Chair of Hydrology
University of Freiburg

Lennart Schmidt

Controls of flood magnitude

A Germany-wide analysis using parametric and non-parametric approaches

- Master thesis -
Freiburg im Breisgau, June 2018

Supervisors:
Prof. Dr. M. Weiler, University of Freiburg
Prof. Dr. S. Attinger, Helmholtz-Centre for Environmental Research - UFZ, Leipzig

Contents

List of Figures

List of Tables

Abstract

With two severe flooding events in the last two decades, flood research has become a major field of hydrological science in Germany. Nation-wide studies have focused on the detection of trends, seasonal flood patterns and the influence of macro-climatic circulation patterns. However, no integrated nation-wide approach that investigates flood generation with respect to all relevant factors, especially preconditions in soil moisture, has been reported. After compiling a nation-wide dataset of 29.247 flood events at 373 gauging stations, this study applies an all-encompassing approach to identify major controls of flood magnitude by region and season. Precipitation, soil moisture and temperature of 0-7 days prior to the flooding event are examined to identify differences in the temporal regime of preconditions. It is shown that, in summer, the control of precipitation and soil moisture on flood magnitude is stronger. This is due to the higher variability of the hydrologic system, i.e. higher precipitation intensity of convective events and a greater range of wetness states in soil moisture. In catchments in the southern regions, flood magnitude is primarily controlled by mean precipitation of the 1-3 days prior to the event and topography. In flat regions towards the north, response times are higher (3-7 days) and rainfall-runoff transformation is more complex: Catchment characteristics like catchment area, drainage density and land cover have a stronger influence on flood magnitude. The applied algorithm, RandomForest, is clearly superior to a traditional Generalized Linear Model, which highlights the suitability of machine-learning approaches for the analysis of hydrological extremes.

Keywords: *Flood magnitude, Peak-over-threshold, Preconditions, RandomForest, Machine-learning, Germany*

Zusammenfassung

In Anbetracht der schweren Hochwasser in den Jahren 2002 und 2013 hat die Hochwasserforschung in Deutschland in den letzten zwei Jahrzehnten an Bedeutung gewonnen. Bisherige deutschlandweite Studien befassten sich v.a. mit Trendanalyse, der Klassifizierung von Regionen anhand ihrer Hochwasserregime sowie dem Zusammenhang von Hochwasserereignissen und großskaligen athmosphärischen Zirkulationsmustern. Es wurde jedoch bisher keine Studie veröffentlicht, die auf nationaler Ebene alle für die Hochwasserentstehung relevanten Faktoren berücksichtigt, insbesondere im Bezug auf Vorbedingungen der Bodenfeuchte. Diese Studie analysiert einen landesweiten Datensatz von 29.247 Hochwasserereignissen an 373 Pegeln. Mithilfe des RandomForest-Algorithmus wird der Einfluss klimatischer und hydrologischer Vorbedingungen sowie diverser Einzugsgebiets-Parameter auf den Hochwasserabfluss quantifiziert. Um Erkenntnisse über die zeitlichen und räumlichen Zusammenhänge zu gewinnen, werden vier verschiedene Großregionen jeweils im Sommer und Winter untersucht und die Vorbedingungen zu verschiedenen Zeitpunkten vor dem jeweiligen Ereignis erfasst. Der Zusammenhang zwischen den klimatischen und hydrologischen Vorbedingungen und dem Hochwasserabfluss ist im Sommer ausgeprägter als im Winter, was auf eine höhere Variabilität des hydrologischen Systems im Sommer zurückzuführen ist: Konvektive Niederschlagsereignisse und ein ausgeprägteres Regime der Bodenfeuchte schlagen sich direkt in der Stärke des Hochwassers nieder. In den Einzugsgebieten (EZGs) im Süden wird der Hochwasserabfluss maßgeblich vom mittleren Niederschlag der 1-3 dem Ereignis vorausgehenden Tage sowie von topographischen Faktoren bestimmt. Im Norden hingegen ist die Topographie weniger ausgeprägt und die Abflussbildung komplexer, so dass mehrere andere EZG-Parameter einen Einfluss auf den Hochwasserabfluss haben, unter anderem die EZG-Fläche, Bodenbedeckung sowie die Einzugsdichte. Hier ist es der mittlere Niederschlag der vorausgehenden 3-7 Tage, der den Hochwasserabfluss maßgeblich beeinflusst. Im Vergleich mit einem klassischen Generalisierten Linearen Modell ist die Genauigkeit des RandomForest-Algorithmus weitaus hher, was die Eignung von Machine-Learning-Algorithmen zur Analyse von hydrologischen Extremereignissen verdeutlicht.

Stichworte: *Hochwasserabfluss, Vorbedingungen, RandomForest, Machine-Learning, Deutschland*

Chapter 1

Introduction

Mankind has always been threatened by natural hazards of various kinds – Among these, fluvial floods. They can pose a significant, large-scale hazard to human life and property. In Germany, floods are the natural hazard that causes the greatest economic damage and several large flooding events in the last two decades have directed public interest towards flood research. Several studies have been published that analyzed trends and seasonal patterns of flood occurrence at a national scale. Also, the influence of macro-climatic circulation patterns on flood generation has been investigated for the whole of Germany. However, no study has yet investigated flood generation with respect to all relevant physiographic factors and preconditions in soil moisture at that large a scale. This master thesis aims at filling that gap by compiling and analyzing a dataset of about 29.000 flood events as to the influence of hydro-climatic preconditions and multiple physiographic factors. The investigation is carried out by region and season to investigate spatial and temporal patterns. Both parametric and non-parametric modeling techniques are applied and their performance is compared to evaluate non-parametric algorithms as tools for the analysis of hydrological extremes.
This thesis is structured in the following way: The first chapter provides an overview of the current state of flood research with a focus on Germany and concludes with the principal research objectives of the study at hand. The following chapter illustrates the methods that are employed to compile the flood dataset and presents the theoretical background of the modeling approach that is applied. The results that are obtained by modeling are first presented and then discussed with respect to the principal research objectives. The final chapter closes with an overview of the main findings and potential future extensions of this study.
Throughout this work, "floods" refer to fluvial floods, only. The terms "dynamic" and "static" are used in the sense of time-variant and time-invariant, respectively.

1.1 Research to date

1.1.1 Theory of Flood Generation

This section serves to outline the current conceptual knowledge on the processes that play a role in generation of fluvial floods, following Patt and Jüpner (2001) and Maniak (2013). Generally, flood generation depends on three processes: Runoff generation, runoff concentration and flood routing. The component of baseflow is commonly disregarded as it is known to make up less than 10% of peak flow.

- **Runoff generation:**
  This is the process that transforms event precipitation into effective precipitation, i.e. the share of precipitation that enters the stream as direct runoff. Direct runoff is defined as surface runoff and interflow in the soil zone. The ratio of precipitation and direct runoff is determined by the infiltration and retention characteristics of the catchment. This, in turn, is influenced by both **dynamic and static** factors. The dynamic ones are amount and intensity of precipitation and preconditions in soil moisture and snow cover. The static ones are vegetation, land cover, topography, soil and geological properties and groundwater levels. In theory, there are two flood types that are linked to area-dependent runoff generation mechanisms: In small catchments $<100\text{km}^2$, strong precipitation events, that are limited in spatial extent but cover most of the catchment area, lead to "Hortonian surface runoff": As precipitation intensity exceeds infiltration capacity of the soils, large quantities of water reach the stream in short time via surface runoff, leading to (flash) floods. In large catchments, these local extreme events do not show in the streamflow response at the outlet, as

they only cover a limited fraction of the catchment. In these larger catchments, runoff generation is linked to long, enduring rainfall, possibly in combination with snow melt or wet preconditions. At some point, soils are saturated so that interflow is activated and infiltration rates have become low, so that "saturation excess surface runoff" is generated.

- **Runoff concentration:**
  In each catchment, the response of streamflow to direct runoff is different due to its **static** characteristics. Depending on the shape and area of the catchment as well as its topography, roughness and soil properties, different regions within the catchment drain at different speed - Thus, shaping the flood wave. For details, readers are referred to the model of an impervious, tilted surface as presented in Maniak (2013).

- **Flood Routing:**
  Once effective precipitation has reached the stream as direct runoff, the shape of the flood wave changes on its way downstream. This behavior is mainly influenced by **static** factors such as slope, roughness and length of the main channels, incoming tributaries and the morphology of the channel and its banks – including anthropogenic changes to the latter. The presence of lakes in the catchment is an important factor as well.

1.1.2 Typology of Floods

Merz and Blöschl (2003) analyzed an extensive dataset of about 11.500 flood events in 490 Austrian catchments in order to classify floods into 5 major types: Long-Rain, Short-Rain, Flash, Rain-on-Snow and Snowmelt floods. The classification was based on timing of the floods, storm duration, rainfall depths, snowmelt, catchment state, runoff response dynamics and spatial coherence. Table 1.1 presents these flood types and their respective characteristics. This typology has since been accepted as common terminology in flood research in Germany and Austria (e.g. Freudiger et al., 2014; Nied et al., 2014). Consequently, this terminology will also be applied in the study at hand.

Table 1.1: Flood typology and the respective flood characteristics (source: Merz and Blöschl, 2003).

Process Type	Long-Rain Floods	Short-Rain Floods	Flash Floods	Rain-on-Snow Floods	Snowmelt Floods
Timing of floods	no pronounced seasonality	no pronounced seasonality	floods and extreme rainfall mainly in summer or late summer	mainly occur at the change between cold and warm periods	floods in spring to summer
Storm duration	long duration (>1-day)	duration of several hours to 1-day	short duration (<90 min), high intensities	moderate rainfall events can cause large floods	rainfall unimportant
Rainfall depths, snowmelt	substantial rainfall depths	moderate to substantial rainfall	small to moderate rainfall depths	snowmelt and rainfall	snowmelt, no or minor rainfall
Catchment state (SWE, soil moisture)	wet due to persistent rainfall	wet for large flood events	dry or wet	wet, snow covered	wet, snow covered
Runoff response dynamics	slow response	fast response	flashy response	fast or slow response	medium or slow response
Spatial coherence	large spatial extent of storms and floods (>10^4 km^2)	local or regional extent	limited spatial extent of storms and floods (<30 km^2)	limited to areas of snow cover	medium spatial extent of floods

1.1.3 Flood Research

With two severe flooding events in 2002 and 2013, flood research has received increased public attention in Germany over the last two decades. Thus, multiple studies have been published. In order to place this study in the context of flood research in Germany, the following section provides an overview of the current state of research:
As to the definition of what is considered a flooding event, there are two major approaches: The traditional and most-commonly used one is the extraction of the Maximum Annual Flood (MAF), i.e. the highest streamflow peak of each year. While the approach is straightforward in implementation, it is known to introduce a loss of information as several high floods might occur in one year (IHUK, 1999; Lang et al., 1999). Also, vice versa, the MAF of one year can be lower than multiple floods of another year - In that case, it would not accurately represent the probability distribution of flood magnitudes. To account for this, the Peak-Over-Threshold method (POT), also called Partial Duration Series, was introduced: All streamflow values above a certain threshold are collected. This introduces more flexibility

as any number of events per year can be captured but it also introduces further analytical complexity: If not based on expert knowledge of the respective catchment, determination of the true threshold is not evident. Also, for any further analysis, independence of flood events has to be ensured. For both issues, various approaches have been proposed. These were summarized by Lang et al. (1999), who made a major effort in setting up guidelines for POT-estimation.
Regarding the analysis of the resulting flood event dataset, studies vary as to the flood characteristic that is examined: Major focuses of research are flood frequency and magnitude as such, but duration and the shape of frequency or magnitude distributions are investigated as well. With regard to climatic changes in past and future, trend analysis has been a field that has received considerable public interest. For Germany, Petrow and Merz (2009) investigated trends in flood frequency and magnitude in time series of 145 catchments throughout Germany for the time period of 1951-2002. At macro-scale, no clear trends could be detected as these proved to be spatially clustered: A third of all catchments in south-western Germany exhibited an increase in flood magnitude of annual maximum floods. In central Germany, an increase was detected only for maximum winter floods. Also in winter, flood frequency increased at about a third of all catchments. A follow-up study by Petrow et al. (2009) related these trends to changes in macro-climatic circulation patterns: Trends in MAF were linked to the dynamics of flood-inducing circulation patterns. For the latter, a significant increase in frequency and persistence was detected, especially in winter.
As many studies analyze multiple flood characteristics, the following section is grouped by factors that were identified to influence flood generation in general. First, research on dynamic, i.e. time-variant, factors is presented. This if followed by an overview on the state of research regarding static factors, i.e. time-invariant catchment characteristics.
Dynamic factors:
A pronounced link between macro-climatic patterns and floods was also observed by Samaniego and Bárdossy (2007), who applied a non-linear generalized model to 46 catchments of the Neckar and found significant links between macro-climatic patterns and flood frequency/duration, also especially in winter season. The Vb-weather patterns have been linked to high flooding potential in Eastern Germany, e.g. for Elbe and Oder catchments Petrow et al. (2007); Beurton and Thieken (2009). Bárdossy and Filiz (2005) successfully derived flood-inducing circulation patterns from positive increments of discharge time series. Nied et al. (2014) was able to link the flood types as presented in the previous chapter to specific climatic patterns.
Linking macro-climatic patterns to flood events proved to be successful as these carry information of precipitation frequency, intensity and duration at the same time. Thus, they imply information of the wetness state of the catchment prior to a precipitation event. The above flood typology indicates the importance of catchment preconditions in soil moisture and snow-water-estimate (SWE) for flood generation. Therefore, a separate investigation of precipitation, soil moisture and SWE is necessary for further process understanding that could help in prediction of floods (Nied et al., 2013):
Schröter et al. (2015) found out that the flood severity, i.e. the combination of both magnitude and spatial extent, of the nation-wide flood in June 2013 was a result of exceptionally wet preconditions on a large scale. As a result of these, precipitation of moderate strength led to exceptionally high streamflows. In a small catchment in the Ardeche region, Huza et al. (2014) illustrated a similar influence of soil moisture on rainfall-runoff processes by comparison of storm-hydrographs. At mesoscale, Nied et al. (2014) clustered antecedent soil moisture patterns prior to floods in the Elbe catchment and detected a seasonal effect: In winter, occurrence of large-scale floods showed to be more likely at high soil moisture. In summer, the flood-inducing soil moisture patterns were more variable, ranging from dry to wet. Also, the spatial distribution proved to be relevant. The latter has also been shown by Merz and Plate (1997), who proved that spatial variability has a major effect on streamflow response, especially for events of medium magnitude. Also, it is commonly agreed that the influence of soil moisture decreases with flood magnitude (Gutknecht et al., 2002; Merz and Blöschl, 2009*a*; Nied et al., 2017). Other studies focused on the effect of soil moisture on runoff coefficients, i.e. the ratio of direct runoff to precipitation – which in turn represents a major control of flood generation: Merz and Blöschl (2009*b*) showed that catchment moisture conditions control runoff coefficients to a higher degree than does event rainfall. Penna et al. (2011) highlighted the influence of soil moisture on catchment response time via altered runoff coefficients. Soil moisture is also implemented in extreme-flood forecasting systems like the SCHADEX (Paquet et al., 2013). Still, research at large scale is limited due to low data coverage. A way to surpass this limitation is to use modeled soil moisture, as the study at hand does.
Another important variable of catchment pre-conditions is the presence of snow in the catchment. Snow melt influences soil moisture and can cause flood events (Merz and Blöschl, 2003). Despite multiple successful efforts on small scale, modeling of snow processes at a larger scale remains a challenging task due to a high spatial variability and complexity of the processes involved (Blöschl et al., 1990; Rössler et al.,

2014). A common approach is the degree-day-method, introduced by Hargreaves and Samani (1982), and applied in several hydrological modeling frameworks (e.g. Kumar et al., 2013; He et al., 2014). As an extreme, Rain-on-Snow-Events (RoS) have led to floods of strong magnitude (Merz and Blöschl, 2003; Sui and Koehler, 2001). As occurrence is relatively rare and difficult to capture, RoS are the target of intensive research as to the frequency of occurrence in time and space (e.g Garvelmann et al., 2013). Freudiger et al. (2014) simulated RoS-events from historical data for all major river basins in Germany and concluded that basins in medium-elevation mountain ranges are most prone to RoS-events. At these altitudes, highest RoS-flood hazard was identified for the early winter period.

Static factors:

As mentioned in the previous section, static variables have an influence on flood generation through runoff generation and concentration. As to their static nature, they only influence average flood characteristics of a catchment. Bronstert et al. (2017) state that the influence of static variables becomes harder to detect with increasing mean MAF.

Catchment area has been reported to be a major driver of flood characteristics. Merz and Blöschl (2009a) analyzed mean MAF of 459 and found catchment area to be the dominant static variable, showing a negative correlation. A similar relation was identified by Pfaundler (2001), who ran a stepwise multiple regression model on 231 swiss catchments and Uhlenbrook et al. (2002) with a similar approach on catchments in southern Germany.

There is evidence that land cover types influence flood generation: Samaniego-Eguiguren (2003) found a correlation of increasing winter flood frequencies with increasing proportion of permeable land and decreasing proportion of forest. These results are in line with Uhlenbrook et al. (2002), who found similar functional relationships between flood magnitude and the above land cover classes. Kuraś et al. (2012) detected an increase of flood magnitude and frequency as a result of large-scale deforestation, that proved to be stronger at higher return periods. Above studies analyzed catchments at the meso-scale. In a comprehensive commentary, Blöschl et al. (2007) hypothesizes that the effect of land cover on flood generation is scale-dependent, i.e. it decreases with increasing catchment size.

Of soil characteristics, measures of infiltration capacity have been identified as influential in several studies (e.g Castellarin et al., 2001; Pfaundler, 2001). Notably, the changes of infiltration rates at different hillslope conditions have recently been subject to scientific debate as studies have led to contradictory results (Morbidelli et al., 2018).

Still, there is consensus of a positive correlation of slope and flood magnitude (Maniak, 2013). A study by Chifflard (2006) gave clear results: At slopes of $>6°$, runoff generation is governed by slope instead of soil moisture as result of higher flow velocities. Regarding morphologic properties of a catchment, drainage density is known to have a significant effect on flood magnitude in form of a positive correlation (Wharton, 1994). Thus, research focus on drainage density has evolved towards identification of possible threshold values of drainage density that allow for classification of ungauged basins as to their variability in flood magnitude (Pallard et al., 2009). Related to this is the influence of catchment shape on flood magnitude: It is known to be present in form of a negative correlation and is often included in flood frequency analysis of ungauged basins using the "shape factor", R (Dyck and Peschke, 1995; Zrinji and Burn, 1994).

