of the GPD, as μ is included in both the estimated return periods using the GPD and the calculated return periods of the observations.

5.1.1.3 Flood stage exceedance and infinite return periods

Results in 4.2.4 showed, that the calculation of return periods of flood stage triggering flows resulted in infinite values for some flood stages. The return period was calculated using formula (3.14) and a non-exceedance probability P_U, obtained using the cdf previously fit to our POT time series. The closer P_u is to one, the higher the return period calculated. For the return period to be infinite, $1 - P_U$ has to be zero, which was the case with a non-exceedance probability of one.

When calculating P_U, the bottom term in formula (3.9), takes into account, that the maximum observed POT value, may be exceeded. Therefore, the maximum POT discharge value does not have a $P_U = 1$, but the cumulative non-exceedance probability monotonically approaches $P_U = 1$ with increasing discharge values Q. For some stations, the difference between the maximum observed Q_{POT} in the POT time series and Q_{Stage} of flood stage triggering flows was very high. The highest Q_{POT} already had a high non-exceedance probability, an even higher Q_{Stage} resulted in a value of P_u being very close to or even reaching one. This in turn led to the return period becoming very high or even infinite.

The value of μ was not relevant for this, meaning the chosen threshold did not influence whether a station´s return period was infinite or not. The threshold only limited the lower bound of discharge values chosen, and, therefore, events per year, it at no influence on the highest peaks.

A comparison of stations with infinite return periods and stations where flood stage triggering flows were never reached or exceeded in the chosen time period, showed a pattern. For those stations, where the return periods of all stages were infinite, neither flood stage was never triggered. An exception to this is explained in detail later. When the moderate and major flood stages had infinite return periods, neither of both flood stages was triggered, and additionally, for half of those stations, the minor stage was also never triggered. Of those stations with infinite return periods for the major stage, for all except one station, the stage triggering discharge was never reached. The exception for the major stage was station ´01500000´, Figure 5.1 depicts the GPD fit to the POT data. It shows, that for a P_u below 0.75, the GPD estimates the actual discharge very well. Above that, however, it first slightly overestimated and later significantly underestimates Q. $P_U = 1$ is reached way below the highest POT value of Q_{POT} = 34 mm/d. The same was true for the estimation of return periods (Figure 5.2), while the GPD estimated the discharge well for return periods below two years, above 15 years it significantly underestimated Q. This underestimation led to the infinite return period for the major stage, because while Q_{Stage} = 32 mm/d was smaller than the max POT discharge Q_{POT} = 34 mm/d, the fitted GPD already reached $P_U = 1$ at around 24 mm/d.

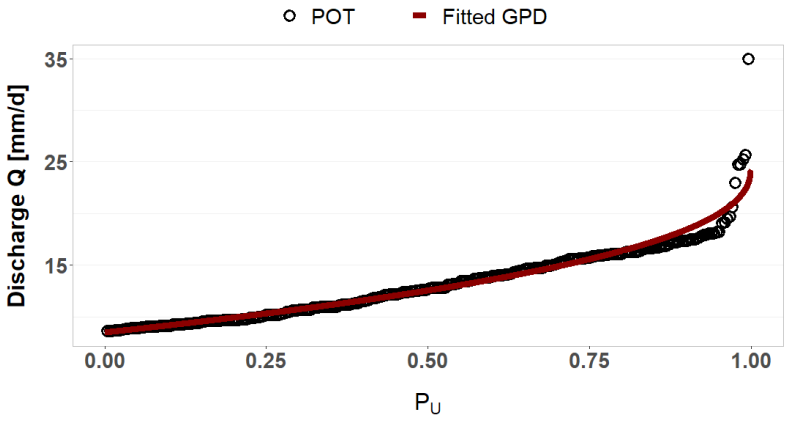


Figure 5.1: POT data for station ´01500000´(black) and fitted GPD (red): Cumulative non-exceedance probability P_u

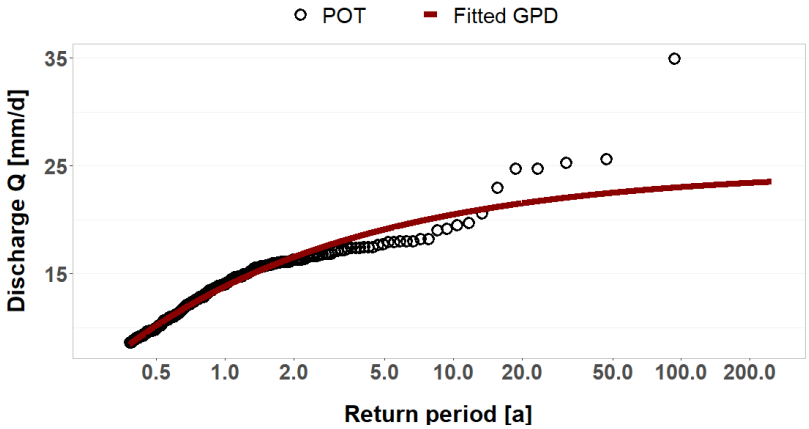


Figure 5.2: POT data for station ´01500000´(black) and fitted GPD (red): return period in years

5.1.2 Regression

For the ordinal logistic regression to be used, the proportional odds assumption had to be valid. The Brandt test was executed along with a visual evaluation of the assumption. Calculation results confirmed the proportional odds assumption for all stages, however, visual evaluation was less conclusive. When plotting the odds, they were deemed proportional if the categories of a parameter had a similar value, depicted as a similar distance from zero on the x-axis (right end of the plot). Figure A-1 in the appendix shows the visual results exemplary for the minor stage and parametes of the final minor model. While the distance was similar for most parameters, especially for HLR the difference was rather large. Unfortunately, no exact numbers could be found for the distance, above with the assumption is not valid. But seeing as the distance was very similar for those parameters, that were selected for the final regression models, and the calculated results confirmed the proportional odds assumption, ordinal logistic regression was deemed applicable.

5.2 Research questions

5.2.1 Question 1: Classification evaluation

The results of the classification showed, that the impact-based flood stages matched the statistical thresholds on average for only 8% of stations, totaled over all stages. For the minor and moderate stage, a majority (71% and 52%) of stations was classified as below. For the major stage, the majority (62%) of stations were classified as *above*.

As stated in the introduction, the conference paper by Anderson (2016) is the only comparative analysis of impact-based and statistical thresholds found in the literature. She found a strong relationship between statistical and impact-based thresholds, which was different from the one given by the APRFC (Alaska-Pacific River Forecast Center) (NOAA, 2019). In Anderson´s analysis, the major flood stage was best approximated by a return period of 100 – 500 years and the moderate flood stage by a return period of 25 – 50 years. She relates the high difference between recurrence intervals and flood stages for the major stage to the rarity of major events and the scope of available gauge data in Alaska.

Comparing the median return periods calculated for each flood stage to Anderson´s findings, the moderate stage median of 9.3 years is below the 25 – 50 years range she found, while the median for the major stage of 184 years lies in the lower 20% quantile of her given range of 100 – 500 years.

To better relate her findings to the ones of this thesis they are classified according to 3.2.3. Both the moderate and major stages are classified *above*, the moderate stage by about ten years, the major stage by 50 – 400 years. For the major stage, this matches the findings of this analysis, with the majority of CONUS stations being classified as *above*. For the moderate stage, however, results obtained here indicate a lower return period of the stage triggering flow, opposite to her findings. For

the minor stage, no results were reported by Anderson. When comparing the results, one has to keep the different study sites in mind, especially in regards to the population density of the analyzed watersheds. Seeing as flood stages are based on impacts on anthropogenic infrastructure and population, a higher population density means an increased number of targets for flood impacts and increased causal effects.

Using the GAGES II database (Falcone et al., 2010; Falcone, 2017), the two study sites are compared to try and explain the differing results. Anderson (2016) used Alaskan catchments for her analysis, as no catchment IDs are given, all Alaskan catchments included in GAGES II are used for the comparison. The 87 catchments in Alaska have a mean residential population of 0.8 residents/km^2, the 727 catchments in CONUS used for this thesis have a mean of 61 residents/km^2. When it comes to the percentage of developed area in the watersheds, for Alaska the maximum is 9%, with a mean of 0.3%. For catchments in CONUS, the maximum is 97% and the median is 7% of a catchment being developed area.

This might explain why Anderson´s results would classify the moderate stage as overall *above*, seeing as in less populated areas there is also less developed area and with that less infrastructure to be damaged by a flood. That in turn leads to flood stages being triggered at higher discharge values, resulting in higher return periods than the reference return periods given by the APRFC. Transferring this to the catchments of CONUS, more development in the catchment and around the stream leads to lower flows causing more damage than in undeveloped areas. Lower flow means lower return periods, hence, why the minor and moderate stages are classified as below for the majority of stations and their return periods are below APRFC reference return periods.

Anderson also mentioned outliers in the relationship between statistical and impact-based thresholds, though no explanation for them was found by her. Similarly, the flood frequency analysis performed here also resulted in outliers, leading to a wide range of calculated return periods for flood stage triggering flows. The methodical reasons behind this wide range have already been discussed in detail in 5.1.1.2 and 5.1.1.3, here only the implications for the classification are explored. The poor fit of the GPD, especially at higher return periods leads to faulty classifications of stations. For the minor stage, that error is the smallest, as the plots show a rather good fit to the data here. For the moderate stage, both underestimation and overestimation can be observed in the plots, for the majority of stations, the deviation is limited to an acceptable extent. However, for the major stage, the poor fit leads to a grave overestimation of the classifications. The return periods of discharge values are overestimated because the discharge is underestimated.

For the major stage, one has to keep in mind, that at many stations the discharge level necessary to trigger said stage was never reached in the time series. This leads to very high and in some cases even to infinite return periods as explained in 5.1.1.3. While the exact numbers calculated for the

return periods might not be valid, the classification resulting from them can be. Because if a discharge value is never reached and as for some stations, much higher than the largest POT value Q_{POT_max}, its return periods must be very high, compared to the return period of the maximum POT. Considering that the smallest maximum return period of the POT data T_{POT_max} is 76 years, with a median of 101 years, that leaves 397 stations with a maximum POT return period above 100 years. If the stage triggering flow is now higher than the maximum POT flow, for more than half of the stations that should automatically lead to the classification of *above*, without fitting a distribution. Looking at the classifications of those stations with $T_{POT\_max} > 100$ and $Q_{Stage} > Q_{POT\_max}$, only 113 are actually classified as *above*. For 418 stations the major stage was never reached in the time series, and they were classified as *above*, including the 113 stations extracted before ($T_{POT\_max} > 100$ and $Q_{Stage} > Q_{POT\_max}$). So, 305 of the 418 stations were classified as *above*, despite only upholding one of the criteria mentioned before ($Q_{Stage} > Q_{POT\_max}$). Examining the difference between the major stage triggering discharge and the maximum POT value for those stations gave a median difference of 70 mm/d and a 25% quantile of 16 mm/d. This difference was large enough, to lead to high return periods as explained in 5.1.1.3. These results prove that despite the poor fit of the GPD, the classification *above* can still be assumed as valid for the majority of stations, even if the calculated return periods are overestimated.

The classification of *below* and *in* for the major stage on the other hand is more dependent on the correct estimation of the return period flows. For the classification as *in*, stage triggering flows can only take up a limited, much smaller range of flow values, compared to *above* and *below*. Poor estimation of that range can quickly lead to over or underestimation of the classification. For the classification as *below*, it can be observed, that $Q_{Stage} < Q_{POT\_max}$ was true, which is reasonable considering the return periods of Q_{POT_max}.

To summarize there is a high estimation inaccuracy of the fitted GPD compared to the calculated return periods of the POT data, which leads to possible faulty classification. This error is smallest for the minor stage, seeing as the GPD estimates low return periods well. The moderate stage is both undecimated and overestimated, with an overall acceptable deviation for most stations. For the major stage, the estimation error is highest. However, despite the significant underestimation, the error of classification is balanced out and the results can be used for further analysis.

5.2.2 Question 2: Classification variability

5.2.2.1 Spatial variability

When plotting the classification for all flood stages separately, patterns could be seen both across all three stages and varying with increasing flood stage. Those patterns are now compared to the hydrologic landscape regions shown in Figure 3.2.

Stations of the minor stage that were classified *below* mostly exhibit a humid or sub-humid climate, but they can also be found in areas with semi-arid climates and impermeable soils. This is true for both stations in mountain areas and stations located along the Atlantic Plain and central Lowlands. Stations classified *above* are located in regions of semiarid mountains (Colorado, Utah, New Mexico, Wyoming) and humid mountains with permeable soils in the Appalachian Highlands. Stations classified as in can be found in both humid and arid climates.

For the moderate stage an increase in stations classified *above* is visible. Stations in humid and sub-humid areas are still mainly classified as *below*, as seen in the diagonal zone from Illinois to Texas and the center of the Appalachian Highlands. Stations classified *above*, as for the minor stage, exhibit an arid and semi-arid climate. However, there is an increase in stations classified *above* located in humid areas, that were previously predominantly classified *below*.

With the majority of stations being classified *below* for the major stage, they are now located in both areas with a humid and arid climate. Those stations that remained *below*, are still in regions with humid or sub-humid climates, the most prominent locations being the Appalachian Highlands, Central Lowland, and the western Atlantic Plain.

The majority of catchments with a humid or sub-humid climate were classified *below*, the higher precipitation in those regions leads to higher discharge and with that a more frequent exceedance of flood stage triggering flows. Additionally, in regions with impermeable soils, lack of percolation leads to an increase in surface runoff flowing to the streams and with that an increase in discharge. In arid regions the precipitation is lower than the evapotranspiration, the lower amount of rainfall combined with the high evapotranspiration leads to less runoff and with that less frequent exceedance of flood stage triggering flows and the classification as *above*. The permeability of soils adds to that, as percolated water does not directly contribute to stream discharge values (not taking exfiltration into account). The climate characteristic of the HLR regions seems to have the biggest influence on the spatial distribution of the classification, while the permeability of soil and bedrock is only secondary.

5.2.2.2 Hydro-climatic parameters

The flood stage threshold is determined by the area around the stream, depending on urbanization, population, and anthropogenic land use. The recurrence of flood stage exceedance is determined by the overall amount of discharge at a gauge, which depends on those characteristics of the entire catchment, that influence runoff formation processes. The classification of stations is dependent on the latter, as it states how often a flood stage triggering flow is reached. In the following, results from literature are compiled on catchment characteristics that influence flood formation processes and those that can be used to determine flood hazard.

Anderson (2016) anecdotally analyzed, whether the outliers she found could be explained using catchment characteristics such as mean annual discharge, population, or ice effect, though no relationship was found. Merz et al. (2014) stated that a traditional view is that catchment characteristics such as topography, geology, and meteorology influence the formation of floods. Hollis (1975) stated that the effect of urbanization is decreasing with increasing flood magnitudes, as the importance of interception, depression storage, and infiltration decreases. O'Driscoll et al. (2010) confirmed these findings, that urbanization leads to an increase in smaller peak flows, while the increase is less pronounced for flows with higher return periods. According to O'Driscoll et al. (2010) urbanization also leads to an increase in discharge variability and flashiness.

Saharia et al. (2017) performed an analysis using data from the NWS flood event database, to examine the variability of temporal and spatial flood characteristics in CONUS. They used the Köppen-Geiger climate classification to investigate the influence of climatic regimes and other parameters such as basin area and relief on flood magnitudes and flooding rise time. They found precipitation to be the primary driver of floods, with the magnitude of peak discharge depending on the causative rainfall. The catchment area also influenced the unit peak discharge, although in mountainous areas the relief had a greater impact than the basin size.

When thinking of influences on the recurrence of flood stage exceedances, the determination of flood stage flow values has to be kept in mind as well. If the area around the gauge is highly developed a lower flow value will have a greater impact, than if the area was only sparsely populated. The lower the flood stage flow value, the more often it will get exceeded, which is why characteristics influencing the flood risk also influence the classification applied in this thesis. Assessing the flood hazard of a catchment, Balica et al. (2009) used a large number of catchment criteria for calculating a flood vulnerability index. They divided into social, economic, environmental, physical components, each containing different catchment characteristics. Parameters such as percentage of urban area, land use, rainfall, and river discharge were used to determine the exposure (damage potential) to floods.

The results of the correlation analysis (4.4.1) showed only a weak correlation between the classification and catchment parameters, meaning that no single catchment characteristic can sufficiently explain the spatial variability of the classification. Examining the correlation coefficients per flood stage showed, that for the minor stage the annual mean precipitation had the highest correlation to the indicator, while for the moderate and major stage it was land-use characteristics. One could hypothesis from this, that for the minor stage and with that flows of smaller recurrence the precipitation is more important than land use. For higher stages and with that higher return periods, land use and its influence on runoff processes become more important than precipitation. However, the correlation coefficients of those parameters are only differing by small amounts and

the correlation itself is very small. So, while a pattern is visible in the coefficients and they are all significant, the correlations are too similar to reliably draw conclusions from them.

The literature review showed that many catchment characteristics influence the formation of runoff, reaffirming the results of the correlation that no single parameter can sufficiently explain the relationship between statistical and impact-based thresholds. To model the relationship a combination of parameters needs to be used. A closer examination of the selected model parameters is carried out in the following chapter.