Large-scale analyses:

The studies that were presented above vary in scale and as to the number of flood-relevant factors included. To our knowledge, there are only few studies that both analyzed flood generation on a large scale and included a set of factors that can be regarded as representative of all major controls that are relevant for flood generation. Nied et al. (2017) took a significant step towards an all-encompassing approach of flood generation analysis: After identifying flood-relevant soil moisture patterns and flood-inducing macro-climatic patterns in two studies, a complete flood risk modeling chain for the Elbe catchment was introduced (Nied et al., 2013, 2014, 2017): It combines the above information with hydrological and hydraulic modeling to assess the effect of changes in hydro-climatic conditions on various flood characteristics. With respect to flood magnitude, the major findings were that the number of gauges in the catchment undergoing at least a 10-year flood is governed by weather patterns. Floods of lower magnitude, i.e. of a 2-year return period, proved to be controlled by soil moisture patterns. This study is unique in the sense that it included all relevant dynamic and static variables and was applied on a large scale. All other studies that were found were either limited in spatial scale or in number of relevant variables. Uhlenbrook et al. (2002) did include multiple physiographic variables and an estimate of catchment wetness but was limited to South Germany and did not include any snow process estimates. Samaniego and Bárdossy (2007) set up a statistical model that included all relevant models but the analysis was limited to the Neckar catchment and was more of a methodological nature. Merz and Blöschl (2009a) analyzed the control of all relevant hydro-climatological preconditions and static factors on mean MAF – Thus, MAF was reduced to its static component. Also, the study region was the whole of Austria, not

Germany.
On national scale, there are four studies available that analyzed flood events. As presented, Petrow and Merz (2009) performed a Germany-wide trend analysis that could be linked to changes in macro-climatic patterns by Petrow et al. (2009). Uhlemann et al. (2010) compiled a dataset of 80 trans-basin floods, i.e. floods that affect a multitude of basins at the same time. 64% of these occurred in winter, including the ones of highest severity. Beurton and Thieken (2009) classified floods in German basins into three regions by seasonality: A central-western region of precipitation-driven winter floods that are influenced by the Atlantic Ocean. Towards the east, spring and summer floods occur more frequently – these were linked to Western- and Vb-circulation patterns. In the south, a multi-modal regime was identified with both winter floods due to snow melt and summer floods due to cyclonic weather patterns. In conclusion, it can be said that the influence of macro-climatic patterns on flood generation and the respective seasonal changes have been studied sufficiently. However, no study at national scale has yet taken apart the influence of macro-climatic patterns into the component of precipitation and catchment state. Also, theoretical knowledge of the influence of static variables has not been verified on that large a scale.
This study aims at filling the research gaps as presented above by applying an "all-encompassing" approach, i.e. including the majority of dynamic and static factors, to the whole of Germany. Preconditions in soil moisture and snow cover are accounted for and functional relationships of static factors are investigated to verify common hydrological knowledge.

1.1.4 Machine-Learning Algorithms

Both algorithms that are applied in this study, RF and GLM, are well-established in the scientific community. However, not both to the same extent in the field of hydrology. GLM needs little comment, as an extension of the ordinary linear regression to non-normal error distributions, it remains the standard parametric approach for analysis of covariate influence (Hastie et al., 2009). It has been applied to a multitude of hydrological problems, see Naghettini (2016) for details.
With emerging computational power, non-parametric machine-learning algorithms (ML), also referred to as data-driven models, gained popularity in hydrological sciences at the start of the millennium. This was mainly due to the fact that these algorithms can handle large datasets and have proven to give reliable estimates even if the data is subject to noise and measurement errors, as is often the case in hydrological modeling (Solomatine and Ostfeld, 2008). In contrast to physical models, ML-techniques do not require prior physical knowledge. Unlike regression techniques, no assumptions are necessary regarding the functional relationship of input and output or error term distribution. In addition, non-linear relationships as well as interactions between predictors are implemented automatically (Solomatine and Ostfeld, 2008). The primary prerequisite of ML-techniques is the availability of a substantial amount of data.
One of these is the the RandomForest algorithm (RF). It was introduced by Breiman (2001) and has since received considerable attention, especially in remote sensing and ecological sciences. While procedures like cross-validation may be applied for finding the optimal combination of RF parameters, RF is known to produce robust results at standard parameter settings with low tendency to overfit to the training data. This is due to the substantial amount of randomization during model calibration (Hastie et al., 2009).
In the past five years, several applications of RF in the field of hydrology have been reported. Among these were prediction of streamflow (e.g. Shortridge et al., 2016), flow duration curve estimation at ungauged sites (Booker and Snelder, 2012), streamflow classification (Peñas et al., 2014), -regionalization (Gudmundsson and Seneviratne, 2015), and forecasting of urban water demand (Herrera et al., 2010).
Several studies have performed a comparative evaluation of the performance of multiple MLs. Shortridge et al. (2016) applied five MLs and GLM for streamflow prediction in a seasonal watershed in Ethiopia. When predicting streamflow directly, RF reached highest accuracy in 4 out of 5 sub-catchments and successfully detected non-linear runoff behavior of wetlands. When predicting streamflow anomalies, i.e. normalized to monthly longterm averages, GLM produced good results, too. Lima et al. (2015) investigated the performance of multiple MLs on 9 different environmental datasets, including three hydrological ones. All MLs proved to be suitable for prediction and of higher accuracy than GLM, RF was shown to be fastest in calibration on large datasets. Booker and Snelder (2012) applied both stepwise linear regression and RF for estimation of flood duration curves in ungauged basins. RF proved to be more accurate but the study revealed a major drawback of the RF method (and most other MLs): The domain of predictions is limited to the one of calibration data. Thus, if the response is not naturally constrained in domain, RF predictions might be inaccurate for extreme values.
Despite intensive research, few studies were found that applied RF for flood research, none for investigation of flood magnitude. Wang et al. (2015) used both RF and Support-Vector-Machines (SVM) for mapping of flood risk in the Dongjiang River Basin, China. Based on historical flood records and data

on several physiographic variables, flood risk maps were produced with an average classification error rate of 8.76%. RF and SVM were similar in accuracy but the authors highlighted RF's ease of operation compared to SVM. Similar probabilistic classification approaches were applied by several authors, e.g. Lee et al. (2017) and Terti et al. (2017). Regression-type models that predict flood frequency were set up by Zhao et al. (2018) and Sadler et al. (2018). The majority of studies as listed above profited from two major features of RF: The calculation of variable importance, i.e. the explanatory power of a predictor as assigned by the model, and Partial Dependence Plots (PDP) – These show the functional relationship of the response variable to the respective predictor as it was mapped by RF. The combination allows for interpretation of model structure similar to GLM but extended to non-linear relationships. Therefore, RF combines the benefits of MLs with tools that allow for causal investigation of the respective system. As hydrological research has only recently discovered MLs like the RF, the study at hand represents a novel approach insofar as ML-techniques are applied for flood research, more specifically for causal research on flood magnitude. The aim is to investigate the suitability of RF for analysis of hydrological extremes by comparing its performance to an old-but-gold approach, the GLM.

1.2 Research Objectives

Principal research objectives of the proposed study are:

- **Controls of flood magnitude**
  Identification of factors that control flood magnitude on national scale. In order to gain insight on the influence of pre-event conditions, the dynamic factors precipitation, soil moisture and temperature are sampled at different time periods prior to a flood event. Functional relationships are examined and contrasted with literature.

- **Analysis by region and season**
  The analysis is run on distinct regions of Germany that are known to exhibit different physiographic conditions. This spatially-explicit analysis allows for an examination of the respective mechanisms that control flood magnitude in each region, separately. Also, winter and summer floods are investigated separately to investigate different flood generation mechanisms in each season.

- **Comparative investigation of model performance**
  Two algorithms are applied, one parametric and one non-parametric: GLM and RF. First, model performance is evaluated and model results are validated against each other. Second, suitability of the approaches as tools for analysis of hydrological extremes is compared as to their performance in different environments.

Chapter 2

Methodology

This chapter outlines the methodology of this study. The principal aim of study was to identify environmental factors that control flood magnitude in Germany. Thus, flood events were sampled from streamflow timeseries of 374 gauging stations across Germany (ch. 2.1.2). As the influence of hydro-climatic preconditions was to be assessed, these were sampled from timeseries of precipitation, soil moisture and temperature for each catchment, separately (ch. 2.1.5). The derived preconditions and a set of catchment attributes were then used as inputs, i.e. predictors, for two statistical models, RF and GLM (ch. 2.6). The aim of this approach is to simulate the observed flood magnitudes at each of the 374 gauging stations, taking into account the respective preconditions and catchment attributes. This enables the interpretation of the resulting models as to the explanatory power of each of the inputs. A predictor that has high explanatory power over simulated flood magnitude is assumed to exert strong control over flood magnitude in reality. By calibrating separate models by region and season (chs. 2.2, 2.3), patterns of controls of flood magnitude in time and space can be derived. As this chapter describes the modeling process, flood-relevant factors that are included in the models are referred to as predictors or variables. In interpretation, i.e. when moving from model to reality, the term factors will be applied again.
First, the dataset is presented. This is done in the following order: After giving some general information on the dataset, the sampling scheme to derive flood magnitude Q_f is illustrated. Next, all predictors that were used in the analysis are listed. Also, the sampling and geoprocessing operations that were carried out to derive them are specified. Following this, the process of sub-setting the data by region and season is presented. Following this, the modeling approach is explained. It is split up into model calibration, validation and interpretation. For each of these, the respective methods are presented.
All data processing is performed in R (R Core Team, 2018). Packages that are not included in the base version of R are cited as such.

2.1 Dataset

2.1.1 General Information

The dataset consists of time-series of streamflow, Q at daily resolution at 374 gauging stations that are spread across Germany. For each of the corresponding catchments, similar time-series of the dynamic predictors precipitation, effective precipitation, soil moisture and air temperature are included. Also, data regarding 10 catchment attributes of various types, i.e. static predictors, is available for each of the catchments. As displayed in figure 2.1, the corresponding catchments cover all of Germany except the south-west and larger parts in the north. Catchment size varies between 100 and 8469 km^2 with the majority (85%) below 1000 km^2. Mean catchment size is 739 km^2. The time period of reference is 1950-2010 with a mean coverage of all stations of 65%. Here, "coverage" refers to the share of complete records both in streamflow and predictor records. As can be seen in figure 2.2, the majority (75%) of all stations have a coverage of 50% or more. Both Q and all predictors that were supported in volumetric units were area-adjusted, i.e. converted to mm.

2.1.2 Sampling of Flood Magnitude

As mentioned in chapter 1.1, there are different ways to define a flood. Next to the use of maximum annual streamflow, MAF, all streamflows that exceed a defined threshold value can be considered a flood event. This threshold value can be determined either by an expert or by using percentiles to separate

Figure 2.1: Map of catchments that were included in the study.

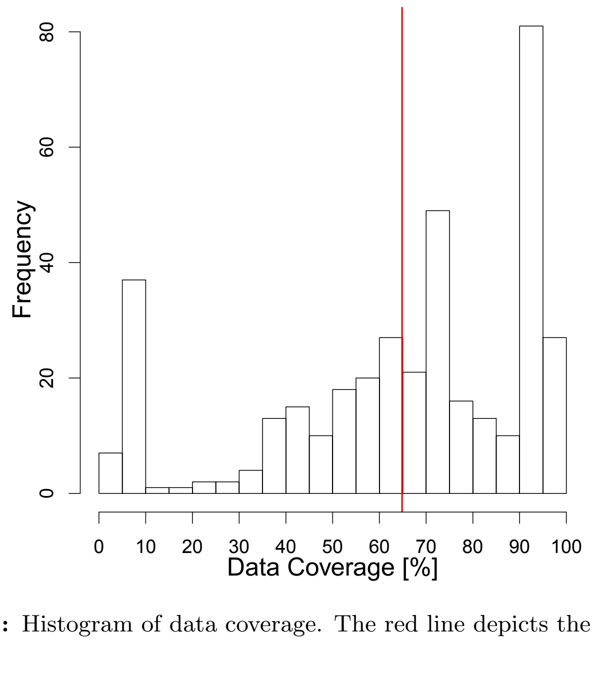


Figure 2.2: Histogram of data coverage. The red line depicts the mean of 65%.

exceptionally high values (Lang et al., 1999). The latter was applied in this study. By applying a percentile threshold on each of the 374 catchments separately, this study uses the term flood in a relative sense: No matter how large the difference is between average streamflow conditions and high flow events, a period of exceptionally high subsequent streamflows is considered a flood event. This means that no difference is made between floods in different catchments, even though the shape of the hydrograph might be distinctly different. In order to account for the differences in shape, the dataset was divided into 4 regions of distinct physiographic characteristics (see chapter 2.2). Figure 2.3 illustrates how all values of Q, that exceed the threshold-value of a X%-percentile were grouped into one flood event. Of this event,

the flood magnitude, i.e. the maximum value, Q_f was extracted. Table 2.1 displays unit, symbol and source of both Q and Q_f.

Table 2.1: Unit, symbol and source of streamflow Q and flood magnitude Q_f.

Name	Unit	Symbol	Source
Flood Magnitude	mm	Q_f	derived from Q by POT-analysis
Streamflow	mm	Q	GRDC (2010), EWA (2010)

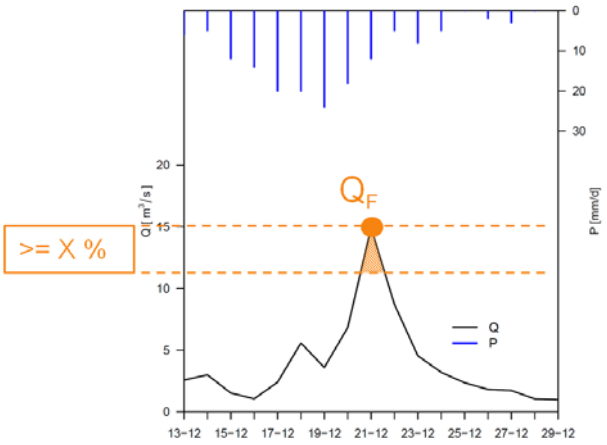


Figure 2.3: Sampling procedure to derive Q_f from daily streamflow Q. The black line depicts daily mean streamflow Q, blue bars depict daily mean precipitation, P. All subsequent daily streamflows (here: 3) that exceed the X%-percentile threshold were grouped as one flood event (orange polygon) and the flood magnitude, Q_f, was extracted.

Following the approach used by Samaniego-Eguiguren (2003), the 95%-quantile was used as a base reference. Bezak et al. (2014) advises to try different threshold values as the right threshold depends on the respective dataset. So, for a more rigid separation, 97% and 98%-percentiles were applied as well. The choice of which percentile to use was based on two criteria: First, the plausibility of event durations. This should not exceed 20 days which was considered to be the longest physically plausible flood. More importantly, a visual analysis of the time-series was carried out to examine which threshold separates strong flood events best. This means that secondary flood waves both before and after the main event as well as multiple peaks in direct succession should best not be included. Depending on the duration of the resulting floods, the number of flood events at each station varies between gauging stations.
In statistics, a major requirement to any dataset is that its samples are independent of each other, see e.g. Hastie et al. (2009) or Wood (2006). In order to ensure this, flood events were filtered once more as to the difference in time between each two subsequent events. Kundzewicz et al. (2005) suggest a time distance of 5 days in between two events for catchments <45.000km^2 to ensure independence. Here, this approach was extended to 7 days, as this is the time period of predictor sampling. If there were several events with less than 7 days in between them, the one with the highest Q_f was kept.

2.1.3 Predictors

The predictors that are included in the models are listed in table 2.2, along with their respective unit, the symbol that will be used throughout this study and the data source. Four of these are dynamic, i.e. time-variant. Here, they are present as daily mean values over the respective catchment area. The others are static, i.e. time-invariant, as they consist of only one value per catchment over the whole time period. The predictors with the source marked as "mHm" are model outputs of the "Mesoscale Hydrologic Model" that was developed by the Department of Computational Hydrosystems at the Helmholtz Centre for Environmental Research in Leipzig, Germany. For more information on the simulation of these the reader is referred to Samaniego et al. (2010) and Kumar et al. (2013). The mHm-model was run on the whole of Germany and data for each of the 374 catchments was extracted for this study.

Table 2.2: Unit, symbol and source of predictors that were included in the models of this study.

Type	Name	Unit	Symbol	Source
Dynamic	Precipitation	mm	P	DWD (2015)/mHm
	Effective Precipitation	mm	P_{eff}	mHm
	Soil Moisture	fraction	SM	mHm
	Air Temperature	°C	T	DWD (2015)/mHm
Static	Aridity Index	fraction	AI	derived from data
	Mean Annual Precipitation	mm	P_{ann}	derived from data
	Catchment Area	km^2	$Area$	mHm, see Zink et al. (2017)
	Land Cover Classes	%	$Forest$, $Impermeable$...	CORINE (EEA, 2010)
	Mean Altitude	m	$Altitude$	derived from DEM (see 2.1.4)
	Mean Catchment Slope	°	$Slope$	
	Mean Channel Slope	°	$ChSlope$	
	Maximum Flow Path Length	km	$FLMax$	
	CV of Flow Path Length	-	$FLCV$	
	Drainage Density	fraction	DD	

Choice of predictors was based on common knowledge of the factors that affect runoff generation and concentration (see previous chapter). Of these, as many as available were included. However, the process of flood routing is accounted for at a basic level only as to limitations in data availability and scope of this study. For some of the predictors, additional information is needed:

- **Effective Precipitation P_{eff}:** Normally, effective precipitation is defined as the component of rainfall that reaches the stream as direct runoff (Patt and Jüpner, 2001). Here, however, the term describes the combination of precipitation, snow accumulation and -melt as modeled by mHm from observed precipitation. Snow melt and accumulation were modeled using a modified version of the degree-day method by Linsley (1943). Details of the snow modeling routine can be found in Samaniego et al. (2010).

- **Soil Moisture SM:** The soil moisture with respect to porosity, i.e. the fraction of saturated water content, averaged over the whole soil column. Soil column depth was derived from soil maps, maximum soil depth was 1.8m. For more details, the reader is referred to Samaniego et al. (2013).

- **Air temperature** T: Temperature, too, is a model estimate by mHm, averaged over each of the catchment. The principal motive to include T in the analysis was to account for runoff generation and concentration processes on frozen soils, as these are not accounted for in mHm. The effect of T on snow melt and accumulation is already included in P_eff. The effect of evapotranspiration was assumed to be negligible for the present flood modeling approach due to the short time period that is evaluated. From here, T will simply be referred to as "temperature".

- **Aridity Index:** An estimate of the average "dryness" or ,vice versa, "wetness" of a location. Following Middleton and Thomas (1992), it was calculated from average annual observed precipitation, P_{ANN}, and average annual potential evapotranspiration PET_{ann}:

$$AI = \frac{P_{ann}}{PET_{ann}} \tag{2.1}$$

  PET_{ann} was estimated in mHm, using the Hargreaves and Samani method (Hargreaves and Samani, 1982). See Kumar (2010) for details. AI has proven to be a reliable estimate of average catchment conditions (Blöschl et al., 2013). In this study, it was included to serve as a control variable for average catchment wetness: The preconditions of P_{eff} and SM as well as Q_f contain both a static and dynamic component. They all vary in average values between different catchments, but in each catchment they are time-variant. One of the principal objectives of this study was to quantify the influence of preconditions on flood magnitude, i.e. the link between the dynamic components of both Q_f and preconditions in each of the individual catchments. To make sure that any explanatory power that is assigned to preconditions by the models only refers to its dynamic component, AI was included to account for the static component of preconditions, i.e. average wetness. Annual precipitation, P_{ann} was included as well to compare which one performs better as a control variable.

2.1.4 Geoprocessing

In hydrological science, it is widely acknowledged that the geometric characteristics of a river network play an important role as to the shape, timing and peak magnitude of a flood wave - see e.g. Maniak (2013) or Patt and Jüpner (2001). In order to take geometric properties into account, six predictors were derived using the "arcpy" and "numpy" packages in python (Rossum, 1995).
A 100x100m digital elevation model (DEM), obtained from the German Federal Agency for Cartography and Geodesy (BKG, 2010), and the respective slope raster were at hand and were averaged to give mean *Altitude* and mean *Slope*. Also, flow direction and flow accumulation rasters were available. Commonly, the elevation dataset is used to derive the direction of water flow in each cell. This, in turn, allows for the computation of flow accumulation for each raster cell, i.e. the number of upstream cells that drain into the respective cell. The latter is also referred to as the support area. This was used to extract the stream network of each of the basins. Doing this, a critical step is to find a valid threshold of support area values that defines stream or non-stream-cells. The goal here is to identify a stream network that is as close as possible to the real one. The support area threshold is specific for each catchment as it depends on various factors such as slope, geology, vegetation and soil and geomorphological properties. Tarboton et al. (1991) state that the ideal approach is a visual comparison with topographic maps. As this was not feasible for all 374 catchments, a different approach was needed that could be applied to all catchments in an automated fashion: Using a support area threshold of 1km^2, i.e. 100 cells, synthetic stream networks were created - i.e. stream networks with a higher number of streams than there are in reality. On these synthetic stream networks, Strahler stream order was assigned (Strahler, 1957). In order to approximate the true number of streams, only the streams of the three highest Strahler orders were kept. Visual comparison of spot samples of the resulting stream networks and topographic maps indicated a reasonable similarity. In some cases there are two main streams that merge in direct proximity (<= 500m) of the outlet, thus delineating a main stream of negligible length. In these cases, the stream network was corrected by including the four highest Strahler stream orders.
From the resulting stream networks, $ChSlope$ and DD were calculated. The latter is defined the following way:

$$DD = \frac{l}{Area} \tag{2.2}$$

where l denotes the sum of channel lengths and $Area$ is catchment area. DD and $ChSlope$ were included because they serve as a measure of runoff concentration and Flood Routing, respectively. Thus, they affect flood magnitude (Maniak, 2013).
The relation of flow length distribution and precipitation event duration determines the shape of the flood wave, thus controlling flood magnitude (Patt and Jüpner, 2001). Therefore, flow length from each cell in the catchment to the outlet cell was calculated and two relevant shape parameters of flow length distribution were derived: The maximum value $FLMax$ and the coefficient of variation, $FLCV$. Regarding the latter, the following principle holds: Flow lengths are more similar the more circular the catchment shape is. Therefore, all areas drain at the same time, increasing peak flows (Maniak, 2013).

2.1.5 Sampling of Preconditions

As mentioned in chapter 1.2, the principal focus of this study is to quantify the effect of preconditions on flood magnitude. Thus, following the identification of the date of Q_f, dynamic predictors P, P_eff, SM and T were sampled to extract the preconditions prior to the flood event. The respective sampling scheme is displayed in figure 2.4: Going back in time from the day of Q_f, the means of predictor values over multiple time periods, Δt, were extracted. These time periods range from $\Delta t = 0$ d, i.e. the predictor value on the day of Q_f, to Δt=7 d – Which gives the mean value of the 7 days prior to the flood event. In the following paragraphs, the predictors are denoted by their name and the respective time interval, e.g. SM_0, SM_1...SM_7.

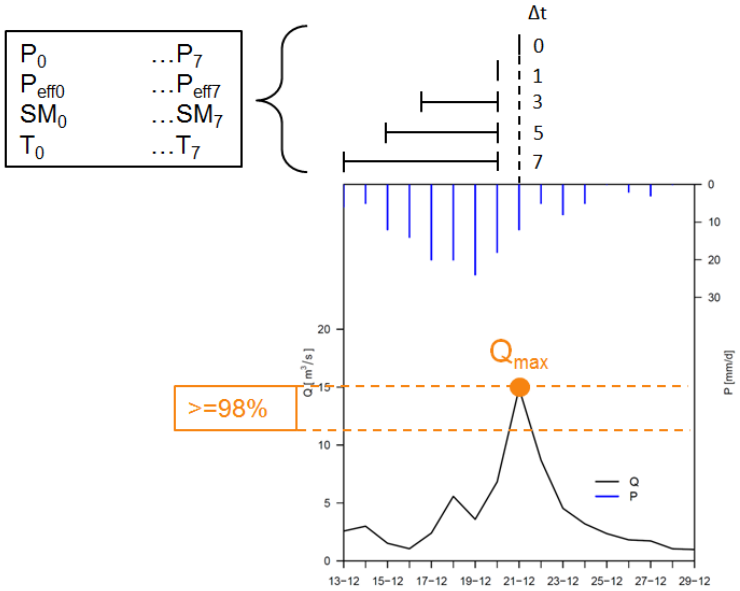


Figure 2.4: Sampling procedure to derive preconditions from predictor time-series and the respective denomination. Δt denotes the time interval in days of which mean values are obtained.

2.2 Regional Subsets

In order to be able to detect spatial variations in the effect of preconditions it is important to define regions of catchments that are similar in their general response characteristics i.e. in static variables that influence the flood generation processes mentioned in chapter 1.1.1. Therefore, the dataset was split into subsets following the classification of "Natural Regions" of the German Federal Institute of Regional Studies (Meynen et al., 1953). These are based on geomorphological, geological, hydrological, ecological and pedological criteria. Thus, they are assumed to represent the entity of static factors, even those that are not included in the actual dataset like soil or vegetation properties. This classification has been applied in both administrative and research context for 5 decades now and is regarded as the standard classification of Germany's natural regions. Of the five main regions that were originally defined, two were merged into one: Alps and its foothills. This results in the four regions as depicted in fig. 2.5. Merging the alps and its foothills into one region was necessary to ensure sample sizes that are comparable throughout the regions. Also, with respect to the number of predictors that are included in the models, minimum sampling size was set to 1000 datapoints. This would not have been met in the Alps-region as a single region. The resulting datasets are referred to as "regional datasets" in the following paragraphs.

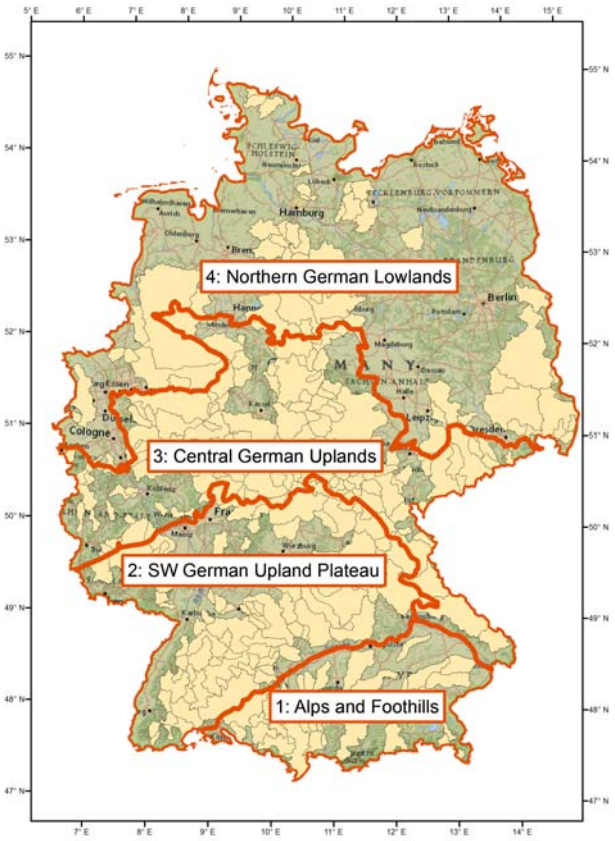


Figure 2.5: Map of the study regions and the respective catchments.

2.3 Regional-Seasonal Subsets

In addition to an analysis in space, the aims of this study include an analysis of variations in flood generation processes between summer and winter. Following the classification of a hydrological year in Germany, winter season was defined from 1st of November to 30th of April and summer season from 1st of May to 31st of October. The resulting datasets are referred to as "regional-seasonal" datasets throughout this study and are labeled according to region number and season, e.g. 1S, 1W, 2S and so on. Accordingly, the respective models are refereed to "regional-seasonal models" 1S, 1W, 2S etc.

2.4 The RandomForest Algorithm

This study applies Generalized Linear Models (GLM) and RandomForest (RF) as statistical models. As GLM is widely-used, the background of RF, only, will be briefly explained:
RFs are ensembles of decision trees that are trained on data to form a robust nonparametric model capable of handling large nonlinear, noisy, fragmented, or correlated multidimensional data for classification and regression (Liaw and Wiener, 2002). It extends classical decision tree methodology by a combination of bootstrapping and random variable selection (Breiman, 2001). The RF algorithm includes three main steps (compare fig. 2.6):

1. Draw n bootstrap samples from the data set.
2. Grow a decision tree for each bootstrap sample ("Trees 1, 2, B") in the following way: The data is split into subsequent subsets at so-called nodes that split the data best. At each node, only a random subset of the available predictors is considered and the best splitting criterion, i.e. a specific value of one of the predictors, is determined. Here, the objective function is the reduction in total variance before and after the split. This is done for n trees on different bootstrap samples with different predictors at each node due to random variable selection.
3. Predict new data from the majority vote, i.e. the mean of all trees' predictions for regression.

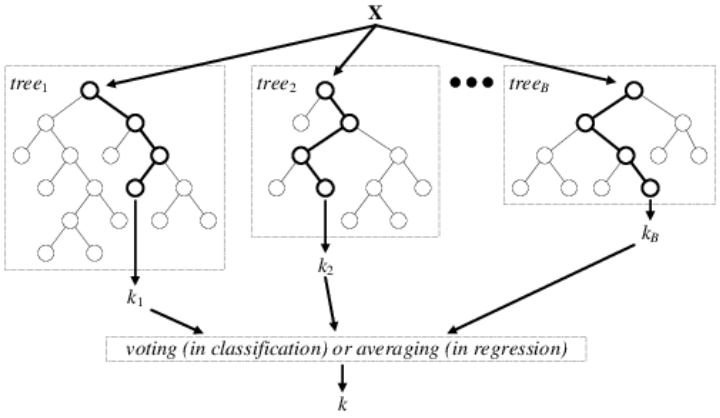


Figure 2.6: Simplified structure of RandomForest algorithm (source: Pedregosa et al. (2011)).

The main advantages of RF are that it is not based on any distribution functions and therefore does not adhere to any assumptions. Also, it automatically performs variable selection. It can detect non-linear relationships in the data that other piecewise linear regressions are not able to detect and has been experienced to be fast in training (Hastie et al., 2009; Shortridge et al., 2016).
The main disadvantage is that RF is purely data-driven, so a considerable amount of data is needed to ensure an adequate model fit. Also, with common number of trees of more than a 100, a detailed examination of these is infeasible. RF is implemented in R in the package "randomForest" by Liaw and Wiener (2002). The parameters of random forests are the number of "out-of-bag" in each bootstrap, the number of randomly selected predictor variables at each node, the number of trees and the trees' complexity (Breiman, 2001). This study applied the standard parameter values of the above package.

2.5 Analysis of Collinearity

Prior to model calibration, hierarchical clustering using Pearman's Squared Correlation Coefficient, r^2, was applied to assess collinearity among predictors. This was executed by the use of "Hmisc"-package (Harrell, 2018) for all 4 regional datasets, separately. RF is insensitive to collinearity and GLM as well – if applied on PCA-transformed data as in this study (ch. 2.6). Therefore, this analysis only serves the purpose of data exploration to facilitate interpretation of results.

2.6 Model Calibration

2.6.1 General Information

For each of the 8 "regional-seasonal datasets", both RF and GLM were calibrated – Giving a total of 16 models. Model calibration and validation were carried out on separate subsets of each of the 8 regional-seasonal datasets - Training and test data. Following the procedure as proposed by Hastie et al. (2009), the ratio of test vs. training-subsets was $\frac{3}{4}$ vs. $\frac{1}{4}$. Balanced sampling was applied, i.e. randomized splitting was repeated 50 times and the pair of training and test data that were closest to each other in standard deviation and mean value were used for the analysis. For both model algorithms, a feature selection algorithm was applied to identify the model structure that produces the best goodness-of-fit.

2.6.2 Principal Modeling Approach

The general equation of RF model structure is:

$$\hat{Q}_f = f(x_1, x_2, ...x_i), \tag{2.3}$$
$$\text{with} \quad x_1^i \in P_{eff0...7}, SM_{0...7}...$$

RF was calibrated using "Recursive Feature Elimination" (rfe) from caret-package (Kuhn, 2008), an iterative reduction of model complexity. Starting from a full model that included all 40 predictors, the 5 least relevant predictors (by variable importance) were excluded at each iteration. Model performance of the resulting models was assessed by 10-fold Cross-validation using Root-Mean-Squared-Error (RMSE) as objective function. This serves two purposes: Firstly, the best-performing model structure is selected. Secondly, model robustness is tested by the use of Cross-Validation – thus, reducing the risk of Overfitting. The general model structure of GLM is:

$$\hat{Q}_f = \beta_0 + \beta_{11}x_1 + \beta_{12}x_1^2 + \beta_{21}x_2 + \beta_{22}x_2^2 + ... + \beta_i x_i + \beta_{ii}x_i^2 + \epsilon \tag{2.4}$$
$$\text{with} \quad x_1^i \in P_{eff0...7}, SM_{0...7}, ...$$
$$\text{and} \quad \epsilon \sim Normal(0, \sigma^2)$$

where $\beta_0^i i$ are the regression coefficients and ϵ is the error following a gaussian distribution with mean zero and variance σ^2.
For GLM, Principal Component Analysis (PCA) was performed on each of the training datasets, prior to model calibration. This approach was used to account for multi-collinearity as present in the dataset (ch. 3.4). PCA applies orthogonal transformation on the data to give n_{PC} orthogonal, i.e. non-correlated principal components (PC) in an alternative, multi-dimensional feature space. These explain $s\%$ of the data's variance. For further details, see Dunteman (1989). In this study, $s = 95\%$ was applied and the resulting n_{PC} PCs were used as inputs for the GLM. Next, the GLM was calibrated using "Stepwise Backwards Selection" by BIC (stepBIC) to selected which PCs to include in the model. This procedure is similar to rfe, only that, at each iteration, each of the predictors is taken out of the model and only the least important one is dropped. Also, it uses "Bayesian Information Criterion" (BIC) as objective function and no Cross-Validation is performed. However, the BIC is known to effectively reduce model complexity, thus minimizing the risk of Overfitting (Dormann, 2017). BIC was chosen instead of the commonly-known AIC as it penalizes complex models more heavily, leading to simpler, thus more robust models (Hastie et al., 2009).

2.6.3 Alternative Modeling Approaches

Adding to the principal modeling approach as presented above, alternative approaches were applied:
RF:

- **P instead of Peff:** RFs were fit using a similar model structure but using observed P instead of modeled P_eff. This was done to quantify whether including a mHm-model estimate of snow accumulation and melt with its own inherent inexactitudes leads to higher model performance.
- **Scaling of catchment subsets:** Chapter 1.1.1 states that the model structure has to account for the general level of wetness across catchments. In the principal modeling approach, this is done by including control variables. Another approach is to standardize each catchments' flood and predictor data to $[0, 1]$ with 0 being the lowest and 1 the highest observed record. This standardized approach was applied to find out whether it would outperform the principal approach.

GLM: The Generalized Linear Model is called "generalized" because it allows for other error distributions than the normal distribution. As to the skewness of Q_f, lognormal- and gamma-distribution were applied in alternative calibrations to quantify whether the distribution of residuals would improve. Also, for physical plausibility, a truncated gaussian distribution was applied that limits the predicted response to $[0, \infty]$.

2.7 Model Validation

Analysis of model performance is carried out for each of the 8 datasets, separately. Using the unseen test data, model performance is evaluated using the Coefficient-Of-Determination (R^2). In addition to goodness-of-fit measures, analysis of residuals is part of the model validation process.This is carried out by plotting residuals against predicted values in each region and season.

2.8 Model Interpretation

Interpretation of results was based on the best model algorithm. However, clear trends or results of single regional-seasonal models of the second best algorithm were included under consideration of the respective Goodness-of-Fit. For this, a threshold of $R^2 > 0.7$ was set.
The analysis was performed on three scales: The regional scale compares averages between the regions, the regional-seasonal scale includes seasonality and, adding the time intervals of preconditions, Δt, results were interpreted on regional, seasonal and temporal scale.