5.2.3 Question 3: Model evaluation

5.2.3.1 Accuracy

Stepwise model selection and model evaluation were used to choose catchment characteristics for classification prediction models. With a prediction accuracy between 61% (moderate stage) and 73% (minor stage), it is possible to model the relationship between statistical and impact-based thresholds. A closer examination of the prediction reveals, that the classification *in* was never correctly predicted by any of the three models. This is most likely due to only a small number of stations being classified as *in* (8%), which is not enough for the models to properly reproduce the category. It can also be seen in Figure 4.7 and was examined in the analysis of the spatial variability in 5.2.2, that the stations classified as *in* do not share similar locations or HLR characteristics. This increases the difficulty of correctly predicting this category.

The difficulty to estimate categories that have fewer observations can also be seen for the classification of *below* and *above* for all three stage models. For the minor model, the majority of observations used to train the model were *below*, which led to an estimation accuracy of 95% for the *below* category but an underestimation of the other categories. For the major model, a majority of *above* was observed as the station category, resulting in estimation accuracy of 97% for the *above* category, with the other categories being overestimated. The observations used to train the moderate model were more balanced in the number of stations classified *above* and *below*. While there still is a tendency to underestimate, the model correctly predicts 87% of *below* and 40% of *above*.

The clear influence of data distribution across the three categories is especially apparent for the *in* category. When the model was trained with a majority of *below* observations, the majority of predictions for stations classified as *in* were *below*. The opposite is true if the majority of observations were *above*, resulting in the overestimation of the *in* category, the majority being predicted *above*. More balanced observations in regards to the categories led to a smaller difference between over- and underestimation of the *in* category.

This shows that to increase the prediction accuracy across all classification categories, the observations used to train the model need to be more evenly distributed across all categories.

5.2.3.2 Parameters

When analyzing the spatial pattern of the classifications across all stages, climate variables of the HLR regions could be used to explain the observed variability. It is, therefore, reasonable, that of the catchment characteristics selected for the regression, the mean annual precipitation (PPTAVG) is included in the regression models for all stages.

In the previous comparison of the results with Anderson (2016), the percentage of developed area in the catchment (DEVLP) was named as a possible explanation of the differences. Looking at the regression models, that parameter is not included in either of the three models. It is possible that seeing as we have a higher overall percentage of developed area in our catchments, the importance of the parameter is smaller than the combination of other catchment criteria. Looking at the correlation between other catchment characteristics and DEVLP, the parameter might have been replaced with another that is correlated with DEVLP but has a higher influence on the indicator. There is a moderate correlation between DEVLP and PLANT, a parameter that is included in two of the final models. Additionally, there is also a moderate correlation between DEVLP and HLR, HLR being included in all three AIC models. Taking the almost perfect correlation between DEVLP – RIP_DEV into account, parameters correlated with RIP_DEV should also be useable as a replacement of DEVLP.

Comparing parameters used in the AIC and BIC models, the correlation between parameters is examined to try to explain the different selections. HLR was included in all AIC but in no BIC model. Examining the correlation between HLR and other parameters, there is a moderate correlation between HLR – FOREST and HLR – SNOW. For the minor stage, the correlation HLR – SNOW may have led to the inclusion of the SNOW parameter, in the model. For the moderate stage, a moderate correlation between RIP_DEV and FORREST with PLANT and a strong correlation between HLR – PLANT might explain the replacement of the three parameters with PLANT for the BIC model. The moderate correlation between PLANT and FOREST led to an alternation between the two parameters in the 100 best models of the stepwise model selection. This is especially visible for the major BIC model, as one of the two is included in all models, but never both.

The model selection process showed, that a large number of catchment characteristics can be used to model the relationship between statistical and impact-based thresholds, with only small differences in accuracy. It also showed that those parameters impacting runoff processes have a bigger influence on the classification than those determining the flood stages. Neither RIP_DEV nor DEVPL was included in the final models.

Before considering the hydrological reasonability of the coefficient signs, the effect of the parameters on the response variable is examined. PPTAV, PLANT, and DRAIN all have an increasing effect on the $logit(\mathrm{P}(y \leq r|x))$. This means the higher the values of those predictors, the higher the log-odds

of being in a category or below. For FOREST and SNOW the opposite is true, increasing values decrease the log-odds. To translate this to probabilities of being in a category (*below* < in < *above*): an increase in PPTAV, PLANT, and DRAIN results in an increased probability of the predictor falling in a lower category, while an increase in FOREST and SNOW results in an increased probability to fall in a higher category.

To get back to the hydrologic reasonability, the influence of the parameters on the probability and with that the category makes sense considering flood formation processes. Seeing as precipitation is the primary driver of floods, it is reasonable that the PPTAVG parameter has the highest regression coefficient (Saharia et al., 2017). An increase in precipitation leads to an increase in runoff and with that higher discharge values. Those, in turn, result in flood stages being triggered more often than the assigned return period range would suggest.

Seeing as snowpack is a natural water reservoir, an increased percentage of snow of the total precipitation reduces winter floods and results in less low flow values in spring and summer due to snowmelt. Davenport et al. (2020) found, that an increase in rain fraction leads to larger peak discharge values, a shift from snow to rain increases the flood risk. Over all analyzed watersheds, rain-dominated catchments had floods with significantly larger peak discharge flows than snow-dominated basins. The higher the SNOW parameter value, the higher the percentage of snow of the total precipitation. With snow decreasing flood peak magnitude and flood risk, it makes sense that increasing parameter values increase the probability of falling in a higher classification category, as the return periods of high flows are increased.

Several studies reported that with increasing drainage area, the discharge also increases (Furey and Gupta, 2005; Curran et al., 2016; Saharia et al., 2017). In the model, an increase in drainage area increases the probability to fall in a lower classification category and with that indicates decreased return periods meaning increased frequency of peak discharge values.

PLANT refers to the percentage of catchment area used for agriculture. Schilling et al. (2014) modeled the conversion of cropland to perennial vegetation, to examine the effect on floods. They found, that an increase in perennial vegetation reduced the number of flood events and the frequency of severe floods. Hounkpè et al. (2019) modeled different land-use scenarios focusing on changes in land use and land cover. They found that land-use changes significantly affect the magnitude and frequency of floods. The expansion of agriculturally used areas and the decrease of natural vegetation, such as forests, led to an increase in flood characteristics. Those findings match our model parameter, as an increase in PLANT increases the probability to fall in a lower category, meaning stage triggering flows are exceeded more often.

The effect of forests on floods varies in the literature. Bathurst et al. (2017) state, that the runoff-reducing effects of forests are most significant for small storms and are increasingly less effective as

precipitation increases. According to them, changes in forest cover have little effect on the peak discharge of larger floods with return periods of ten years or longer. Alila et al. (2009) on the other hand found, that in some catchments, forests may reduce flood frequency over all flood magnitudes. Rogger et al. (2017) commented on the contradicting effects of forests in literature, naming varying methods and gaps in research as the reason. In a more recent study, Bathurst et al. (2020) found, that forested catchments have a lower peak flow magnitude for given a flood frequency and a higher return period for given flow magnitude. In our model, FOREST increases the probability to fall into a higher classification category, therefore, increasing the return periods of certain magnitude flows. Bathurst et al. (2020) also stated that the relevance of land cover decreases for the largest events on record. However, in this thesis FOREST is included in the major model, therefore, affecting higher return periods and flows partially much larger than the highest flow value of the time series. This contradicts the findings of Bathurst et al. (2020), as forests are only relevant for high return periods, and not included in the other models.

Curran et al. (2016) performed regression analysis, using catchment characteristics for estimating flood magnitude and return periods for ungauged catchments in Alaska and Canada. Their results showed that the log of the drainage area (DRAIN) was the strongest explanatory variable, followed by the log of the mean annual precipitation (PPTAVG). They used the same parameters for estimating discharge corresponding to different annual exceedance probabilities, only varying the parameter coefficients. It is important to note, that their regression equation was developed for ungauged catchments, where peak flows are not significantly affected by urbanization, impervious surfaces, and flow regulation measures.

The regression models built in this analysis do not predict a discharge value but classify whether certain discharge values are *below*, *within*, or *above* an assigned range of return periods. Nonetheless, both regression models aim to describe flood formation processes. In this thesis the question is, which parameters increase the frequency of flood stage triggering flows, leading to the classification of *below*, and which parameters decrease the frequency, classifying a station as *above*.

In the final models selected here, PPTAVG has the highest parameter coefficient and with that is the strongest explanatory variable, while DRAIN is only included in one of the three models. Curran et al.'s (2016) models are designed for catchments with small anthropogenic impacts, while the catchments analyzed in this thesis show great variability of anthropogenic changes (e.g. developed area). Additionally, their models predict flood magnitude corresponding to a return period, while the models here classify the flood frequency of certain flow values, this explains the difference in parameter selection.