2.8.1 Variable Importance

RandomForest:
The model outputs of RF allow for an intuitive interpretation of the explanatory power of predictors by calculating "Variable Importance" (VI). There are two ways of doing this (Hastie et al., 2009): "VI by increase in node purity" and "permuted VI by increase in MSE". Here, MSE refers to Mean-Squared-Error. The first one sums the explained variance attributed to a certain predictor over all trees. While providing a direct representation of the model structure, it is known to be biased by scale of measurement and levels of categorical variables because the underlying Gini gain splitting criterion is a biased estimator and can be affected by multiple testing effects (Strobl et al., 2007).
Therefore, it is generally advised to use permuted VI by increase in MSE (Breiman, 2001; Strobl et al., 2008; Parr et al., 2018): It is calculated from out-of-bag samples, i.e. the ones that were excluded by bootstrap procedure during model fit. In these, values of the respective predictor are permuted while keeping the response unchanged. Next, the increase in MSE in comparison to non-permuted values is calculated. This measure serves as a mini-sensitivity analysis that follows the assumption that the decrease in prediction accuracy is stronger the more important a predictor is for the model. This method is regarded to be more robust as the above due to its randomized approach. As advised by Strobl et al. (2008), the non-standardized version of this variable importance measure was used.
GLM:
In GLM-analysis, predictors are usually standardized or normalized before calibrating the model, so that all predictors are on the same scale and regression coefficients can be interpreted directly as to their sign and absolute value. Here, predictors were transformed into n_{PC} principal components. These PCs were used as actual model inputs for GLM and were assigned coefficient values. As such, these are difficult to interpret. However, as PCA is a linear transformation, coefficient values were retransformed according to the load of each predictor on each PC. As this retransformation goes back to the original input scale, predictors were normalized to $[0, 1]$ before PCA-analysis. Thus, each of the 40 predictors and its second-order polynoms got assigned a coefficient value of either negative or positive sign, indicating the respective explanatory power and the direction of the functional relationship. For better interpretability, first and second order polynoms‘ coefficients were summed. By doing this, a measure similar to the variable importance of RF was at hand. Therefore, it is termed "variable importance" throughout this study, as well.
For both RF and GLM, variable importances had to be comparable among the 8 regional-seasonal models. Therefore, they were scaled to the sum of each models‘ importances, termed "relative variable importance", given in %. Depending on the scale of analysis, variable importances were averaged by region, season or predictor.

2.8.2 Partial Dependence Plots

In order to visualize the functional relationship of predictors and target variable as fitted by the models, partial dependence plots (PDPs) were computed for RF, types using the package "pdp" (Greenwell, 2017). The partial dependence function of the target variable $F_s()$ on predictor x_s is defined as follows (Friedman et al., 2008):

$$F_s(x_s) = E_{x\setminus s}[F(x_s, x_{\setminus s})] \tag{2.5}$$

where $x_{\setminus s}$ denotes all other predictors included in the model, so $x = x_s + x_{\setminus s}$. $E_{x\setminus s}$ is the expected value over the joint marginal distribution of all predictors $x_{\setminus s}$. Partial dependence functions can be estimated from data by:

$$\hat{F_s}(x_s) = \frac{1}{N}\sum_{i=1}^{N} F(x_s, x_{i\setminus s}) \tag{2.6}$$

where $\{x_{i\setminus s}\}_1^N$ are the N data points of $x_{\setminus s}$. Thus, partial dependence functions are derived by averaging the effect of all $x_{\setminus s}$ over every single value of x_s.
For GLM, the value and sign of each predictors' first- and second-order polynomial were given, so there was no need to compute partial dependence plots. For comparability, PDPs of each regional-seasonal model are standardized to the respective mean of observed Q_f – Giving the relative deviation from mean in %. Tick marks of observed datapoints are included in the plots for validation to ascertain reliability of the respective areas of the plots.

2.9 Estimation of Prediction Uncertainty

In order to estimate the uncertainty of predictions, a method was needed that can be applied to both RF and GLM. GLM itself offers a convenient estimate: Based on the standard error of its coefficients, a confidence interval of its predictions can be derived. Non-parametric approaches like RF do not offer anything similar, the only way to derive estimates of prediction uncertainty is by randomization as in Cross-Validation or Bootstrap (DiCiccio and Efron, 1996). Therefore, a bootstrap procedure was applied by performing the following steps (Efron and Tibshirani, 1994):

1. Draw 100 bootstrap samples of training data: Random sampling with replacement
2. Fit models to each of the bootstrap samples, applying the model structure that was ranked best by feature selection algorithms rfe and stepBIC
3. Predict test data with each of the 100 models. Visualize lower and upper bounds for each of the data points

This procedure gives an estimate of how the predictions would change if the training data was a different realization of its population. As training and test data originate from the same population, it is an estimate of how well the respective model represents and predicts the joint population of both test and training data - i.e. the complete dataset.

Chapter 3

Results

3.1 Sampling of Flood Magnitude

Table 3.1 presents the size of the candidate datasets that were derived by applying 95%, 97% and 98% thresholds in POT-analysis as well as the share of durations above the physically-plausible threshold of d >20d. For the 95%-percentile, the share of events >20d is considerably higher than with the other thresholds. To visualize the performance of the three thresholds, exemplary annual time-series of selected stations are depicted in figure 3.1. Here, 98%-threshold outperforms 97%-threshold in singling out the actual flood event from secondary flood waves before or after the main peak. Consequently, the dataset of the 98%-threshold is used - Thus, the highest 2% of all daily streamflows of each of the 374 catchments were extracted and grouped into flood events. This resulted in 37912 flood records. After filtering out independent events and excluding events with a duration >20d, the final dataset of all floods at all stations consists of 29247 records.

Table 3.1: Statistics of the resulting flood datasets for different percentile thresholds.

Percentile X	**95%**	**97%**	**98%**
n floods	68918	49988	37912
Fraction with Duration>20d	1.2%	0.6%	0.3%

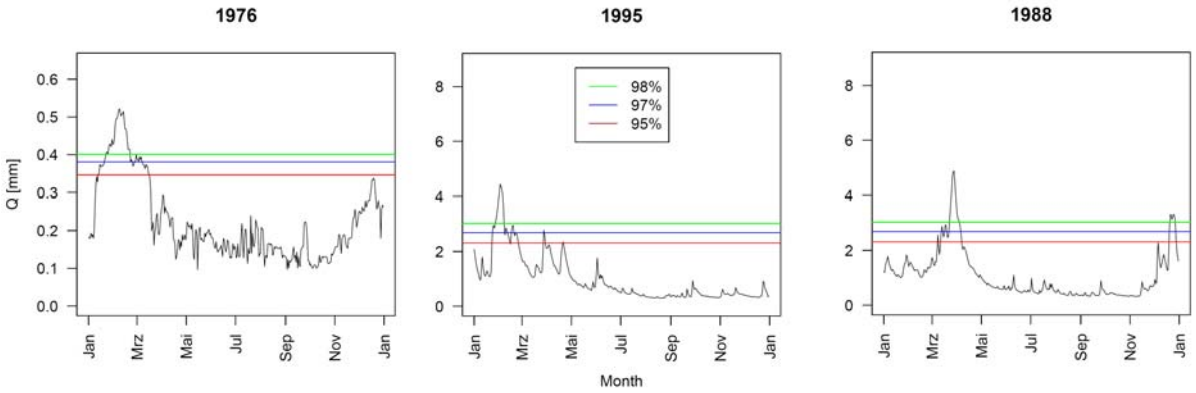


Figure 3.1: Performance of different percentile thresholds: Exemplary annual time-series of the gauging stations "Große Tränke, Spree" (left) and "Herrenhausen, Leine" (center and right).

3.2 Dataset

3.2.1 General Information

This section presents the relevant meta-data on the four resulting regional datasets as given in table 3.2. These are displayed once more in figure 3.2. It can be seen that there is a south-to-north gradient:

- **Size of Regions:** The regions vary in size with the southernmost region 1 making up only 10% of overall area, and northernmost region 4 covering almost half of the area.
- **Spatial Coverage:** In regions 1,2 and 3 catchments are distributed evenly and spatial coverage is similar. In region 4, the catchments are clustered in places and overall gauge density is only half the one of the other regions.
- **Data Coverage:** Data coverage in time-series is similar, a little lower in region 4.
- **Event Duration:** Region 1 exhibits shorter floods than the other regions. In region 4, average flood duration is considerably higher, almost twice as high.

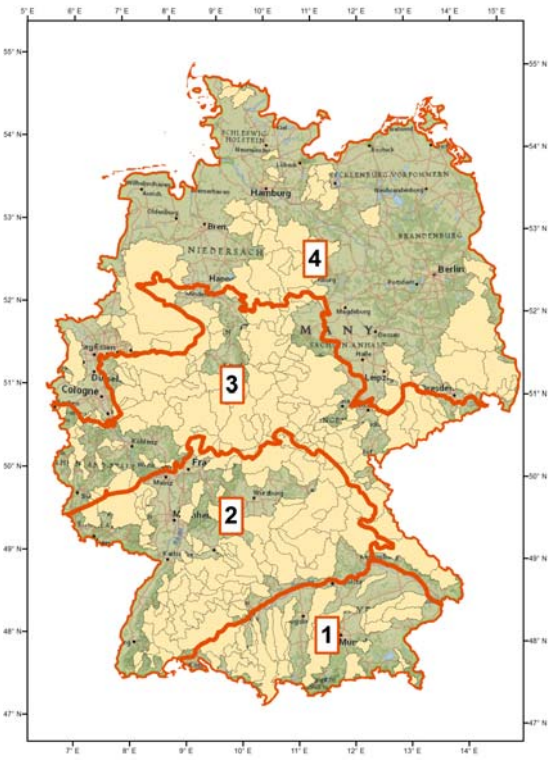


Figure 3.2: Map of the study regions and the respective catchments, duplicate.

Overall, the sampling procedure identified 29247 events at 374 stations. However, these are not spread evenly over the regions: The amount of events in each region and at each gauging station is only partly related to spatial and data coverage. This is due to the nature of data processing: At each station, a fixed share (2%) of maximum flows was extracted. Subsequent maxima that belong to one flooding event were classified as such. Consequently, the duration of events varies. Region 1 exhibits floods of lower duration than the others. In region 4, on the other hand, average flood duration is almost twice as high. This results in an average number of events per station that is only half of the one in region 1. Adding this to low spatial and data coverage, overall number of events in relation to the area is low in region 4. Both regions 1 and 4 account for about 17% of all events. Regions 3 and 4 attribute to 34 and 33% of all detected events.
Separation into summer and winter subsets of each region resulted in seasonal subsets of lower sample size: Except for region 1, only about 20 % of floods occur in summer (see table 3.3).

Table 3.2: Meta-data of the regional datasets. The rightmost column shows either total or mean values. "p.s" means per gauging station.

Region		**1**	**2**	**3**	**4**	**Total**
Name		Alps and foothills	SW German upland plateau	Central German uplands	North German lowlands	
Area	km^2	37499	75304	89270	155800	357874
Area	%	10	21	25	44	100
Stations	-	44	121	125	83	373
Station density	km^{-2}	0.0012	0.0016	0.0014	0.0005	0.0012
Mean Catchment size	km^2	415	724	681	1227	762
Mean coverage	%	67	65	69	56	64
Mean duration	d	2.2	3.0	3.2	3.6	3.0
Mean events p.s.	-	114	81	77	58	82
Events	-	5022	9821	9612	4792	29247
Events	%	17	34	33	16	100

Table 3.3: Summary statistics of the regional-seasonal datasets.

Region	Season	Type	n	%	Type	n	%
1	Summer	Training	1919	51	Test	672	54
1	Winter	Training	1847	49	Test	584	46
2	Summer	Training	1601	22	Test	564	23
2	Winter	Training	5764	78	Test	1892	77
3	Summer	Training	1331	18	Test	435	18
3	Winter	Training	5878	82	Test	1969	82
4	Summer	Training	664	18	Test	244	20
4	Winter	Training	2930	82	Test	954	80

3.3 Dataset Analysis

Generally, the differentiation into natural regions by Meynen et al. (1953) is reflected in distinct characteristics among the regional datasets. Many variables exhibit a gradient from south to north, both in average values and variability. The following paragraph gives an overview of the spatial patterns, starting with Q_f and dynamic predictors, then presenting static variables that show a clear gradient and finishing with the static variables that give an inconsistent pattern. As the distributions of Q_f, P_{eff}, *Area* and flood duration are left-skewed in all regions, "average" refers to median value as depicted in the respective box-and-whisker plots. These boxplots are set to show whiskers up to 1.5 of Inter-Quartile-Range. For comparison, the terms small vs. large, low vs. high etc. are used in a relative sense, i.e. only contrasting between the regions but not referring to absolute values. For clarity, variables that do exhibit a clear south-to-north gradient in average values are depicted in figure 3.3.

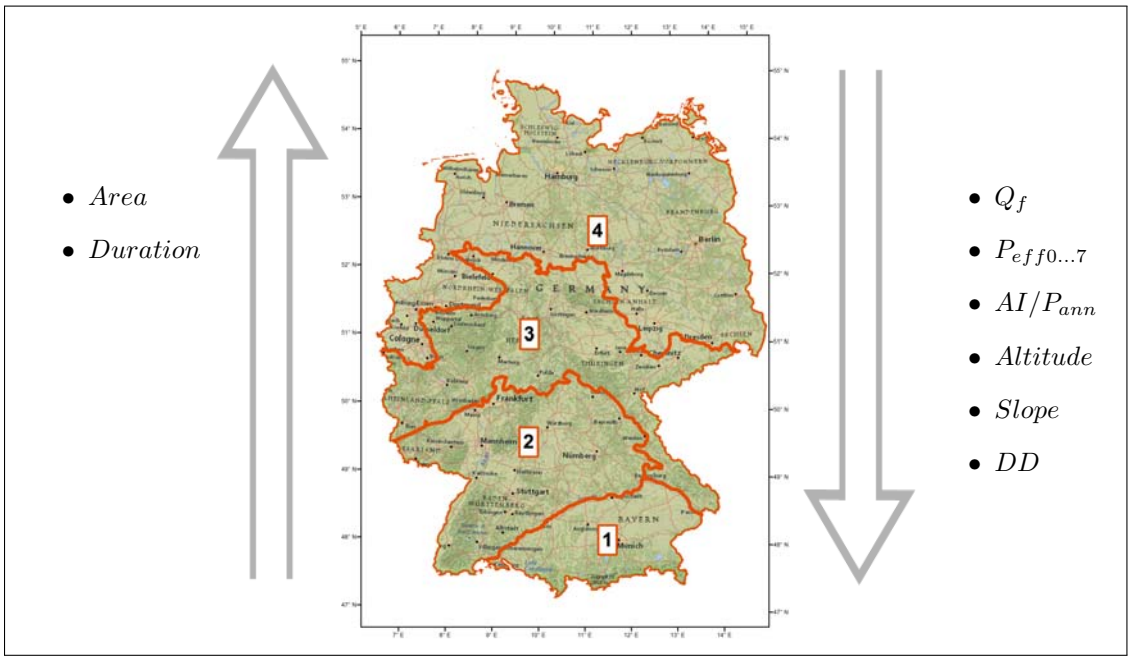


Figure 3.3: Variables that exhibit a clear gradient across the study area. The direction of the arrow reflects an increase in average values.

All relevant information on the spatial pattern of Q_f are depicted in figure 3.4. Q_f is highly left-skewed in all regions and average values decrease from south to north. Also, variability of Q_f is highest in the south, decreasing towards the north (see boxplot). When analyzing the per-station means of Qf as depicted in the map, it can be seen that high values generally cluster in mountainous regions such as the Alps, Black Forest, Bavarian Forest, Ore Mountains and the Renish Uplands.

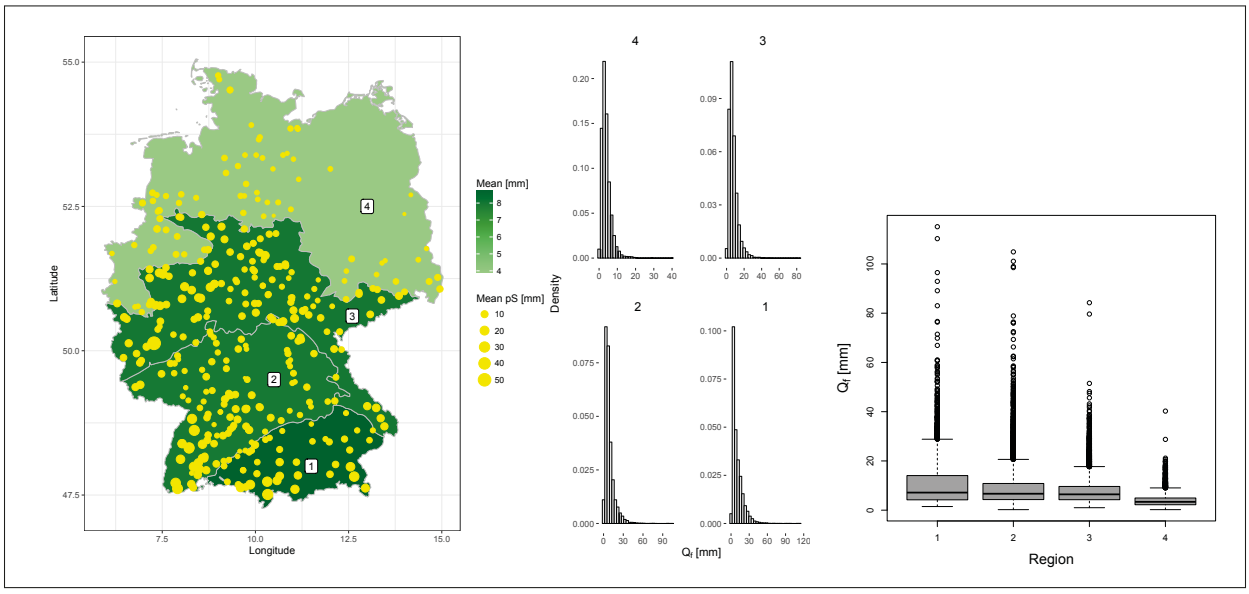


Figure 3.4: Statistics of Q_f across the regions. Left: Map of mean per region and per station ("pS", yellow dots). Center: Histogram of Q_f across regions, note that x- and y-axes are not standardized. Right: Boxplot of Q_f across regions.

A similar trend is seen for pre-event P_{eff} where all $P_{eff0...7}$ show a similar pattern (fig. 3.5): A decrease from south to north in both average values and variability. In pre-event SM, there is no clear gradient but all estimates $SM_{0...7}$ exhibit the same spatial pattern: regions 1 and 4 have higher average values than regions 2 and 3 with region 4 showing strong variability (fig. 3.5). As to T it is to be noted that there is no gradient but all $T_{0...7}$ are similar in their spatial pattern: Region 1 exhibits higher averages and variability than the other regions (fig. 3.5).

Several static variables decrease in average values from south to north:
General wetness as represented by AI and P_{ann}, with strong variability in the southernmost region. Here, high per-station means cluster in mountainous areas similar to Q_f (fig. 3.6). *Altitude* and *Slope* decrease both in mean values and in variability from south to north. DD also decreases towards the north with low values in region 4 and high variability in regions 2 and 3.

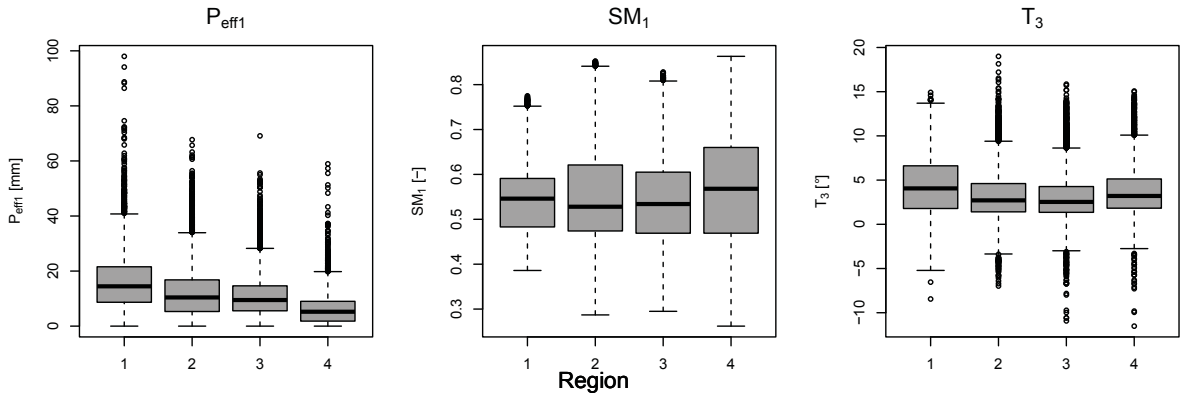


Figure 3.5: Boxplots of P_{eff1}, SM_1, T_3. These are representative in their spatial pattern of $P_{eff0...7}$, $SM_{0...7}$ and $T_{0...7}$.

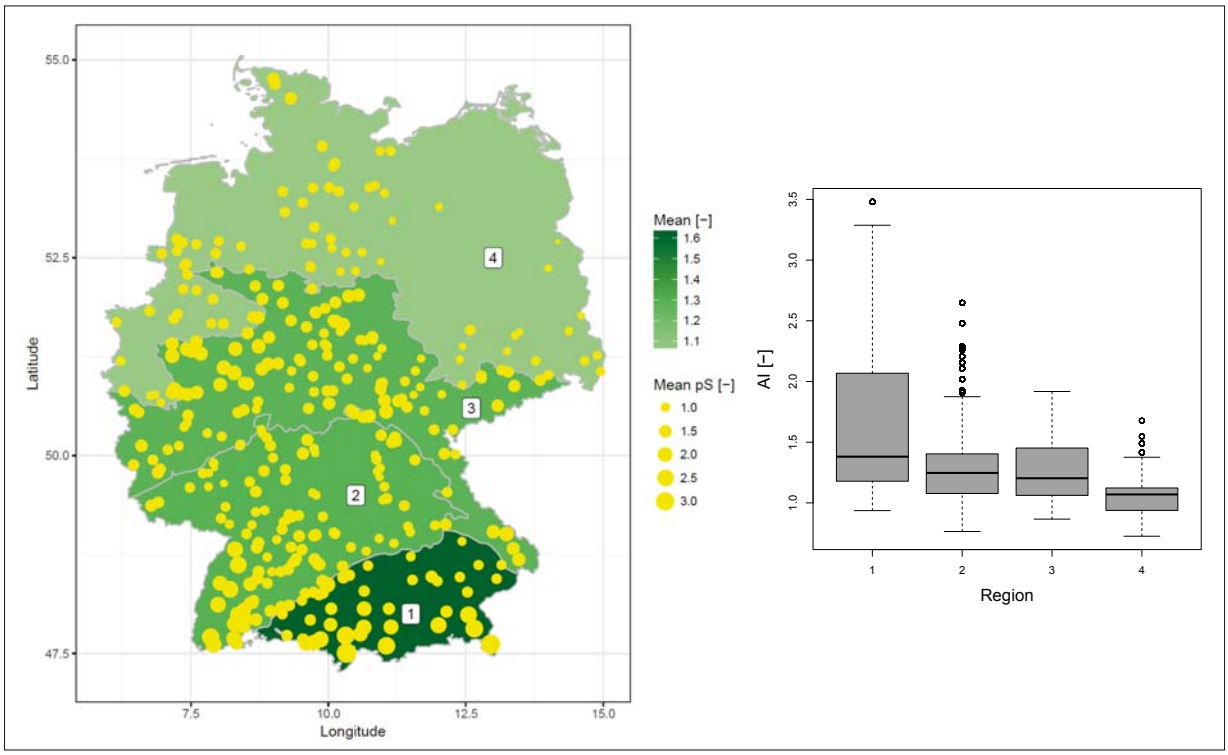


Figure 3.6: Left: Map of means of *AI* per region and per station ("pS", yellow dots). Right: Boxplot of *AI*.

Several variables exhibit an opposite gradient, increasing from south to north: This is the case for *Area* (fig. 3.7) and *Duration*. In the southernmost region, the majority of floods last one day whereas in the northern regions they last 2 to 3 days with a large share of values spreading up to 8 days (fig. 3.8).

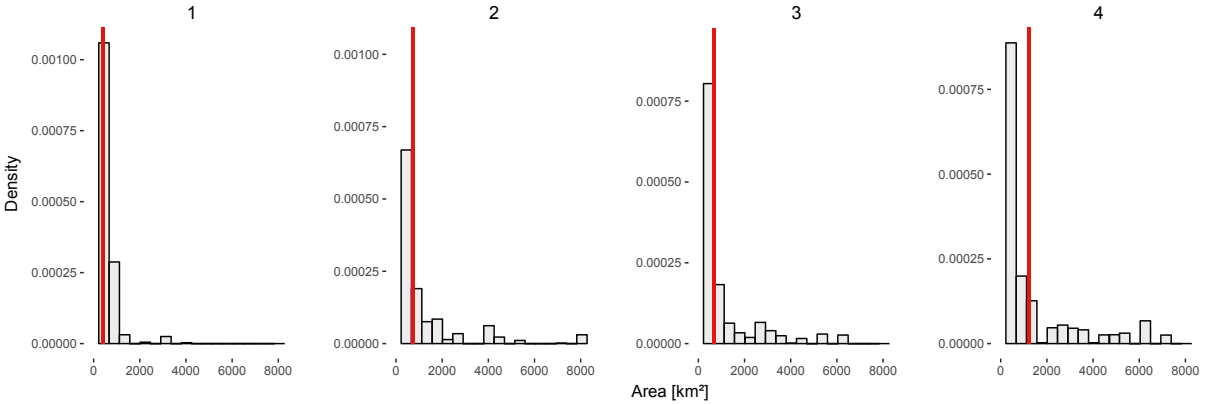


Figure 3.7: Histograms of *Area* across the regions. Red line depicts the mean value.

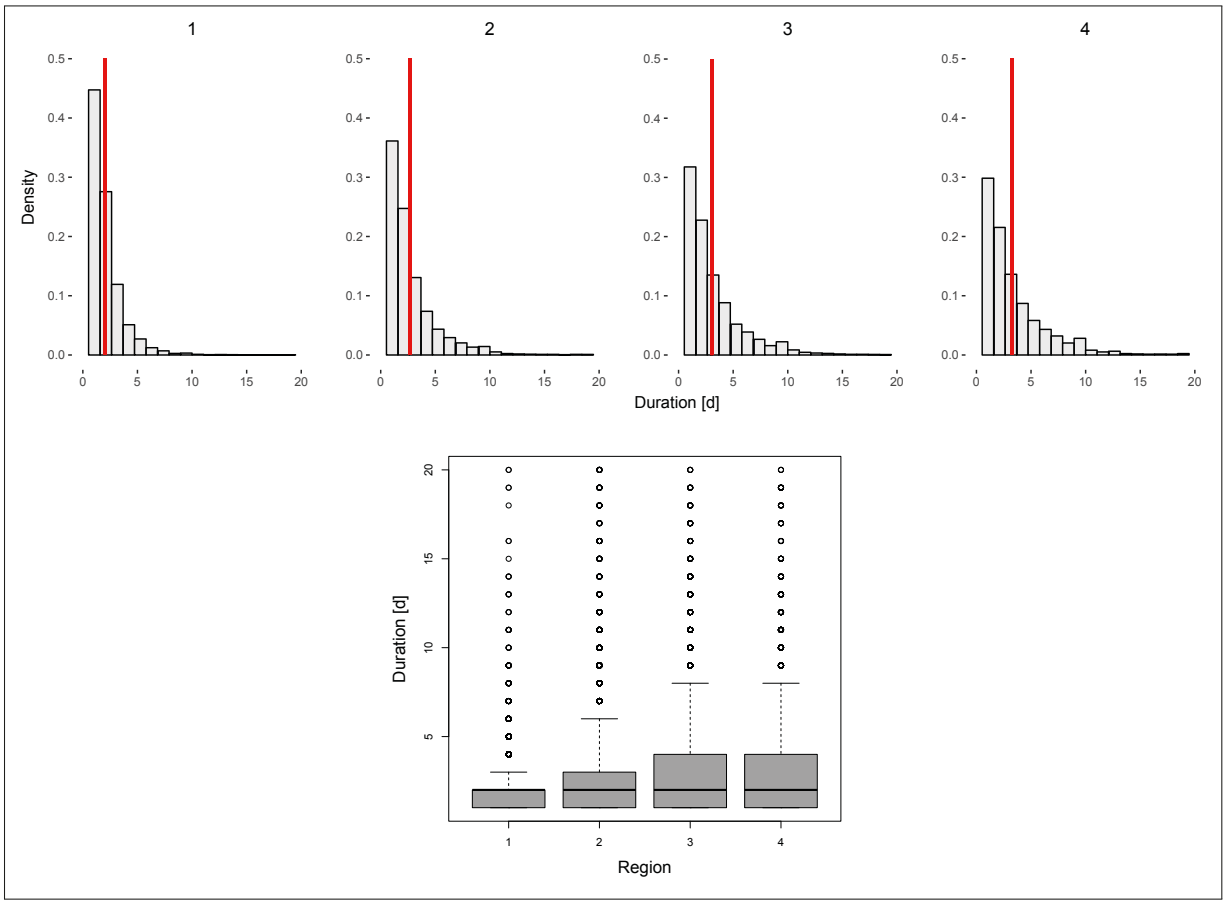


Figure 3.8: Histograms (top) with means (red line) and boxplot (bottom) of *Duration* across the regions.

Some static variables show an inconsistent pattern:
Land-cover is one of these: Regions 2 and 3 are dominated by *Forest*, whereas regions 1 and 4 exhibit high values of *Permeable*. *ChSlope* exhibits high values and variability in regions 2 and 3. Values of $FLMax$ are high in regions 1 and 4.

3.4 Analysis of Collinearity

Analysis of collinearity was run on each of the regions, separately. As all regions show a similar pattern, only the one of region 2 is displayed here (fig. 3.9), the complete set of collinearity cluster plots can be found in the appendix.

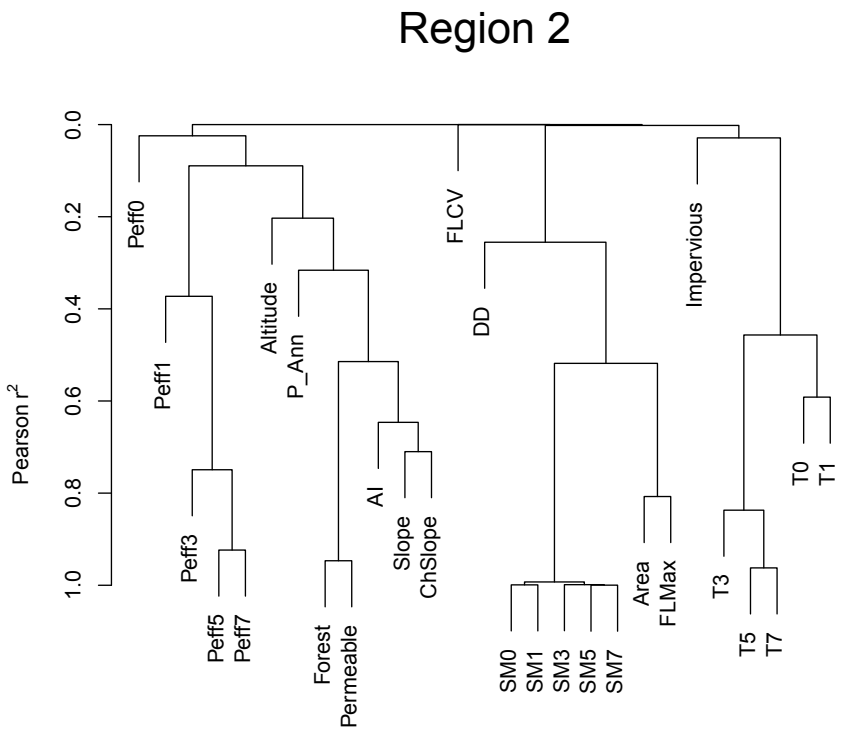


Figure 3.9: Clusters of collinearity among predictors.

Major collinearity exists between the following predictors:

- P_{eff3} - P_{eff7}
- $Forest$ vs. $Permeable$
- $Slope$ vs. $ChSlope$
- SM_0 - SM_7
- $Area$ vs $FLMax$
- T_3 - T_7

3.5 Model Validation

3.5.1 General Information

Generally, RF and GLM models of the structure listed as "principal modeling approach" show higher accuracy on training data than the "alternative approach" (see chp. 2.6). RF is applied to raw (i.e. not scaled per catchment) input data and included P_{eff} instead of P. GLM fits a standard gaussian distribution. Both models show similar accuracy in calibration and validation, so here only the validation performance on test data is presented (tab. 3.4). Performance metrics on training data can be found in the appendix.
Generally, RF outperforms GLM on all but one regional-seasonal datasets, both in R^2 and RMSE. For both models, accuracy is highest in the southernmost region 1, decreasing towards the north, lowest in region 3. Also, except for region one, accuracy on summer datasets is considerably lower than on winter ones. Setting a value of $R^2 >= 0.7$ as the threshold for acceptable model accuracy, RF lies below this value in three regions. GLM, on the other hand, meets this criterion only in two regions.
Figure 3.10 displays observed vs. predicted- and residual plots. Here, only some exemplary plots of both RF and GLM are displayed, the complete set of analysis plots can be found in the appendix. The models of 1S and 3S are displayed as they are the ones of highest and lowest accuracy. As can be seen in these plots, most models are accurate in low and medium values. Towards the upper end of the value range, all models exhibit heteroscedascity that gets stronger the lower model accuracy is. It is in high values, also, that prediction uncertainty increases for both models – With smaller, more heterogeneous uncertainty bands of GLM.

Table 3.4: Model accuracy by R^2 of RF and GLM on test data across regions and seasons.

Region/Season	**1s**	**1w**	**2s**	**2w**	**3s**	**3w**	**4s**	**4w**	**Mean**
RF	0.86	0.78	0.69	0.78	0.49	0.73	0.58	0.75	0.71
GLM	0.74	0.8	0.47	0.59	0.38	0.56	0.40	0.44	0.55

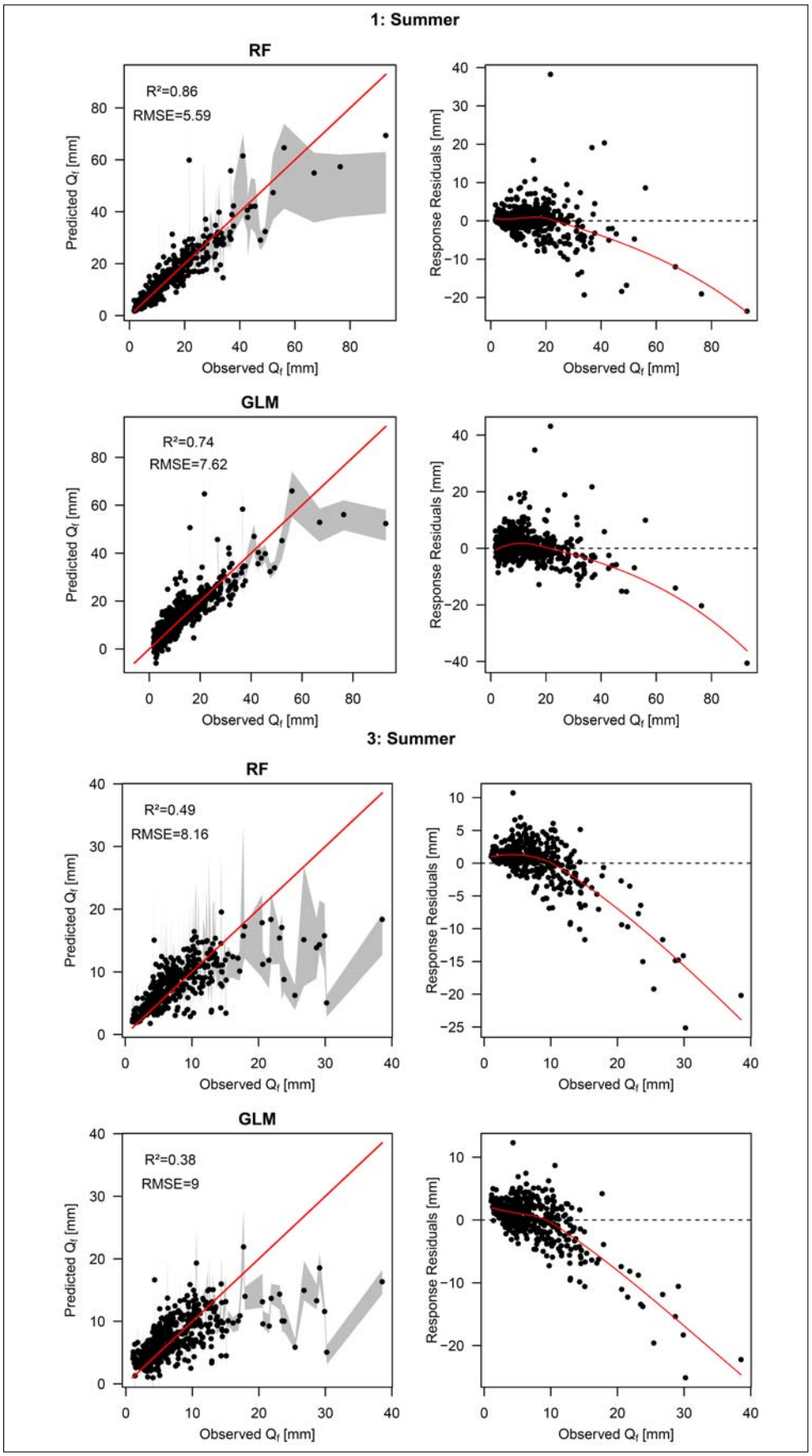


Figure 3.10: Analysis plots of RF and GLM on 1S and 4W test data. Left: Predicted vs observed Q_f. Red line depicts ideal 1:1 fit. Grey areas depict the lower and upper prediction bounds of a 100-fold bootstrap procedure. Right: Residuals on response scale. Red line depicts a LOESS-regression on residuals to visualize goodness-of-fit in lower values of Q_f.