5.2.4 Question 4: Impact-based vs statistical thresholds

The analysis has shown, that recurrence intervals are not a good indicator of how much impact a flood will have on the area around the stream. Only very few stations were classified as within (*in*) the return period range assigned to a flood stage. The calculated return periods of flood stage triggering flows exhibited a wide range of values, exceeding figures both reasonably estimated and comprehensible. Statistical thresholds are, therefore, no alternative to flood thresholds based on the assessment of flood impacts.

The normalized discharge (mm/d) at which a flood stage is triggered strongly differs across catchments, as it depends on the anthropogenic use of the area around the stream. In an area with higher population density, urbanization, or exposure of infrastructure the flood stage will be assigned a lower stage value compared to anthropogenically sparsely used areas. The number of exceedances of a flood stage and with that its recurrence is also vastly different across catchments, as it depends on the catchment characteristics, that influence runoff formation processes, which vary in space.

In addition, the non-stationarity of said catchment characteristics can lead to changing frequency and magnitude of flood events over time, making recurrence intervals an unreliable indicator for flood impacts. While the frequency of flood stage exceedances changes, the actual gauge height assigned to a flood stage does not. That is unless changes to the area around the stream are made, such as an increase in anthropogenic use and infrastructure, making it more vulnerable to flooding and, therefore, lowering the flood stage triggering flow.

Statistical thresholds determine theoretical flows, that can, due to their dependence on the discharge time series, vary with time. Impact-based thresholds are practical flow values, based on observations of flood impacts that, if the area around the stream is not modified, always stay the same.

In Germany, many gauges have a color-coded indicator classifying the water level, that can be accessed online. In Baden-Württemberg, the indicator gives a range of return periods, a statistical threshold, that the discharge level corresponds to. This indicator does not signal how much flooding is to be expected or what actions need to be taken unless that is explicitly stated elsewhere or remembered from experience.

The classification of flood stages in the US clearly states what impacts are to be expected, when a certain water level is exceeded. The meaning and extent of impact caused by each flood stage in theory, stays the same, even if the flood stage gauge height changes. The same is true for the majority of gauges in Germany, which also apply an impact-based classification of the water level.

While statistical thresholds are easier to obtain since they only require a time series of discharge, a classification of floods based on their impacts is preferable. Impact-based thresholds are less vulnerable to non-stationary catchment characteristics influencing runoff formation processes.

Additionally, they are easier to comprehend for residents of the impacted area, as they directly name expected flood impacts.

5.3 Implications

It has been proven above, that statistical thresholds are not a good alternative to impact-based thresholds, but what does that mean for gauges where statistical thresholds have to be used because no impact-based thresholds are available? The previous classification of stations in *below*/*in*/*above* will be used to explain the implications arising from the found relationship between statistical and impact-based thresholds.

If a station was classified *below*, it meant that $Q_{Stage} < Q_{T\_lower}$ and $T_{Stage} < T_{upper}$. Using Q_{T_lower} for the construction of flood protection measures and as the flood stage triggering flow, would lead to an underestimation of floods. Actual impacts would already happen at lower flow values and with that statistically more often as well. Flood protection measures would be effective too late, as flood impacts are already caused by lower flows. The same is true for the assigned flood classification, as corresponding impacts would already happen before the flood stage is reached, misinforming and potentially endangering residents.

The classification of a station as *above* meant that $Q_{Stage} > Q_{T\_upper}$ and $T_{Stage} > T_{upper}$, resulting in an overestimation of flood impacts. Impacts to the area around the stream happen at higher flows and with that statistically less often. Flood protection measures would be in effect at flows not causing flooding, their construction being an unnecessary waste of money and recourses. In theory established measures should of course also protect against higher floods, making them not ineffective, their extend, however, would be inappropriate. Considering the warning function of the flood categories, residents would be urged to avoid certain areas or evacuate too soon, leading to unnecessary concerns and actions taken.

Stations classified *in*, exhibit flood stage triggering flows within the flood stage corresponding return period range, making flood protection measures that are built based on Q_{T_lower} to Q_{T_upper} reasonable. However, the possibility of under- and overestimating floods is also given here. It is important to carefully consider the selection of a return period from the given range T_{lower} to T_{upper} to assign the flood stage to, considering the resulting flows Q_{T_lower} to Q_{T_upper} differ by up to 53 mm/d. So even if the flood stage triggering discharge falls within the range, it does not guarantee that the flood measures based on a selected return period are appropriate.

When a classification of the relationship between statistical and impact-based thresholds is available, the return periods assigned to flood categories can be adjusted accordingly. As presented in this thesis selected catchment characteristics can be used to model said relationship. However, the models as shown here can only determine a classification category relative to T_{lower} and T_{upper}, they do not

indicate how much lower or higher a flood stage needs to be set, compared to the assigned return periods.

5.4 Methodical considerations

The following subdivision follows the scheme proposed by Yen (2002), who considered the uncertainties of flood frequency analysis. That scheme is applied to discuss the FFA and the implications of the results, as well as the selected catchment characteristics.

5.4.1 Natural uncertainty

For the analysis of this thesis, stationary conditions were assumed, meaning the statistical properties (mean, variance) of the time series do not change with time. However, the natural conditions are not stationary, exhibiting trends of time dependant mean or variance (Bauer, 2021).

Multiple studies have found trends in the flood time series, results both showing increasing and decreasing flood properties with strong spatial variability across all catchments of CONUS (Hirsch and Archfield, 2015; Mallakpour and Villarini, 2015; Archfield et al., 2016). Spatial patterns of increasing and decreasing frequency can also be observed in the flood stage exceedances (Slater and Villarini, 2016). The influences of selected catchment characteristics on runoff processes and both flood magnitude and frequency have been described in 5.2.3.2. This clearly shows, that with the discharge time series being proven nonstationary, the catchment characteristics influencing runoff formation processes must also be nonstationary.

Slater and Villarini (2016) named the following dynamic parameters as reasons behind increasing flood frequency: changes in precipitation, snowmelt patterns, land use and cover, antecedent soil moisture, and anthropogenic modifications of the water cycle. The non-stationarity, thus, is not just due to global-scale shifts in atmospheric conditions, like climate change, but also caused on a much smaller scale. The influences of land-use change on the catchment scale have been previously described (Hounkpè et al., 2019), as well as the effects of urbanization (Hollis, 1975; O'Driscoll et al., 2010). Anthropogenic modifications like dams reduce annual peak discharge by up to 90% when comparing the unregulated reach above a dam with the regulated reach downstream of a dam (Graf, 2006; Mei et al., 2017). Retention basins have a similar effect, though on a smaller scale. They reduce flood peak magnitudes but increase the recession time and magnitude at the recession of flood hydrographs (Soong et al., 2009). Modifications in channel capacity and roughness can significantly alter the frequency of floods on a local scale (Slater and Villarini, 2016).

Merz et al. (2014) state, the traditional approach of performing flood frequency analysis under the assumption of a stationary time series needs to be extended to account for non-stationary climate and catchment characteristics that result in changing flood characteristics. Abrupt changes in the discharge time series can be a result of changing land use and cover, gauging practice, or flood

regulation by dams and retention basins. To assess the non-stationarity of parameters an inspection of the discharge record and the characteristics of the catchment is necessary (Villarini et al., 2009).

While stationarity is assumed for the analysis, literature shows that the natural conditions are non-stationary. This results in the calculated flood probabilities not accurately representing the current probability distribution. The time series of the used stations need to be examined for trends and the non-stationary conditions need to be taken into consideration when calculating the probabilities. This also means examining the catchment characteristics for temporal changes.

5.4.2 Model uncertainty

A visual examination of the fitted rating curve to the rating data showed a poor fit for many stations. This results in the conversion from ft^3/s to mm/d being inaccurate. However, since the conversion is used for all compared data, the mistake is the same over all calculations and with that becomes negligible.

As stated in 5.1.1.2, the fitted distribution is not a good approximation of the tails of the POT data. This poor estimation leads to faulty return periods and with that to possibly incorrect classifications of stations mostly for the major stage. While, as shown in 5.2.1, the classification of stations for the major stage is reasonable for most stations, using a better fitting distribution is still advisable to reduce the classification error. As stated by Kidson and Richards (2005) and the US Interagency Advisory Committee on Water Data (USWRC) (1982) the log-Pearson III distribution is the officially recommended distribution for flood frequency analysis on both AMF and POT data of US catchments. Anderson (2016) used if for her analysis, as did Curran et al. (2016).