3.5.2 RandomForest

On average, the use of P_eff instead of P led to an increase in mean model accuracy of 0.02, up to 0.05 in winter (see appendix). For the final model, feature selection by recursive feature elimination led to elimination of 1-8 predictors (tab. 3.5). The models of southern regions required less predictors than the ones in the north.

3.5.3 GLM

Prior to model calibration, training data, consisting of 40 predictors, was transformed by PCA. This resulted in n_{PC}= 13 - 16 principal components to explain 95% of the data‘s variance (tab. 3.5). GLM was fit to these PCs applying the following error distributions: Normal, lognormal, gamma and truncated normal. Of these, the principal model assuming normally-distributed errors reached highest accuracy. As depicted in table 3.5, the number of PCs that were used as model input, n_{PC}, decreased by 0-8. Accuracy og GLM reaches the one of RF only in region 1.

Table 3.5: Model sizes across regions and seasons: Number of predictors that were selected by feature selection algorithms stepBIC and rfe. For GLM, Principal Component Analysis already reduced 40 predictors to n_{PC} principal components.

Model	**1S**	**1W**	**2S**	**2W**	**3S**	**3W**	**4S**	**4W**
RF: n predictors	25	30	20	35	35	39	30	30
n predictors	40	40	40	40	40	40	40	40
GLM: n PCs	11	13	11	14	8	11	6	9
n_{PC}	13	13	15	15	16	16	15	15

3.6 Variable Importance

The following paragraphs present the variable importances as estimated by RF-models. If results of GLM are included, this is mentioned explicitly. GLM variable importances can be found in the appendix. For each of the 8x2 regional-seasonal models, variable importances were scaled to their sum to give "relative importance" (see ch. 2.8). First, the results regarding dynamic variables are presented, followed by the ones for static variables. Variable importances were derived on different scales and results will be presented in the according order: By season, by region and season, and where of interest, variable importances on temporal scale by region and season – i.e., the distribution of importances across different time intervals of preconditions (e.g. $P_{eff0...7}$). As AI and P_{ann} were included as control variables only, these are excluded from average importance calculations.

3.6.1 General Information

In all regions and seasons, average importance of static variables is about as high as P_{eff}, followed by SM and T (see fig. 3.11). Also, there is a distinct pattern between summer and winter: In summer, importances of all dynamic predictors are higher, the ones of static variables are lower. These patterns are reflected on regional scale (fig. 3.12) and will be presented in the following paragraph.

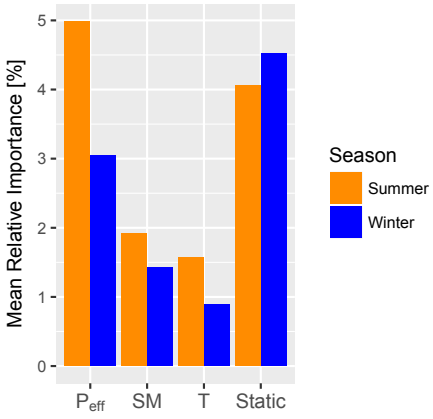


Figure 3.11: Average importance of dynamic and static predictors by season.

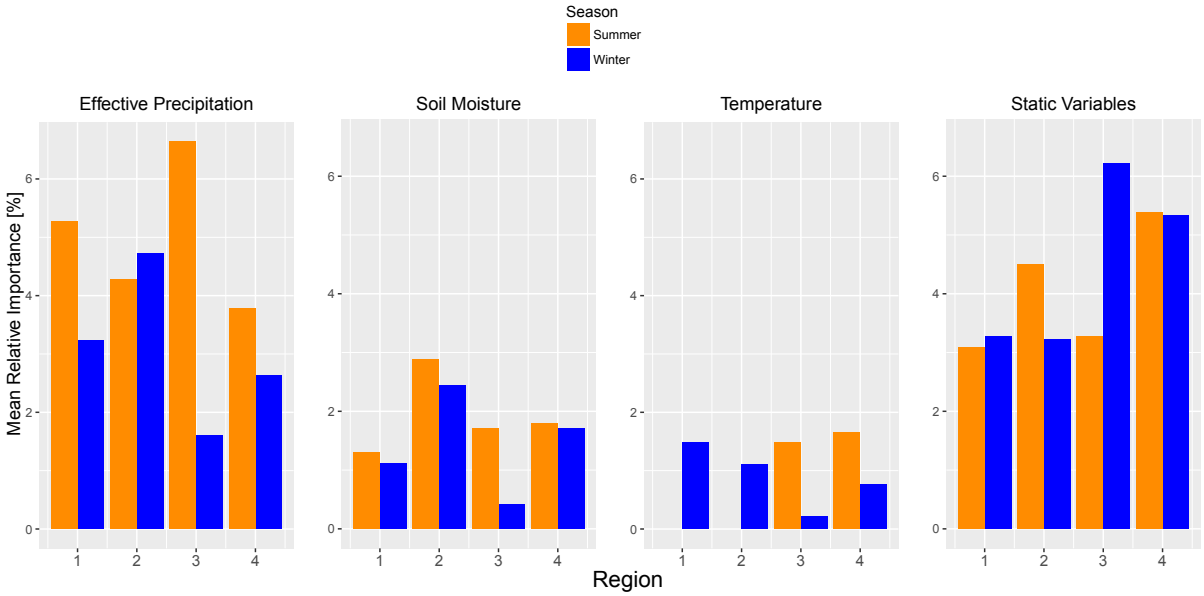


Figure 3.12: Average importance of dynamic and static predictors by region and season.

3.6.2 Dynamic Variables

Effective Precipitation:
As depicted in figure 3.11, P_{eff} exhibits the strongest difference between summer and winter. In figure 3.12, this trend can be analyzed on regional scale: Higher importances in summer can be observed in 3 out of 4 regions, strongest in region 3. Average importance is lower in region 4, supported by GLM. When moving a step further by analyzing the preconditions of P_{eff} on temporal scale (fig. 3.13), the following can be observed: First, all regions show a distinct pattern as to how variable importance of different time intervals changes by season. As before, region 3 shows the strongest variation by season. Second strongest is region 1, especially at $\Delta t = 1$. Second, in the southernmost region 1, preconditions in rainfall shortly before the flood event (1 to 3 days) get attributed highest importances. From south to north, the important time periods move back in time - So that, in region 4, the average preconditions of 5 and 7 days prior to the flood event get assigned highest importance.
Soil Moisture:
As shown in figure 3.11, variable importance of SM is only half of the one of P_{eff} and its seasonal difference is less pronounced. On regional scale (fig. 3.12), it can be seen that region 3 exhibits a stronger seasonal difference than the other regions. Average importance is lowest in region 1. This also reflects in figure 3.14, that shows importance of SM on temporal scale. Here, it is to be noted that 3 out of 4 regions exhibit an increase of importance of SM_0 in summer.

Temperature:
Generally, T gets assigned little importance. As seen in figure 3.12, summer importances are higher. However, this is only the case in the two northern regions 3 and 4. In the southern regions, T is only included in winter models.

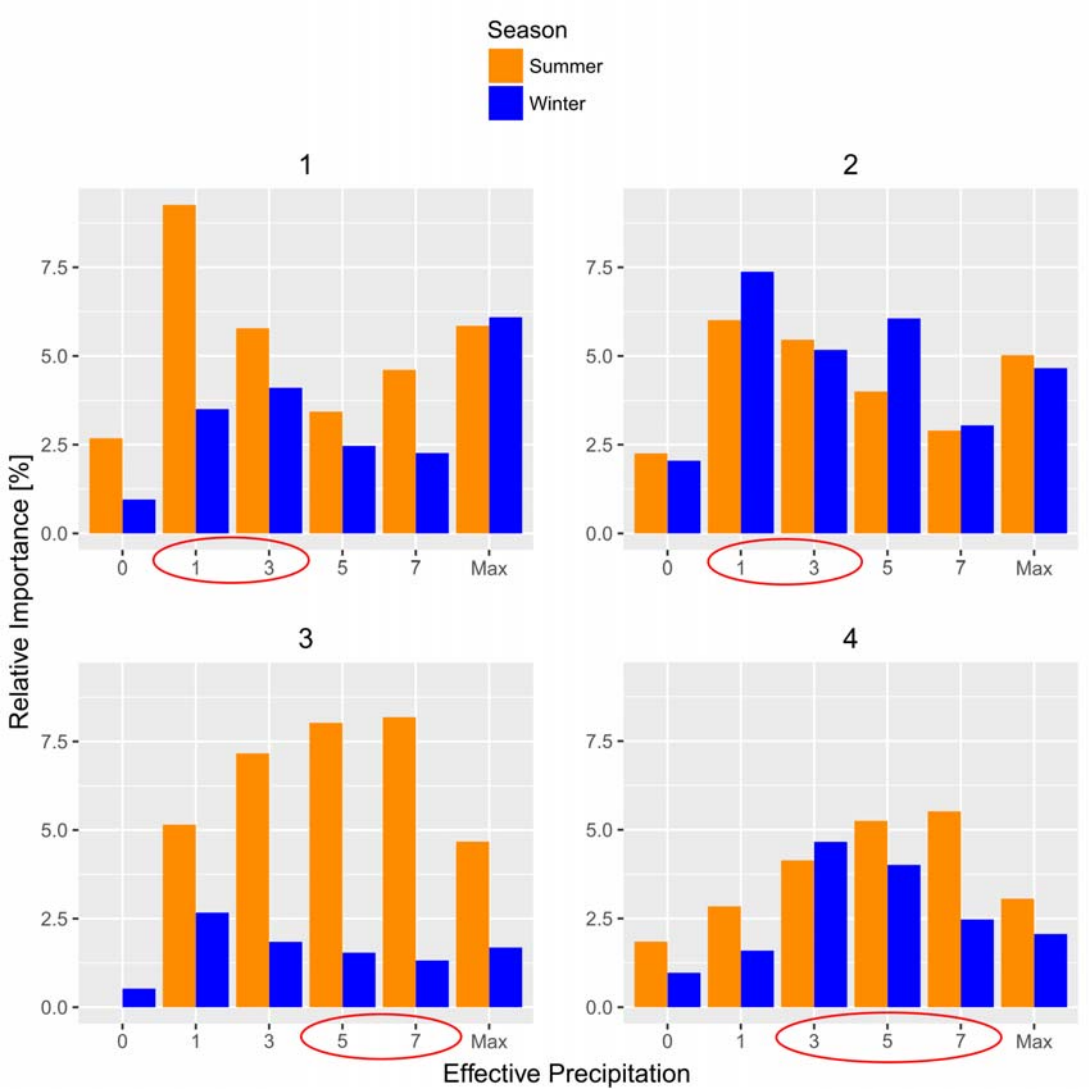


Figure 3.13: Variable importance of P_{eff} by region, season and time interval of preconditions, Δt. The red circles depict time intervals of high variable importance. Intervals that are not shown were excluded from the model by feature selection.

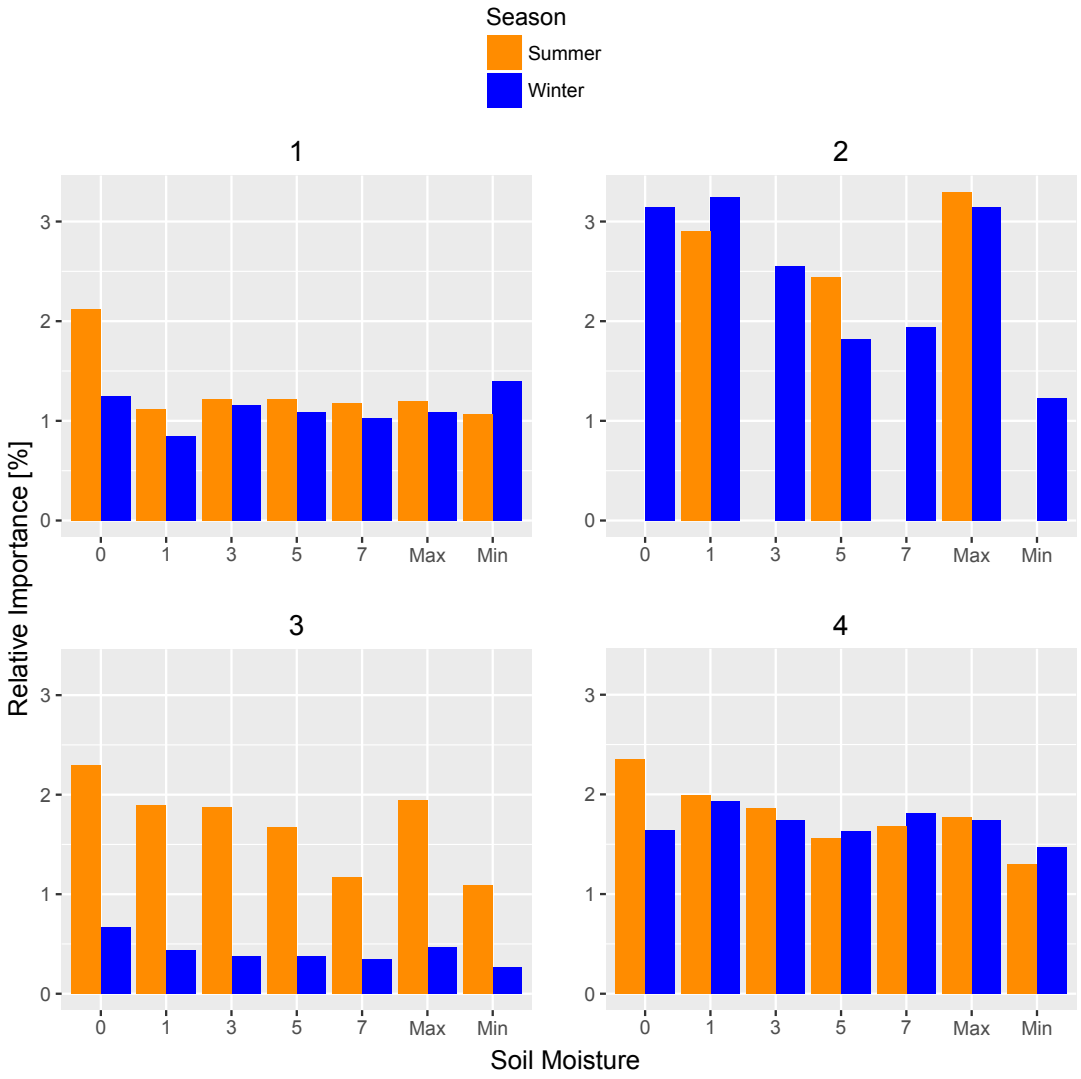


Figure 3.14: Variable importance of SM by region, season and time interval of preconditions, , Δt. Intervals that are not shown were excluded from the model by feature selection.

3.6.3 Static Variables

Generally, static variables get assigned high importance, lower in summer than in winter (fig. 3.11). On regional scale, it shows that this seasonal difference is reflected in 3 of 4 regions (fig. 3.12). In both RF and GLM, static variables get assigned higher importance from south to north.
AI dominates in all regions with highest values in the southernmost region 1 and lower ones in the north (fig. 3.15). A similar pattern can be observed for P_{ann}. This trend is supported by GLM insofar as it attributes highest importance to AI and P_{ann}, as well.
Second highest importances were attributed to $Slope$ and $ChSlope$, especially in the north. This is supported by GLM, as well. From here, the pattern of importances across the regions is more heterogeneous: $Altitude$ has a considerably high importance in region 1. In regions 2 and 4, $Area$ is assigned high importance. As to land-cover derivatives, region 1 shows high values for $Impervious$, region 4 for $Forest$ and $Permeable$. Generally, the estimates of catchment geometry are attributed low influence. Only in region 3 does DD have an influence, in region 4 DD and $FLMax$.
So, generally speaking, it can be observed that the number of relevant static variables increases from south to north.