Curran et al. (2016) also state, that when calculating the discharge corresponding to a certain annual exceedance probability, the skew coefficient determines the curvature of the flood frequency curve. Said coefficient is estimated from flow data, which for short periods of record is an unreliable estimation of the population, as it is very sensitive to extreme events. The US Interagency Advisory Committee on Water Data (USWRC) (1982), therefore, recommends weighing the local skew of a station with a regional skew. This is done to reduce the uncertainty of the skew estimate, as regional skews are assumed to be unbiased. Bulletin 17B (US Interagency Advisory Committee on Water Data (USWRC), 1982) includes a national map of the regional skew coefficients. Using weighted skews is an additional way of improving the fit of the distribution and with that, the estimation of return period flows.

5.4.3 Parameter uncertainty

Literature justifies the selection of the L-moments methods, stating that it gives better parameter estimates of the GPD compared to other methods (Hosking, 1990; Sankarasubramanian and Srinivasan, 1999; Zea Bermudez and Kotz, 2010a). Considering the poor fit to the tails of the POT

distribution, using different methods and comparing the resulting fits might lead to better estimations. GPD estimation methods have been compiled and compared by Zea Bermudez and Kotz (2010b, 2010a).

If the decision is made to use a different distribution altogether, the method used for estimating the parameters has to be changed as well. For the log-Pearson III distribution, the method of moments is recommended by the US Interagency Advisory Committee on Water Data (USWRC) (1982).

5.4.4 Data uncertainty

For the flood frequency analysis in this thesis daily mean discharge values were used. Because of this, the actual maximum instantaneous peak flow was not included in the data. Missing out on the highest peak flows meant either missing or underestimating the exceedance count of a flood stage. While the incorrect number of exceedances of a certain flood stage is not relevant for the classification, the impact on the calculated return periods and flows is. Overall, the maximum POT peaks are too low compared to the actual maximum discharge at the gauge, which leads to an overestimation of return periods and an underestimation of return period flows. A value that is never exceeded in the daily mean POT data, might have been exceeded multiple times in the actual time series. A POT peak that was assigned a return period of for example 100 years, might have a much lower statistical recurrence when using daily maximum flows for the FFA. For the classification that means, when using the POT data stations are more often classified *above*, than would be the case using daily maximum flows.

The catchment characteristics used were taken from the GAGES II database (Falcone et al., 2010; Falcone, 2017). Mean annual precipitation and runoff were given for the time period 1971 – 2000, SNOW for 1901 – 2000, and land use data was from 2006. It is likely, that the data is outdated and does not accurately represent the catchment conditions today. Changes in climate conditions have been documented, as well as increasing urbanization. However, since we are assuming stationarity of the time series, stationarity of catchment characteristics is also assumed. Additionally, the resolution of the data used to determine the catchment characteristics strongly varies, possibly leading to inaccurate representation of the catchments.

6 Conclusion

The key findings of this thesis are as follows:

- No clear pattern was found in the relationship between impact-based and statistical thresholds, as the flood stages exhibited a wide range of return periods. The return period range assigned to the different flood stages could not be confirmed.
- As shown by the at best weak correlations found, no single catchment characteristic could sufficiently explain the spatial variability of the classification.
- The classification and with that the relationship between statistical and impact-based thresholds could be modeled, the prediction accuracy laying between 61% (moderate stage) and 73% (minor stage). A more equal distribution of the data across all classification categories might improve the accuracy further.
- Of the selected parameters in the models PPTAVG, PLANT, DRAIN had an increasing and FOREST, DRAIN a decreasing effect on the indicator category.
- Runoff influencing parameters of the entire catchment were more important in modeling the relationship than flood stage height determining anthropogenic parameters of the stream adjacent area, the latter were not included in the selected models.
- Hydrologists applying statistical thresholds for flood warnings and protection measures must be aware of the discrepancy between theoretical flood thresholds based on statistical recurrence and practical thresholds based on observed impacts.

Future research needs to evaluate the stationary assumption applied in this thesis by examining the discharge time series for trends and the catchment characteristics for their actuality.

When thinking of the implications and applications of the results of this thesis, in the next step, spatial patterns of flood stage triggering flows could be examined disconnected from their return periods. Performing a regression using flood category gauge heights as the response and catchment characteristics as the predictors, to identify characteristics that have a determining influence on the flood stage. If they possess a high enough accuracy, these new models could be applied in areas where no impact-based thresholds are available, where they might result in more accurate flood thresholds than would be the case using statistical thresholds. They could also be applied in areas without sufficient data of the discharge time series available to calculate statistical thresholds, as long as reliable catchment characteristic data is obtainable.

7 References

Agresti, A., 2007. An introduction to categorical data analysis. Wiley-Interscience, Hoboken, NJ.

Akoglu, H., 2018. User's guide to correlation coefficients. Turkish journal of emergency medicine 18 (3), 91–93.

Alaska-Pacific River Forecast Center (APRFC), 2021. High Water Level Terminology. https://www.weather.gov/aprfc/terminology. Accessed October 15, 2021.

Alila, Y., Kuraś, P.K., Schnorbus, M., Hudson, R., 2009. Forests and floods: A new paradigm sheds light on age-old controversies. Water Resour. Res. 45 (8).

Anderson, B.J., 2016. COMPARISON OF ALASKAN FLOOD STAGES: RECURRENCE INTERVAL VS. IMPACT BASED. Geological Society of America.

Anderson, T.W., Darling, D.A., 1954. A Test of Goodness of Fit. Journal of the American Statistical Association 49 (268), 765.

Archer, D., 1998. Flood frequency analysis. In: R.W. Herschy (Editor), Encyclopedia of hydrology and water resources. Kluwer Academic, Dordrecht, pp. 279–288.

Archfield, S.A., Hirsch, R.M., Viglione, A., Blöschl, G., 2016. Fragmented patterns of flood change across the United States. Geophys. Res. Lett. 43 (19), 10232–10239.

Asquith, W.H., 2021. lmomco--L-moments, censored L-moments, trimmed L-moments, L-comoments, and many distributions. R package version 2.3.7. Accessed October 1, 2021.

Bačová-Mitková, V., Onderka, M., 2010. Analysis of extreme hydrological Events on the danube using the Peak Over Threshold method. Journal of Hydrology and Hydromechanics 58 (2), 88–101.

Balica, S.F., Douben, N., Wright, N.G., 2009. Flood vulnerability indices at varying spatial scales. Water science and technology : a journal of the International Association on Water Pollution Research 60 (10), 2571–2580.

Bathurst, J.C., Birkinshaw, S.J., Cisneros Espinosa, F., Iroumé, A., 2017. Forest Impact on Flood Peak Discharge and Sediment Yield in Streamflow. In: N. Sharma (Editor), River System Analysis and Management. Springer Singapore, Singapore, pp. 15–29.

Bathurst, J.C., Fahey, B., Iroumé, A., Jones, J., 2020. Forests and floods: Using field evidence to reconcile analysis methods. Hydrol. Process. 34 (15), 3295–3310.

Bauer, A., 2021. Automated Hybrid Time Series Forecasting: Design, Benchmarking, and Use Cases, Universität Würzburg.

Bayerisches Landesamt für Umwelt (LfU), Landesanstalt für Umwelt Baden-Württemberg (LUBW), 2018. Länderübergreifendes Hochwasserportal. Klassifizierung der Hochwasser-Situation am Pegel. Accessed October 15, 2021.

Bayliss, A.C., Jones, R., 1993. Peaks-over-threshold flood database : Summary statistics and seasonality. Report No. 121, Wallingford, UK.

Bezak, N., Brilly, M., Šraj, M., 2014. Comparison between the peaks-over-threshold method and the annual maximum method for flood frequency analysis. Hydrological Sciences Journal 59 (5), 959–977.

Brant, R., 1990. Assessing Proportionality in the Proportional Odds Model for Ordinal Logistic Regression. Biometrics 46 (4), 1171.

Braun, H., 1980. A simple method for testing goodness-of-fit in the presence of nuisance parameters. Journal of the Royal Statistical Society (42), 53–63.

Brunner, M.I., Seibert, J., Favre, A.-C., 2016. Bivariate return periods and their importance for flood peak and volume estimation. WIREs Water 3 (6), 819–833.

Burnham, K.P., Anderson, D.R., 2002. Model selection and multimodel inference. A practical information-theoretic approach. Springer, New York, NY.

Changnon, S.A., Pielke, R.A., Changnon, D., Sylves, R.T., Pulwarty, R., 2000. Human Factors Explain the Increased Losses from Weather and Climate Extremes. Bulletin of the American Meteorological Society 81 (3), 437–442. http://www.jstor.org/stable/26215116.

Coles, S., 2001. An Introduction to Statistical Modeling of Extreme Values. Springer London, London.

Cunnane, C., 1973. A particular comparison of annual maxima and partial duration series methods of flood frequency prediction. Journal of Hydrology 18 (3), 257–271. https://www.sciencedirect.com/science/article/pii/0022169473900516.

Cunnane, C., 1979. A note on the Poisson assumption in partial duration series models. Water Resour. Res. 15 (2), 489–494.

Cunnane, C., 1989. Statistical Distributions for Flood Frequency Analysis. World Meteorological Organization, Geneva.

Curran, J.H., Barth, N.A., Veilleux, A.G., Ourso, R.T., 2016. Estimating flood magnitude and frequency at gaged and ungaged sites on streams in Alaska and conterminous basins in Canada, based on data through water year 2012. U.S. Geological Survey Scientific Investigations Report 2016–5024.

Davenport, F.V., Herrera-Estrada, J.E., Burke, M., Diffenbaugh, N.S., 2020. Flood Size Increases Nonlinearly Across the Western United States in Response to Lower Snow-Precipitation Ratios. Water Resour. Res. 56 (1).

Dormann, C.F., 2017. Parametrische Statistik. Springer Berlin Heidelberg, Berlin, Heidelberg.

DuMouchel, W., 1983. Estimating the Stable Index alpha in Order to Measure Tail Thickness: A Critique. Annals of Statistics 11, 1019–1031.

Edwards, P.J., Watson, E.A., Wood, F., 2019. Toward a Better Understanding of Recurrence Intervals, Bankfull, and Their Importance. Journal of Contemporary Water Research & Education 166 (1), 35–45.

Falcone, J.A., 2017. U.S. Geological Survey GAGES-II time series data from consistent sources of land use, water use, agriculture, timber activities, dam removals, and other historical anthropogenic influences.

Falcone, J.A., Carlisle, D.M., Wolock, D.M., Meador, M.R., 2010. GAGES: A stream gage database for evaluating natural and altered flow conditions in the conterminous United States. Ecology 91 (2), 621.

Faraway, J., Marsaglia, G., Marsaglia, J., Baddeley, A., 2019. goftest: Classical Goodness-of-Fit Tests for Univariate Distributions. https://CRAN.R-project.org/package=goftest. Accessed October 1, 2021.

Fenneman, N.M., Johnson, D.W., 1946. Physiographic divisions of the conterminous U.S. https://water.usgs.gov/lookup/getspatial?physio. Accessed September 23, 2021.

Furey, P.R., Gupta, V.K., 2005. Effects of excess rainfall on the temporal variability of observed peak-discharge power laws. Advances in Water Resources 28 (11), 1240–1253.

Gall, M., Borden, K.A., Emrich, C.T., Cutter, S.L., 2011. The Unsustainable Trend of Natural Hazard Losses in the United States. Sustainability 3 (11), 2157–2181. https://www.mdpi.com/2071-1050/3/11/2157.

Graf, W.L., 2006. Downstream hydrologic and geomorphic effects of large dams on American rivers. Geomorphology 79 (3-4), 336–360. https://www.sciencedirect.com/science/article/pii/S0169555X06002571.

Hirsch, R.M., Archfield, S.A., 2015. Not higher but more often. Nature Clim Change 5 (3), 198–199.

Hochwasservorhersagezentrale (HVZ), 2021. HVZ- Pegelkarte. https://www.hvz.baden-wuerttemberg.de/. Accessed October 15, 2021.

Hollis, G.E., 1975. The effect of urbanization on floods of different recurrence interval. Water Resour. Res. 11 (3), 431–435.

Hosking, J.R.M., 1990. L-Moments: Analysis and Estimation of Distributions Using Linear Combinations of Order Statistics. Journal of the Royal Statistical Society. Series B (Methodological) 52 (1), 105–124. http://www.jstor.org/stable/2345653.

Hosking, J.R.M., Wallis, J.R., 1997. Regional frequency analysis. An approach based on L-moments. Cambridge Univ. Press, Cambridge.

Hounkpè, J., Diekkrüger, B., Afouda, A.A., Sintondji, L.O.C., 2019. Land use change increases flood hazard: a multi-modelling approach to assess change in flood characteristics driven by socio-economic land use change scenarios. Nat Hazards 98 (3), 1021–1050.

James, G., Witten, D., Hastie, T., Tibshirani, R., 2021. An Introduction to Statistical Learning. Springer US, New York, NY.

Jha, A.K., Bloch, R., Lamond, J. (Eds.), 2012. Cities and flooding. A guide to integrated urban flood risk management for the 21st century. World Bank, Washington, DC.

Kidson, R., Richards, K.S., 2005. Flood frequency analysis: assumptions and alternatives. Progress in Physical Geography: Earth and Environment 29 (3), 392–410.

Landesanstalt für Umwelt Baden-Württemberg (LUBW), 2021. Überflutungsflächen. https://udo.lubw.baden-wuerttemberg.de/public/pages/map/default/index.xhtml?mapId=ee35cde9-a680-4d74-953c-f39abced721f&mapSrs=EPSG%3A25832&mapExtent=401591.7431172887%2C5339213.4171711365%2C410558.12677483493%2C5343740.8510245355&overviewMapCollapsed=false. Accessed October 6, 2021.

Makowski, D., Ben-Shachar, M.S., Patil, I., Lüdecke, D., 2020. Methods and Algorithms for Correlation Analysis in R. Journal of Open Source Software 5 (51), 2306. https://joss.theoj.org/papers/10.21105/joss.02306.

Mallakpour, I., Villarini, G., 2015. The changing nature of flooding across the central United States. Nature Clim Change 5 (3), 250–254.

Mei, X., van Gelder, P.H.A.J.M., Dai, Z., Tang, Z., 2017. Impact of dams on flood occurrence of selected rivers in the United States. Front. Earth Sci. 11 (2), 268–282.

Merz, B., Aerts, J., Arnbjerg-Nielsen, K., Baldi, M., Becker, A., Bichet, A., Blöschl, G., Bouwer, L.M., Brauer, A., Cioffi, F., Delgado, J.M., Gocht, M., Guzzetti, F., Harrigan, S., Hirschboeck, K., Kilsby, C., Kron, W., Kwon, H.-H., Lall, U., Merz, R., Nissen, K., Salvatti, P., Swierczynski, T., Ulbrich, U., Viglione, A., Ward, P.J., Weiler, M., Wilhelm, B., Nied, M., 2014. Floods and climate: emerging perspectives for flood risk assessment and management. Nat. Hazards Earth Syst. Sci. 14 (7), 1921–1942.

Meylan, P., Favre, A.-C., Musy, A., 2012. Predictive Hydrology. A Frequency Analysis Approach. CRC Press, Hoboken.

Naghettini, M. (Ed.), 2017. Fundamentals of statistical hydrology. Springer, Cham.

National Weather Service (NOAA), 2019. NWS Manual 10-950. Definitions and General Terminology. https://www.nws.noaa.gov/directives/sym/pd01009050curr.pdf. Accessed October 12, 2021.

National Weather Service (NOAA), 2021. Map of US river gauges. Observation, Forecast, Inundation. https://water.weather.gov/ahps/. Accessed October 6, 2021.

NOAA National Centers for Environmental information, 2021a. Climate at a Glance: National Time Series. https://www.ncdc.noaa.gov/cag/. Accessed August 17, 2021.

NOAA National Centers for Environmental information, 2021b. Climate at a Glance: Statewide Time Series. https://www.ncdc.noaa.gov/cag/. Accessed August 17, 2021.

O'Driscoll, M., Clinton, S., Jefferson, A., Manda, A., McMillan, S., 2010. Urbanization Effects on Watershed Hydrology and In-Stream Processes in the Southern United States. Water 2 (3), 605–648.

Okoli, K., Mazzoleni, M., Breinl, K., Di Baldassarre, G., 2019. A systematic comparison of statistical and hydrological methods for design flood estimation. Hydrology Research 50 (6), 1665–1678.

R Core Team, 2019. R: A Language and Environment for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Razali, N.M., Wah, Y.B., 2011. Power comparisons of Shapiro-Wilk , Kolmogorov-Smirnov , Lilliefors and Anderson-Darling tests. Journal of Statistical Modeling and Analytics (2), 21–33.

Revelle, W., 2020. psych: Procedures for Psychological, Psychometric, and Personality Research, Evanston, Illinois. https://CRAN.R-project.org/package=psych. Accessed October 1, 2021.

Robson, A., Reed, D., 1999. Flood estimation handbook, volume 3:. statistical procedures for flood frequency estimation. Wallingford (United Kingdom).

Rogger, M., Agnoletti, M., Alaoui, A., Bathurst, J.C., Bodner, G., Borga, M., Chaplot, V., Gallart, F., Glatzel, G., Hall, J., Holden, J., Holko, L., Horn, R., Kiss, A., Kohnová, S., Leitinger, G., Lennartz, B., Parajka, J., Perdigão, R., Peth, S., Plavcová, L., Quinton, J.N., Robinson, M., Salinas, J.L., Santoro, A., Szolgay, J., Tron, S., van den Akker, J.J.H., Viglione, A., Blöschl, G., 2017. Land use change impacts on floods at the catchment scale: Challenges and opportunities for future research. Water Resour. Res. 53 (7), 5209–5219.

Saharia, M., Kirstetter, P.-E., Vergara, H., Gourley, J.J., Hong, Y., 2017. Characterization of floods in the United States. Journal of Hydrology 548, 524–535.

Sankarasubramanian, A., Srinivasan, K., 1999. Investigation and comparison of sampling properties of L-moments and conventional moments. Journal of Hydrology 218 (1-2), 13–34.

Scarrott, C., MacDonald, A., 2012. A Review of Extreme Value Threshold Estimation and Uncertainty Quantification.

Schilling, K.E., Gassman, P.W., Kling, C.L., Campbell, T., Jha, M.K., Wolter, C.F., Arnold, J.G., 2014. The potential for agricultural land use change to reduce flood risk in a large watershed. Hydrol. Process. 28 (8), 3314–3325.

Schlegel, B., Steenbergen, M., 2020. brant: Test for Parallel Regression Assumption. https://CRAN.R-project.org/package=brant. Accessed October 1, 2021.

Slater, L.J., Villarini, G., 2016. Recent trends in U.S. flood risk. Geophys. Res. Lett. 43 (24).

Soong, D.T., Murphy, E.A., Straub, T.D., 2009. Effect of detention basin release rates on flood flows - Application of a model to the Blackberry Creek Watershed in Kane County, Illinois.

Svensson, C., Kundzewicz, W.Z., Maurer, T., 2005. Trend detection in river flow series: 2. Flood and low-flow index series / Détection de tendance dans des séries de débit fluvial: 2. Séries d'indices de crue et d'étiage. Hydrological Sciences Journal 50 (5).

Tavares, L., Da Silva, J., 1983. Partial duration series method revisited. Journal of Hydrology 64 (1-4), 1–14.

Tutz, G., 2021. Ordinal regression: A review and a taxonomy of models. WIREs Comp Stat.

U.S. Census Bureau, 2018. State Area Measurements and Internal Point Coordinates. Unpublished data from the MAF/TIGER database. https://www.census.gov/geographies/reference-files/2010/geo/state-area.html. Accessed August 16, 2021.

Ulrych, T.J., Velis, D.R., Woodbury, A.D., Sacchi, M.D., 2000. L-moments and C-moments. Stochastic Environmental Research and Risk Assessment (SERRA) 14 (1), 50–68.

US Interagency Advisory Committee on Water Data (USWRC), 1982. Guidelines for determining flood flow frequency. Bulletin 17B of the Hydrology Subcommittee. US Department of the Interior, Geological Survey, Office of Water Data Coordination, Reston, VA.

USGS, 2021. What constitutes the United States? What are the official definitions? https://www.usgs.gov/faqs/what-constitutes-united-states-what-are-official-definitions?qt-news_science_products=0#qt-news_science_products. Accessed August 16, 2021.

Venables, W.N., Ripley, B.D., 2002. Modern Applied Statistics with S. Springer, New York.

Villarini, G., Serinaldi, F., Smith, J.A., Krajewski, W.F., 2009. On the stationarity of annual flood peaks in the continental United States during the 20th century. Water Resour. Res. 45 (8). https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/2008WR007645.

Villarini, G., Slater, L., 2017. Climatology of Flooding in the United States. In: G. Villarini, L. Slater (Editors), Oxford Research Encyclopedia of Natural Hazard Science. Oxford University Press.

Wickham, H., 2011. The Split-Apply-Combine Strategy for Data Analysis. Journal of Statistical Software 40 (1), 1–29. http://www.jstatsoft.org/v40/i01/.

Wolock, D.M., 2003. Hydrologic landscape regions of the United States. https://water.usgs.gov/lookup/getspatial?hlrus. Accessed.

Yan, Z., Wang, S., Ma, D., Liu, B., Lin, H., Li, S., 2019. Meteorological Factors Affecting Pan Evaporation in the Haihe River Basin, China. Water 11 (2), 317.

Ye, W., Wang, C., Xu, X., Wang, H., 2018. On seasonal and semi-annual approach for flood frequency analysis. Stochastic Environmental Research and Risk Assessment (SERRA) 32 (1), 51–62.

Yen, B.C., 2002. System and component uncertainties in water resources. In: J.J. Bogardi, Z.W. Kundzewicz (Editors), Risk, Reliability, Uncertainty, and Robustness of Water Resource Systems. Cambridge University Press, pp. 133–142.

Zea Bermudez, P. de, Kotz, S., 2010a. Parameter estimation of the generalized Pareto distribution—Part I. Journal of Statistical Planning and Inference 140 (6), 1353–1373. https://www.sciencedirect.com/science/article/pii/S0378375809002766.

Zea Bermudez, P. de, Kotz, S., 2010b. Parameter estimation of the generalized Pareto distribution—Part II. Journal of Statistical Planning and Inference 140 (6), 1374–1388.

Zeileis, A., Hothorn, T., 2002. Diagnostic Checking in Regression Relationships. R News 2 (3), 7–10. https://CRAN.R-project.org/doc/Rnews/.

Zhou, Q., Leng, G., Feng, L., 2017. Predictability of state-level flood damage in the conterminous United States: the role of hazard, exposure and vulnerability. Scientific reports 7 (1), 5354. https://www.nature.com/articles/s41598-017-05773-4.

Appendix

A.1 HLR descriptions

Table A-1: Hydrologic landscape region (HLR) descriptions (Wolock, 2003)

HLR region number	Description
1	Subhumid plains with permeable soils and bedrock
2	Humid plains with permeable soils and bedrock
3	Subhumid plains with impermeable soils and permeable bedrock
4	Humid plains with permeable soils and bedrock
5	Arid plains with permeable soils and bedrock
6	Subhumid plains with impermeable soils and bedrock
7	Humid plains with permeable soils and impermeable bedrock
8	Semiarid plains with impermeable soils and bedrock
9	Humid plateaus with impermeable soils and permeable bedrock
10	Arid plateaus with impermeable soils and permeable bedrock
11	Humid plateaus with impermeable soils and bedrock
12	Semiarid plateaus with permeable soils and impermeable bedrock
13	Semiarid plateaus with impermeable soils and bedrock
14	Arid playas with permeable soils and bedrock
15	Semiarid mountains with impermeable soils and permeable bedrock
16	Humid mountains with permeable soils and impermeable bedrock
17	Semiarid mountains with impermeable soils and bedrock
18	Semiarid mountains with permeable soils and impermeable bedrock
19	Very humid mountains with permeable soils and impermeable bedrock
20	Humid mountains with permeable soils and impermeable bedrock

A.2 Results

A.2.1 Classification of stations

Table A-2: Count of the classification combinations over all flood stages

Minor stage	Moderate stage	Major stage	Count
below	below	below	220
above	above	above	163
below	below	above	112
below	above	above	84
below	in	above	47
below	below	in	45
in	above	above	35
in	in	above	6
below	in	in	5
in	in	below	2
in	in	in	2
above	in	above	2
in	below	below	1
below	above	below	1
in	below	in	1
below	in	below	1

Table A-3: Count of classification combinations for all stages: starting from the minor stage calculating number and percentage of the following stage classification

Minor	stations	[%]	Moderate	stations	[%]	Major	stations	[%]
below	515	71%	below	377	73.2%	below	220	58.4%
						in	45	11.9%
						above	112	29.7%
			in	53	10.3%	below	1	1.9%
						in	5	9.4%
						above	47	88.7%
			above	85	16.5%	below	1	1.2%
						above	84	98.8%
in	47	6%	below	2	4.3%	below	1	50.0%
						in	1	50.0%
			in	10	21.3%	below	2	100.0%
						in	2	20.0%
						above	6	60.0%
			above	35	74.5%	above	35	100.0%
above	165	23%	In	2	1.2%	Above	2	100%
			above	163	99%	above	163	100%

Table A-4: Count of classification combinations for all stages: starting from the major stage calculating number and percentage of the previous stage classification

Major	stations	[%]	Moderate	stations	[%]	Minor	stations	[%]
below	225	30.9%	below	221	98.2%	below	220	99.5%
						in	1	0.5%
			in	3	1.3%	below	1	33.3%
						in	2	66.7%
			above	1	0.4%	below	1	100.0%
in	53	7.3%	below	46	86.8%	below	45	97.8%
						in	1	2.2%
			in	7	13.2%	below	5	10.9%
						in	2	28.6%
above	449	61.8%	below	112	24.9%	below	112	100%
			in	55	12.2%	below	47	85.5%
						in	6	10.9%
						Above	2	3.6%
			above	282	62.8%	below	84	29.8%
						in	35	12.4%
						above	163	57.8%

A.3 Discussion

A.3.1 Regression

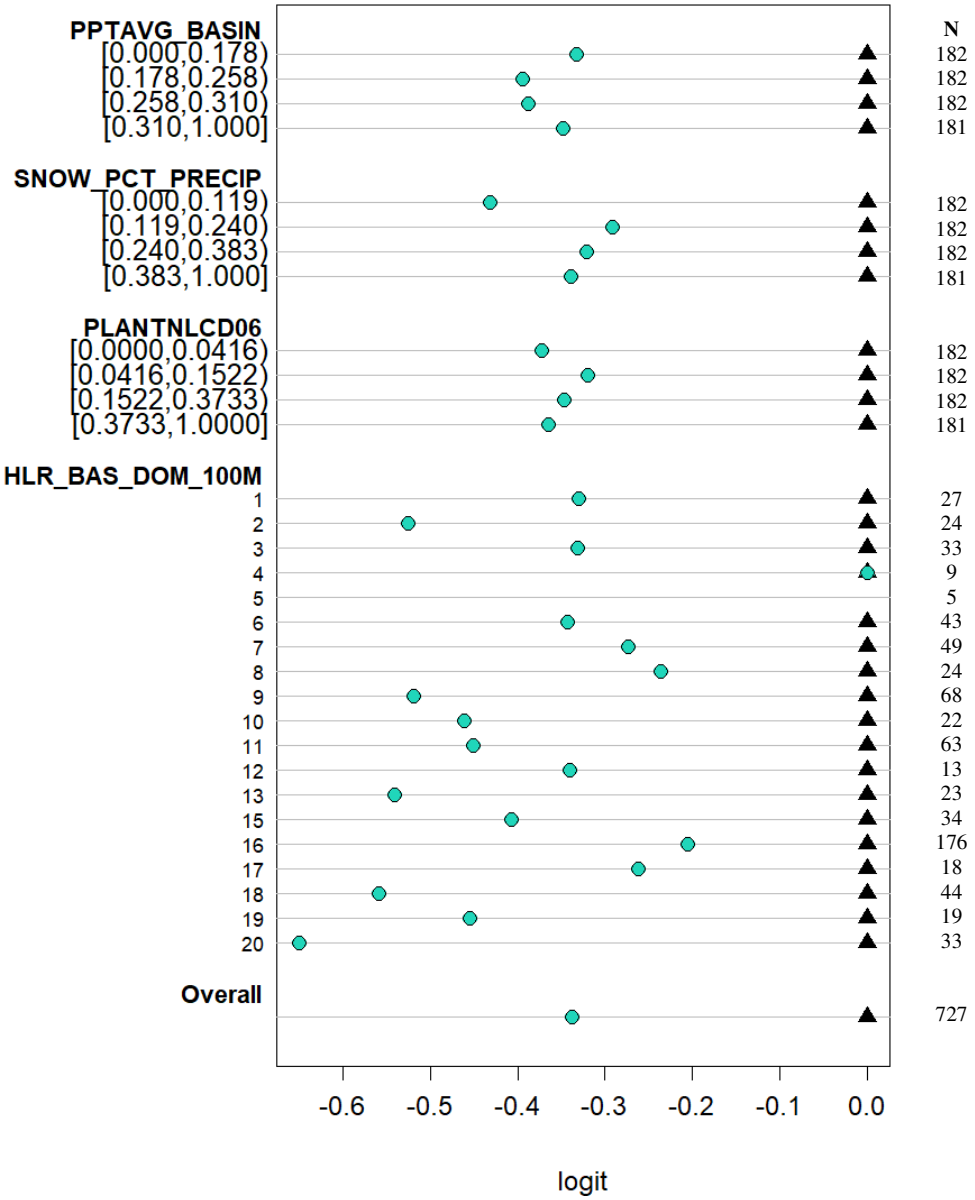


Figure A-1: Plot to test the proportional odds assumption, exemplary for the minor stage and parameters of the final minor model

A.4 Abbreviations

Table A-5: Abbreviations

Notation	Unit	Description
A	[km^2]	Area
AD – test		Anderson-Darling – test
AIC	[-]	Akaike's Information Criterion
AMF		Annual maximum flows
APRFC		Alaska-Pacific River Forecast Center
BIC	[-]	Bayesian Information Criterion
Cdf, F(x)		cumulative distribution function
CLASS	[-]	Reference/non-reference class:
CvM – test		Cramér–von Mises – test
CM		Cumulative models
CONUS		Conterminous United States
$cov(x_1x_2) = s_{x1x2}$	[-]	Covarianz
D	[-]	Kolmogorov–Smirnov statistic
DEVLP	[%]	Watershed percent "developed"
DRAIN	[km^2]	Watershed drainage area
E		Expected value
Edf , Fx, Fy		Empirical distribution function
F(.)		Continuous distribution function
FFA		Flood frequency analysis
FOREST	[%]	Watershed percent "forest"
g(y)		link function
GEV		Generalized extreme value distribution
GPD		Generalized Pareto distribution
HLR	[-]	Hydrologic landscape region
IQR		Interquartile range

k	[-]	Number of groups data is divided in for k-fold cross validation
KS – test		Kolmogorov-Smirnow – test
L	[-]	Likelihood
LENTIC	[%]	Watershed surface area covered by "Lakes/Ponds" + "Reservoirs"
Log(L) = l	[-]	Loglikelihood
Max		Maximum value
Min		Minimum value
MOP		Measure of performance
n		Number of observations
n_{fold}		Number of observations in the k-fold
NOOA		National Oceanic and Atmospheric Administration
NWS		National Weather Service
p	[-]	number of fitted parameters
P(x)	[-]	Probability
P_r	[-]	Probability of individual categories (r)
PLANT	[%]	Watershed percent "planted/cultivated"
POT	[mm/d]	Peaks over threshold
PPTAVG	[cm/year]	Mean annual precipitation
PRECIP_SEAS	[-]	Precipitation seasonality index
Pu	[-]	Non-exceedance probability
Q	[mm/d , ft^3/s]	Discharge
Q_{POT}	[mm/d]	Discharge of POT data
Q_{POT_max}	[mm/d]	Maximum discharge value of POT data
Q_{Stage}	[mm/d]	Flood stage triggering discharge
Q_T	[mm/d]	Discharge of given return period
Q_{T_lower}	[mm/d]	Discharge of T_{lower}
Q_{T_upper}	[mm/d]	Discharge of T_{upper}
r	[-]	Categories of the response

Ri	[-]	Rank of an observation
RIP_DEV	[%]	Riparian 800m buffer "developed"
RMSE	[-]	Root-mean-square error
RRMEDIAN	[-]	Elevation - relief ratio
RUNAVE	[mm/year]	Watershed annual runoff
SNOW	[%]	Snow percent of total precipitation
STREAMS	[km/km^2]	Stream density
T	[years], [a]	Return period
T_{lower}	[years], [a]	Lower end of the return period range corresponding to a flood stage
T_{POT_max}	[years], [a]	Return period of Q_{POT_max}
T_{Stage}	[years], [a]	Return period of the flood stage triggering flow
T_{upper}	[years], [a]	Upper end of the return period range corresponding to a flood stage
USGS		United States Geological Survey
WHG		Water management act
W_n^2	[-]	Anderson-Darling statistic
x	[mm/d]	Flood peak
X		Random variable, vector of observations
x(PU)		quantile function
$x_1, x_2, \ldots, x_n$		random variables
$\bar{x}_1, \bar{x}_2$		Mean of the variable
$X_{k:n}$		Kth order statistic Drawn from a distribution of X
x_i	[mm/d]	Ordered POT data
y	[-]	response of the regression model
y*	[-]	latent variable
α	[-]	significance level
β_i, η_i	[-]	regression coefficients
$\lvert\beta\rvert$	[-]	absolute β value
Θ	[-]	parameters of the distribution

λ	[-]	L-moments
μ	[years], [a]	mean time between two successive POT events
μ_l	[-]	Location parameter
ξ	[-]	Shape parameter
ρ	[-]	Spearman or Pearson's correlation
σ	[-]	Scale parameter
τ	[-]	L-moments ratios
ω^2	[-]	Cramér–von Mises criterion
ζ_k	[-]	intercept of the ordinal regression for the class boundaries

Statutory Declaration

I declare that I have authored this thesis independently and that I have not used other than the declared sources and resources.

Hiermit erkläre ich, dass die Arbeit selbständig und nur unter Verwendung der angegebenen Hilfsmittel angefertigt wurde.

Place, Date Signature