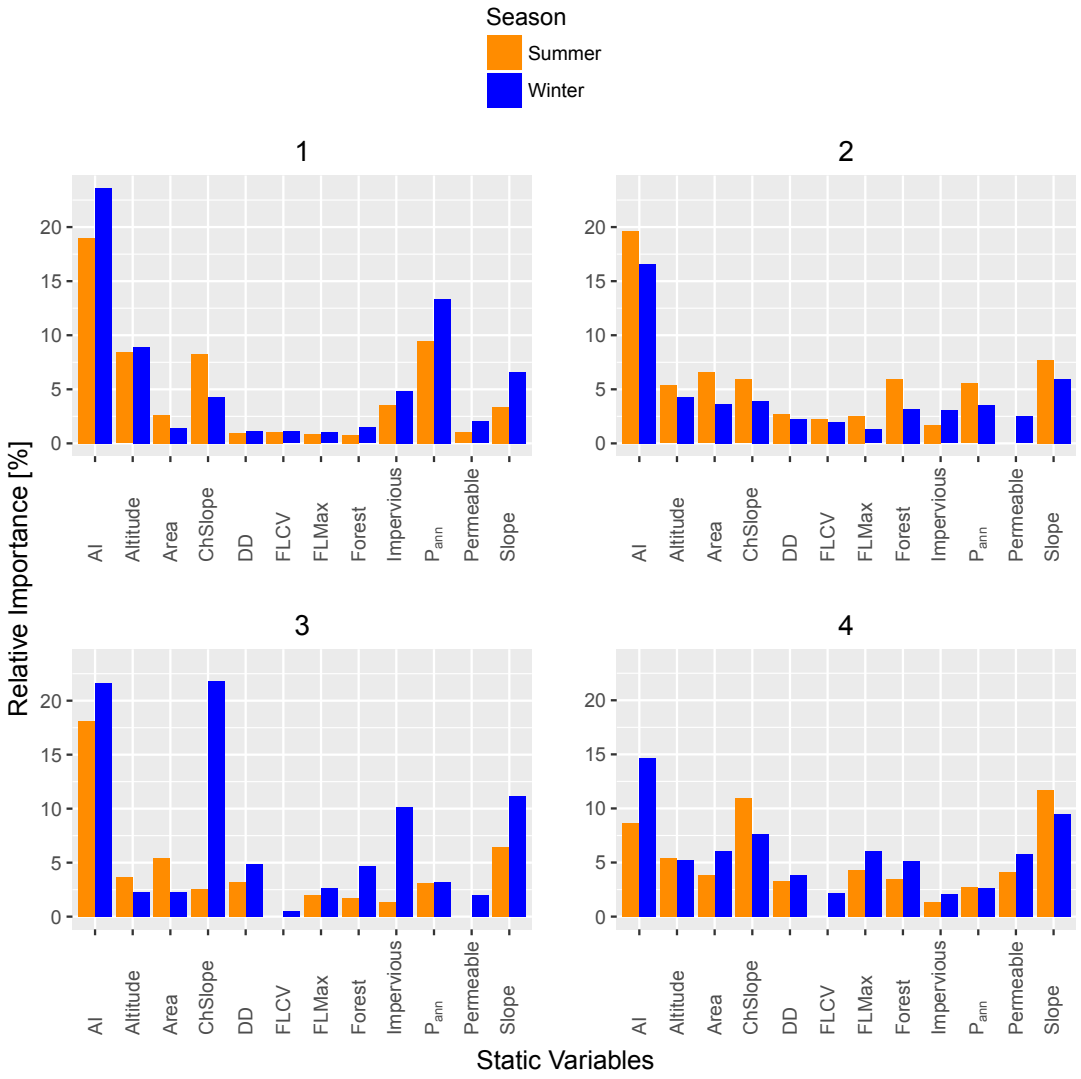


Figure 3.15: Variable importance of static variables by region and season. Variables that are not shown were excluded from the model by feature selection.

3.7 Partial Dependence Plots

The following section presents the results of partial dependence plot analysis (PDP). Thus, they depict the partial response of modeled flood magnitude, $\hat{Q}_f$, on the respective predictor as it was fitted in the models. Only the relevant plots are shown, the complete set can be found in the appendix. PDPs of each regional-seasonal model are standardized to the respective mean of observed Q_f – Giving the relative deviation from mean in %.

3.7.1 Dynamic Variables

Effective Precipitation:
For effective precipitation, only the PDPs of P_{eff1} are displayed (fig. 3.16) as it is representative of $P_{eff0...7}$. Though, the curves are less steep the higher Δt. A positive response of $\hat{Q}_f$ to P_{eff} can be seen in all regions. In summer, $\hat{Q}_f$ increases slower in low values of P_{eff}, but is stronger at high values. In two regions, the domain of P_{eff} extents towards higher values in summer and $\hat{Q}_f$ exhibits a sharp increase.

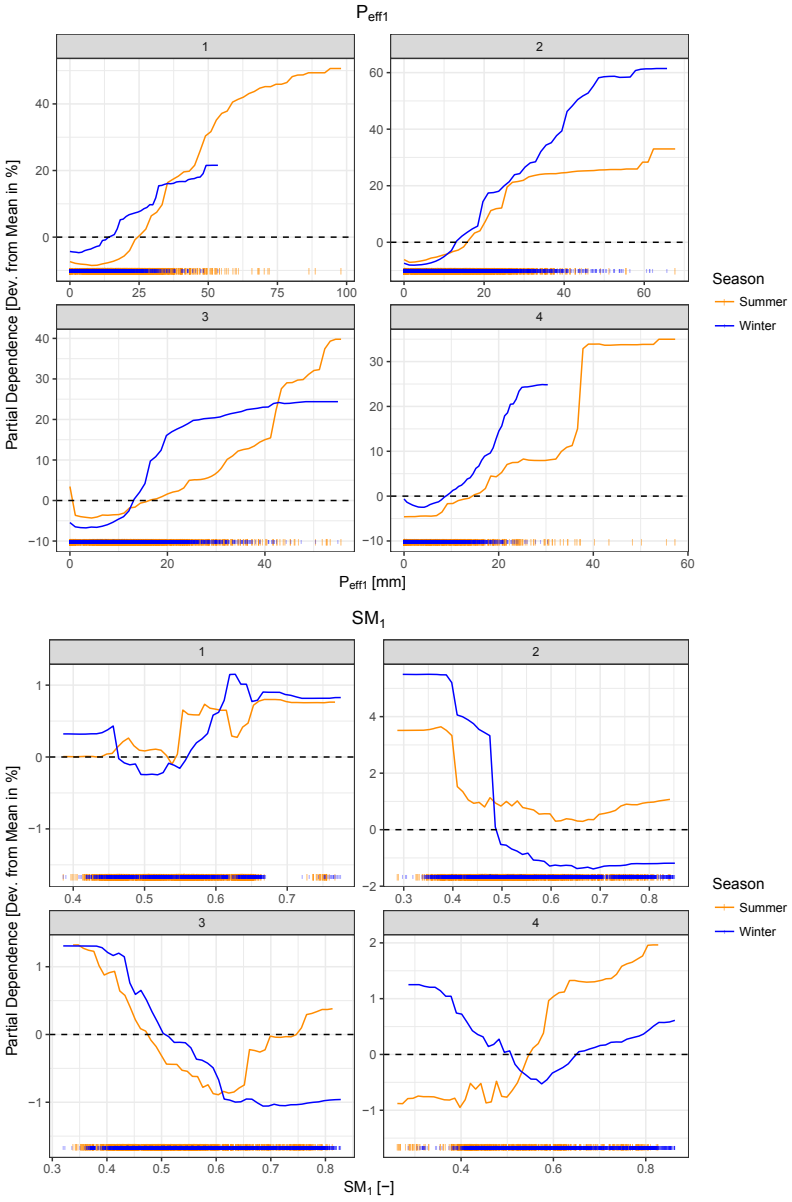


Figure 3.16: Relative partial dependence of $\hat{Q}_f$ on P_{eff1} (top) and $SM1$ (bottom) by region and season. Tick marks display observed values in training data.

Soil Moisture:
The partial dependence of $\hat{Q}_f$ on soil moisture is diverse (fig. 3.16). Once again, SM_1 is representative of $SM_{0...7}$ but more pronounced. While two regions exhibit a negative functional relationship both in summer and winter, the one in region 1 is consistently positive. Region 4 shows both a positive and a negative relationship, depending on the season.
Temperature:
Partial dependence of $\hat{Q}_f$ on T_1 is displayed in figure 3.17. For winter, it is representative of $T_{0...7}$: High values of $\hat{Q}_f$ occur at $T < 0°$and $T > 5°$. In summer, there is an inconsistent pattern. Comparison between the region is not possible as in all regional-seasonal models either one or both of winter and summer components were excluded.

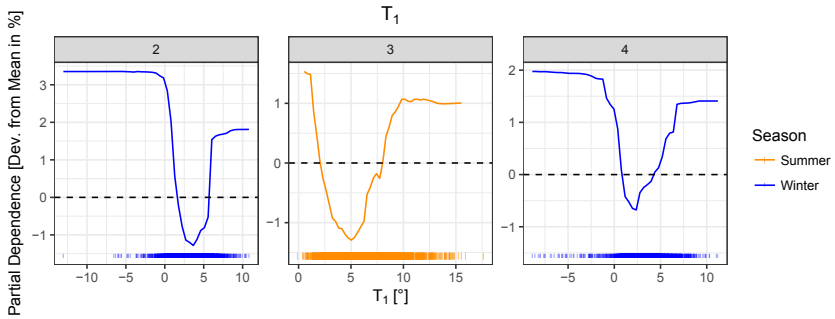


Figure 3.17: Relative partial dependence of $\hat{Q}_f$ on T_1. Tick marks display observed values in training data. If no curve is shown, T_1 was excluded from the model by feature selection.

3.7.2 Static Variables

As the partial dependence plots of static variables vary less by region and season than the ones of dynamic variables, figure 3.18 displays these in a more compact manner: In each plot, two regions are displayed. As static variables vary in domain, the x-axis of these plots was scaled to $[0, 1]$ for comparability. If the functional relationship of two variables is similar, only one of the two is shown.
Partial dependence of $\hat{Q}_f$ on AI (top-left) is positive and, apart from region 2, all regions show a sharp step-like increase of response values over values of AI. PDPs of P_{ann} are similar to this. For $Slope$ (top-right), as well, PDPs show a positive functional relationship with step-like increases. The latter are more distinct, so that instead of a continuous increase only two values of $\hat{Q}_f$ are predicted in regions 1 and 2. A similar pattern is seen for $CHSlope$. Partial dependence on $Area$ (centre-left) gives extremely high values of $\hat{Q}_f$ below a threshold in catchment area in all regions. Above that threshold, $\hat{Q}_f$ drops sharply to a constant level. In unscaled values, this threshold is at ≈ 300 km^2. $FLMax$ exhibits a similar relationship. DD (centre-right) exhibits a linear increase with a sharp increase at the upper end of its domain. As to land-cover estimates, only the ones of highest importance are displayed – i.e. $Forest$ and $Permeable$. $Forest$ (bottom-left) exhibits a non-linear negative relationship with a "re-rise" in some of the regions at high values of $Forest$. $Permebable$ (bottom-right) shows a non-linear positive relationship.

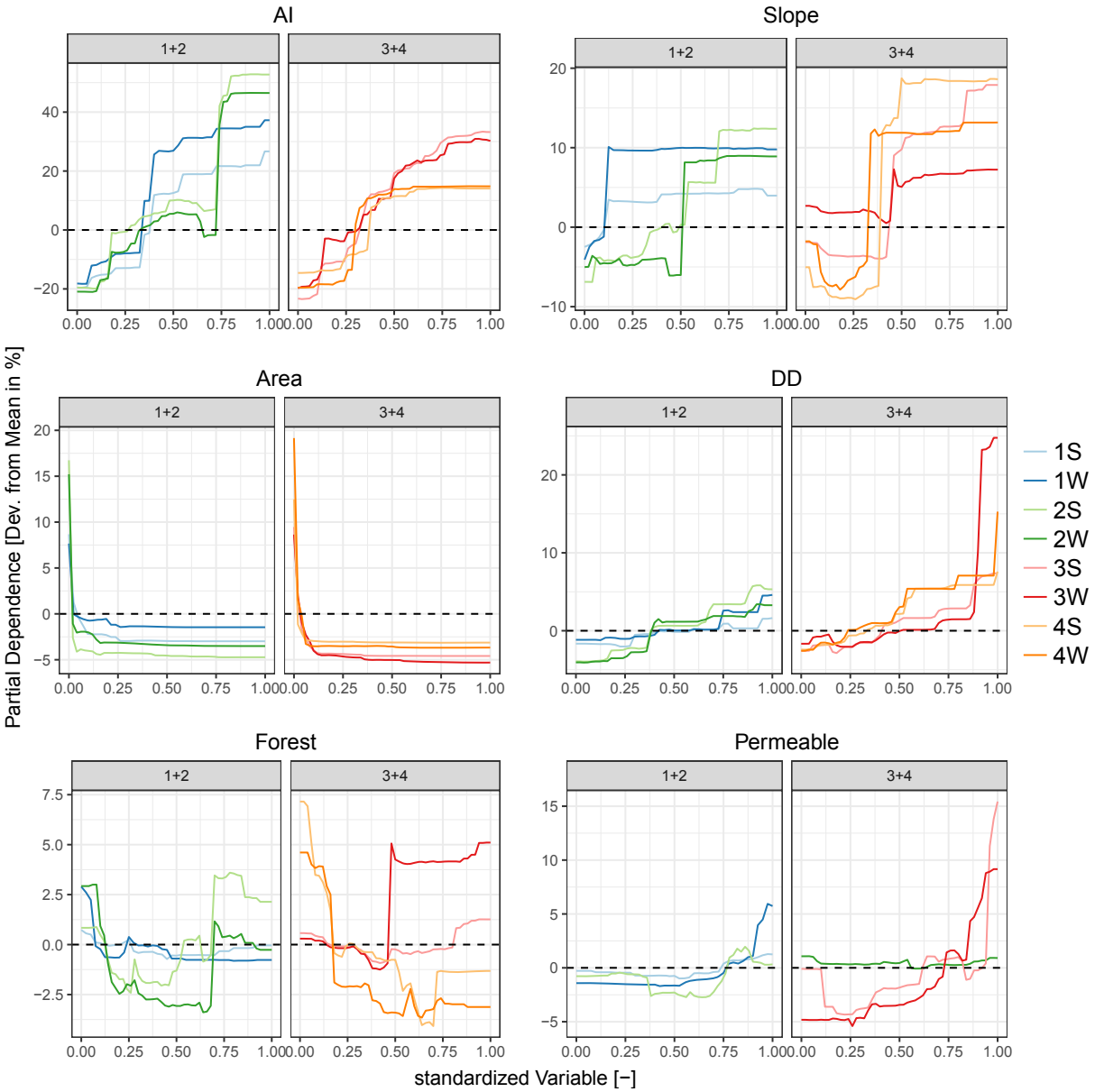


Figure 3.18: Relative partial dependence of $\hat{Q}_f$ on *AI*, *Slope*, *Area*, *DD*, *Permeable* and *Forest*. In each subplot, regions 1+2 are on the left, 3+4 on the right. X-axis was scaled to $[0, 1]$ for comparability. "1S" and "1W" refer to the summer and winter models of region 1, respectively.

Chapter 4

Discussion

In this chapter, the synthesis of all results as presented in the previous chapter is performed: Analysis of the dataset and inherent collinearity, as well as variable importances and partial dependences as obtained from modeling. Doing this, the research objectives as stated in chapter 1.2 are revived: The identification of factors that control flood magnitudes and the analysis of mechanisms of flood generation by region and season. Also, the performance of both GLM and RF is evaluated by region and season, thus comparing their suitability as tools to analyze hydrological extremes.

4.1 Controls of flood magnitude

As mentioned in the first chapter, flood generation can be divided into three processes: Runoff generation, runoff concentration and flood routing. The following section presents this study‘s results with respect to these flood generation processes, focusing on the role of each factor and its seasonality. The second part merges these results on a regional scale, presenting differences in flood generation among the regions. As RFf clearly outperformed GLM in almost all regions, interpretation of results is confined to RF. Results of GLM are only mentioned if they describe clear trends that were detected by both models. Also, accuracy of the summer model of region 3 was not sufficient to allow for full interpretation. Therefore, results of this model were included to a limited extent, only.

4.1.1 Runoff Generation

Control Variables:
In both models of all regions and seasons, Aridity Index AI was identified as the variable of highest influence on flood magnitude Q_f (fig. 3.15). Closely related, annual precipitation P_{ann}, was attributed high importance as well. This validates the modeling approach insofar as both were included as control variables to account for general wetness conditions between catchments. This was done to make sure that any explanatory power attributed to preconditions does not refer to its static component, i.e. spatial variability in average value, but only to its dynamic component of temporal variability in each catchment. As the control variables were actually assigned high importance, this was achieved. The results of PDP-analysis further validate the modeling approach as a positive functional relationship between Q_f and AI/P_{ann} was detected (fig.3.18). The non-linearity of this relationship can be linked to results of a study by Merz and Blöschl (2009*b*) that found a positive correlation of runoff coefficients and annual precipitation: If the share of direct runoff increases with increasing annual precipitation, this adds to give a non-linear increase in flood magnitude. The fact that AI serves as a better control variable than P_{ann}, however, is remarkable as AI is based on model estimates of PET by *mHm* – Itself, based on an empirical formula. As the name suggests, this estimate has mainly been used for classification in arid regions. Using it as a control variable for streamflow extremes could serve as a useful approach for future studies.
Effective Precipitation:
By introducing a control variable, the average wetness was accounted for – So, when comparing mean variable importances of dynamic and static variables in the following paragraph, AI and P_{ann} are excluded. As mentioned in chapter 1.1.1, dynamic variables influence flood generation only, whereas static variables affect both runoff generation and concentration. In this study, P_{eff} and static variables were assigned similar importance, followed by SM and T. All dynamic variables have a higher influence in summer than in winter (fig. 3.11), strongest for P_{eff}. This can be explained in the following way:

The hydrological system is more dynamic in summer: In figure 4.1 it can be seen that summer rainfall preconditions are higher in average values and exhibit a stronger variability. This points at the convective nature of rainfall in summer. Convective events are known to exhibit high intensity, so the link of precipitation to flood magnitude is more pronounced: As precipitation event duration is shorter, less water can infiltrate. Under extreme conditions, precipitation intensity exceeds infiltration capacity so that "Hortonian surface runoff" is generated. The change in response due to convective events is also reflected in variable importances and PDPs of precipitation one day before the flood event, P_{eff1} (fig. 3.16). The domain of P_{eff1} extends towards high values in summer, leading to a response of Q_f that is considerably stronger than in winter – So, precipitation intensity prior to the flood event gets higher in summer due to convective events, thus leading to a non-linear increase in corresponding flood magnitude.
In winter, on the other hand, figure 4.1 illustrates that average soil moisture is higher so saturation is more likely to occur. So, depending on intensity of the precipitation event and preconditions, a weak or a strong precipitation event might lead to a flood of high magnitude. However, representation of preconditions in soil moisture in the models seems to be limited, which showed in PDPs that exhibited various relationships without a structural pattern and fairly little importance that was attributed. As the interaction of soil moisture and precipitation is not fully represented, the signal of precipitation itself is blurred, thus being attributed lower importance in winter.
In winter, snow accumulation and melt make up a large part of the water budget – P_{eff}, as derived from observed P by mHM, does include these processes. The observed increase in model accuracy by including P_{eff} instead of P was to be expected. However, accuracy improved less than anticipated when considering the large influence snow melt actually has as shown by several studies, among these Petrow et al. (2007) and Sui and Koehler (2001). This can be explained by the following: The estimation of the snow-related part of runoff generation is known to be one of the most challenging tasks in hydrological modeling due to high spatial variability, data scarcity and highly non-linear interactions of snow cover, precipitation and temperature (Blöschl et al., 2013) – As the representation of snow in mHm and, thus, in P_{eff} is of rather basic nature, it does not fully capture these processes.
Soil Moisture: Generally, both RF and GLM attribute minor importance to soil moisture and the functional relationships only allow for limited interpretation: High flood magnitudes at low soil moisture may occur, especially during flash floods. However, it is unlikely that these events occur often enough to show up in almost all PDPs. Only in region 1, a consistent positive functional relationship was detected in both seasons (fig.3.16).
In fact, from data analysis a different picture with higher importances and a pronounced seasonality was to be expected: As shown in figure 4.1, soil moisture is lower and more variable in summer - Dry soils with correspondingly high infiltration rates may attenuate flood peaks. Still, wet conditions may lead to saturation excess surface runoff, like in winter. This way, soil moisture would have had explanatory power over its whole domain. Winter soil moisture is less variable. As saturation is more likely to occur, it still exerts a considerable control of flood magnitude. But, once soil is frozen, its signal is more or less extinguished: All rainfall translates into direct runoff and soil moisture does not have any explanatory power.
As a pronounced seasonality could only be observed in the region of lowest model accuracy, this hypothesis could not be supported. However, the lack of explanatory power of soil moisture at the scale of this study has been reported before. While Schröter et al. (2015) detected a major influence of soil moisture when analyzing one single flood event of the Elbe, Nied et al. (2013) performed a catchment-wide analysis of seasonal soil moisture patterns and already at catchment scale, results were more heterogeneous. While the probability of flood initiation proved to be higher in the case of high soil moisture catchment-wide, the spatial distribution and seasonality of the respective soil moisture patterns altered the link to flood initiation so that the respective patterns could only be ranked by their probability of flood initiation. Moving up in scale, a study by Uhlenbrook et al. (2002) on 39 catchments in southern Germany applied a similar regression approach as the study at hand and detected very low to no influence of preconditions in soil moisture. Thus, the study at hand confirms that soil moisture has only limited control over flood magnitudes in a macro-scale model as soil-moisture-precipitation interaction is too complex on event and catchment scale to be captured. The main loss of information at macro-scale is the one regarding spatial distribution of soil moisture: This is known to be highly variable within a catchment (Fohrer et al., 2016) and the importance of it for flood generation is highlighted in several studies (e.g. Merz and Plate, 1997; Nied et al., 2013). Also, a study at this scale has to rely on simulated soil moisture data that is likely to be subject to considerable uncertainty (Samaniego et al., 2013).
Generally, no trends were detected regarding the time intervals in preconditions of SM. This is due to the fact that soil moisture estimates proved to be highly collinear – major changes in soil moisture can only be observed over longer time periods than the one of preconditions as sampled in this study. The only signal that was captured was a high importance of SM_0 in summer, which is essentially a time-lagged

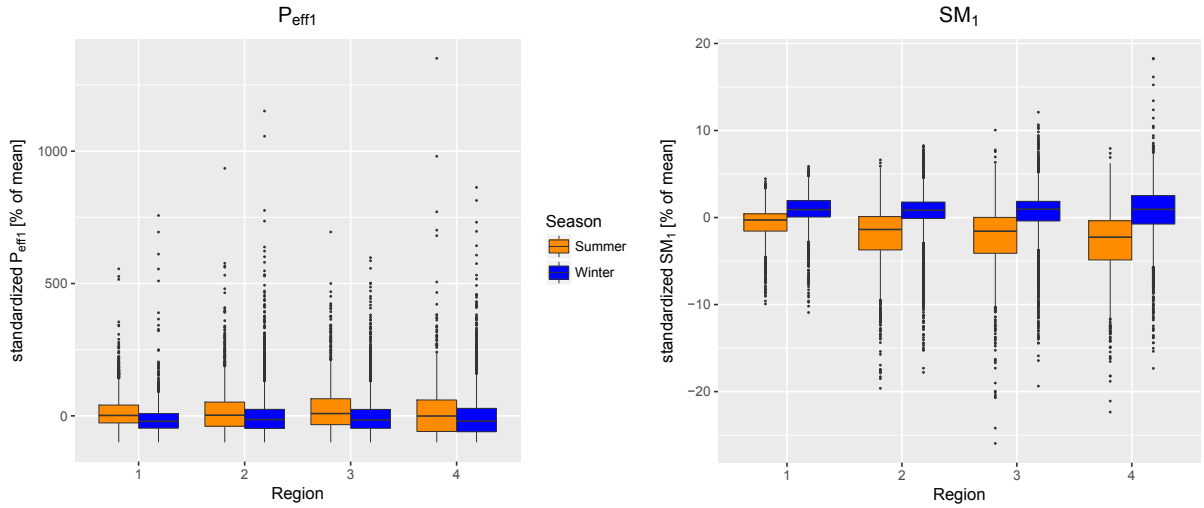


Figure 4.1: Seasonal boxplots of P_{eff1} (left) and SM_1 (right), standardized to each stations‘ mean. These are representative of $P_{eff0...7}$ and $SM_{0...7}$.

signal of P_{eff1} that was attributed higher importance in summer as illustrated above.

Temperature:

Temperature was assigned little importance, in the southern regions only in winter, in the northern regions only in summer. In winter, PDPs indicate plausible functional relationships (fig. 3.17): Comparatively high flood magnitudes at $T < 0°$and $T > 5°$that can be linked to frozen soils and snow melt conditions. As shown in dataset analysis, these are the regions of highest average catchment altitude. It is only plausible that snow and freezing of soils have an influence only in these as catchments in the northern regions lie below 800m. Functional relationships in summer, however, could not be explained from a hydrological perspective.

Static Variables:

As mentioned above, static variables were attributed about as much influence as effective precipitation, lower in summer. Except for land cover, these variables do not underly seasonal changes so the difference between summer and winter is only attributed to the fact that the hydrological system responds stronger to dynamic components in summer, leaving less explanatory power to static variables. This is supported by the fact that functional relationships of these, as visualized in PDPs, exhibit only minor variations between summer and winter.

Of all static variables, only *Slope* and land cover estimates are relevant for runoff generation. *Slope* was attributed highest importance of all static variables (excluding control variables) and PDPs exhibit a positive functional relationship in all regions. While research results are contradictory about the effect of slope on infiltration rates (Morbidelli et al., 2018), it is common knowledge that, regarded as one component, inter- and surface flow increase with increasing slope (Chifflard, 2006). Thus, the results of this study go along with other research.

Infiltration rates are known to be higher in forest environments as a result of retardation of raindrops, decreased rainfall intensity by interception and a higher density of micro- and macropores (Fohrer et al., 2016) – Thus, reducing direct runoff. Open areas, on the other hand, exhibit relatively high surface runoff. Both can be seen in the functional relationships in region 4, where high importance was given to both *Forest* and *Permeable* (fig. 3.18): Predicted flood magnitude decreases with share of forest and increases with share of *Permeable*. Similar results were obtained by Samaniego and Bárdossy (2007). A seasonal signal as a result of different degrees of vegetation cover did not show in PDP analysis. In region 1, "Impervious" is attributed high importance. This finding is consistent with hydrological reasoning as on impervious surfaces at steep slopes, all precipitation is transformed into direct runoff. However, this relationship could not be verified by PDPs.

4.1.2 Runoff Concentration

Slope and land cover are not only relevant for runoff generation, but also for runoff concentration: The higher the slope the lower retention, leading to faster runoff concentration. This results in a more direct translation of precipitation events into flood events as peaks in precipitation are attenuated fairly little on their way downslope. This is further acerbated by high erosive power of surface flow that

creates preferential flowpaths such as gullies. These, in turn, increase flow velocity which increases runoff concentration speed. So, the positive relationship of flood magnitude and *Slope* as detected by PDPs is supported once more. Similar observations have been made by Uhlenbrook et al. (2002) and Samaniego-Eguiguren (2003), amongst others. However, all regions show a sharp, steplike increase at a certain value of *Slope* – this critical threshold value could not be verified by literature research.
Land cover affects runoff concentration in an indirect manner: The process of interception in forest stretches a precipitation peak in time, thus leading to a higher retention of the respective catchment. In open areas, interception is lower, so lower retention leads to a faster runoff concentration, i.e. a more direct streamflow response to precipitation events. So, once more, negative (*Forest*) and positive (*Permeable*) functional relationships are meaningful.
Measures of catchment geometry *Area*, channel slope *ChSlope*, maximum flow length $FLMax$ and drainage density DD were attributed considerable importance in the northern regions. For all regions, *Area* exhibits high flood magnitude for catchment area $<\approx 300km^2$, followed by a very sharp drop to a constant level (fig. 3.18). Generally, a negative relationship is consistent with literature (Pfaundler, 2001; Merz and Blöschl, 2009*a*): In large catchments, flow path lengths span a wider range than in small ones. When a precipitation event of duration d occurs, the areas closest to the outlet drain fastest, response of the areas that are further away is time-lagged. So it is likely that the travel time along the longest flow path in the catchment is longer than precipitation duration, so $FLMax > d$. In this case, there is no point in time in which all areas drain at the same time, thus adding up in peak response. However, this is the case in small catchments. Adding to this, it is more likely for small catchments that the spatial extent of a precipitation event of high intensity covers the whole catchment area. For this reason, Patt and Jüpner (2001) defines two different flood generation mechanisms for catchments smaller and larger than 100 km^2 – Which reflects in the highly non-linear, negative relationship of Q_f to *Area* as detected by the models. Quite logically, $FLMax$ shows a similar pattern.
Drainage density is known to have a positive functional relationship with flood magnitude: The more streams there are, the faster is runoff concentration and the more homogeneous are flow lengths, so that drainage of different areas is more likely to occur at the same time (Pallard et al., 2009). This is exactly what PDPs indicate: A more or less linear positive functional relationship to flood magnitude.
The estimate of catchment shape, homogeneity of flow lengths, $FLCV$, does not have high explanatory power. Probably, a more commonly used estimate like the "shape factor R" (Dyck and Peschke, 1995) is a better representation of catchment shape and should be implemented in future extensions of this study.

4.1.3 Flood Routing

Due to limitations in data availability and scope of this study, flood routing could only be accounted for by including an estimate of channel slope in the analysis. It is commonly regarded as an important control of flood routing (Merz and Blöschl, 2009*a*; Patt and Jüpner, 2001) – The lower channel slope, the more attenuation of the flood wave occurs on its way downstream. So, the models agree to common knowledge as they show higher flood magnitudes at higher channel slopes. However, channel slope and topographic slope seem to carry a similar signal in importances and PDPs. Also, average collinearity is high at $r^2 = 0.8$ (fig. 3.9). Also, all measures of catchment geometry are likely to be subject to a considerable error, which is discussed in the following paragraphs. Even if the estimate of channel slope was exact, there are multiple other factors that influence flood routing like channel roughness, morphology, tributaries, the presence of lakes and tributaries and anthropogenic influence. This makes it unlikely that a distinct signal of channel slope would be captured in the models. Thus, the estimate of channel slope is regarded as a duplicate of topographic slope.

4.2 Regional differences

The following section merges the results of dataset analysis and modeling across the regions to identify major flood controls in the different regions. These are displayed in figure 4.2, once more.

In dataset analysis, a clear south-to north gradient was detected for several variables: Flood magnitude, average wetness and pre-event effective precipitation decrease from south to north. Also, topography gets less pronounced towards the north with lower average altitude, slope and drainage density. In opposite direction, average catchment area increases from south to north. Larger catchments of low slope and drainage density are known to exhibit high retention, resulting in a smoother hydrograph and longer response time than small catchments of strong topography (Patt and Jüpner, 2001). Consequently, a pattern of "flashiness" in response was expected to be observed. In flood duration, this could be seen as average flood duration almost doubles from south to north. Differences in response time were observed in the distribution of importances across the time intervals of effective precipitation (fig. 3.13): In the southernmost region 1, variable importances of effective precipitation were highest at $\Delta t = [1, 3]$, i.e. one to three days prior to the flood event. Moving north, highest importances moved back in time intervals, so that in region 4 these are highest at $\Delta t = [5, 7]$. Apart from hydrological reasoning, these results could be validated by a study conducted by Uhlenbrook et al. (2002), who identified a response time of 3 days for 29 catchments in south-west Germany. These would belong to region 2 of this study – in which a response time of 1 to 3 days was identified, too.

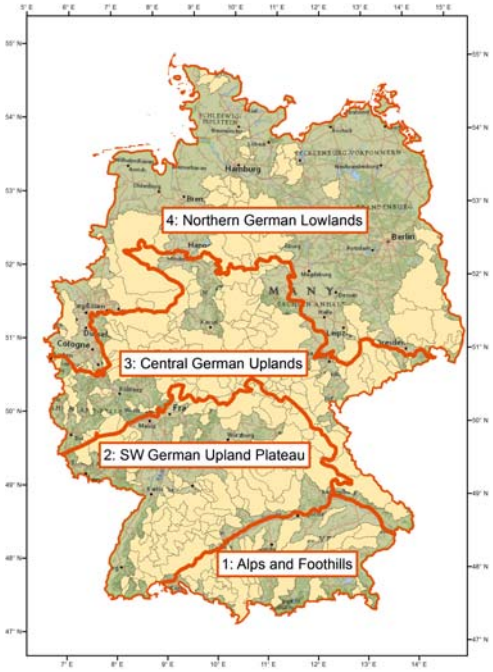


Figure 4.2: Map of the study regions, duplicate.

The following paragraph presents each region separately. In general, a gradient in system complexity was observed: While the southernmost region is largely controlled by topography and dynamics in effective precipitation, land use and catchment geometry play a role as well in the central regions. In the northernmost region, streamflow response is more indirect, so the influence of all static variables is higher and the dynamic component has a lower influence.

Region 1:

Of the catchments in this region, few are in alpine, the majority are located in pre-alpine environments, these are distinctly different in altitude and slope (see boxplots and maps in appendix). Models give a strong weight to this differentiation by assigning high importance to altitude and wetness in form of aridity index and annual precipitation. Naturally, these variables are correlated as shown in collinearity analysis (appendix) and in literature (e.g. Häckel, 2016). The fact that the models give weight to these variables indicates that there is a distinct difference in average flood magnitude between alpine and sub-alpine environments. This was also observed in the map of flood magnitude (fig. 3.4), where high values cluster in the alpine region. Related to this is the high importance of slope and channel slope – As this region comprises catchments of two different topographic types, the model accounts for this by giving weight to slope estimates. The functional relationship in PDPs of all above variables indicate this regime-shift in form of a sharp step-like increase.

As catchments in this region exhibit a flashy response in general, extreme rainfall events have a strong control over flood magnitude. These occur mostly in summer and intensity is high in mountainous regions (Häckel, 2016). This reflects in an remarkably strong increase of importance of P_1 in summer: If convective events occur, these are likely to generate a flood within a day. Soil moisture is attributed fairly low importance in comparison to other regions – As stated by Chifflard (2006), at slopes of >6°, runoff generation is governed by slope instead of soil moisture. So, as there are several catchments in this region of very high slope, soil moisture has less control over flood magnitude in this region than in the other regions – As observed in variable importances. As mentioned before, temperature controls flood magnitude in this region only in winter, as a proxy for snow melt and frozen soils and a resulting increase in runoff generation.

Region 2

As this region comprises catchments that are both mountainous and rather flat, major control is exerted

by topographic characteristics altitude, slope and channel slope. However, land-use and catchment area get assigned explanatory power as well: While catchments are comparatively small, there are a few catchments of large sizes that are accounted for in the models. Land cover of the catchments varies strongly: The area that is covered either by forest or permeable surface varies between 20-80%, respectively. Thus, land cover is a major control in this region, as indicated by variable importance. Even though variable in topography, this region can be regarded as homogeneous in hydrological response characteristics: Catchments that exhibit a flat topography are mainly located in the regions of Swabian and Franconian Jura – Thus, their streamflow response is governed by karst processes and response times are comparatively short.

Region 3

This region is similar to region 2 in several respects: The comprised catchments are variable in topographic conditions like slope and channel slope, catchment area, geometry and land cover. Only catchment altitude is lower and less variable. In that sense, this region represents the transition from low mountain ranges as in region 2 to the lowlands further north. Thus, catchments appear to vary strongly in their flood characteristics, which shows in a density distribution of duration that is already very similar to the one of region 4 (fig. 3.8). Therefore, links between predictors and flood magnitude are heterogeneous which leads to models of low accuracy. Also, variable importances are patchy, indicating a strong seasonality and different response times in summer and winter. As this could not be explained from a hydrological perspective, results of this region were only included in the analysis to a limited extent and should be investigated further.

Region 4

In region 4, topography varies the least, which leads to a smoother hydrograph and longer response times. This implies that the translation of a precipitation event into a flood is less direct than in the other regions, increasing the influence of static catchment characteristics that influence runoff generation and concentration. This can be observed in variable importance (fig.3.12). Almost all variables are attributed considerable control over flood magnitude, strongest are slope, channel slope and area. The latter is particularly meaningful as catchment area varies most in this region. So, excluding region 2, this region is considered as the most complex system which is supported by the fact that models performed considerably worse than in the southern regions.

4.3 Model Accuracy

In general, prediction of flood data is known to be a challenging task as it follows a highly skewed cumulative density function, exhibits heteroscedascity and flood peak measurements are associated with considerable errors (Samaniego and Bárdossy, 2007; Merz and Blöschl, 2009*b*). Considering this, an average model accuracy of $R^2 > 0.7$ was taken as a threshold for an acceptable Goodness-of-fit and suitability for interpretation of results. This was achieved by RF. Of the regional-seasonal models, only the summer models of regions 3 and 4 were below this threshold. GLM performed considerably worse, reaching an $R^2 > 0.7$ only for two regional-seasonal models. Consequently, interpretation of results was limited to RF. Generally, model accuracy was highest in the south in winter, decreasing towards the north and in summer. This can be explained by the complexity of the respective hydrological system: In southern regions, the signal of precipitation is transformed into runoff response more or less directly. In the north, the signal is lagged in time and attenuated, thus harder to predict. In particular, this can be seen in summer: Dry soils attenuate the response in streamflow. As soil moisture proved not to be adequately represented in the model, accuracy decreased in summer, particularly in the more complex system in the north. Southern regions proved to be influenced mainly by topography, so summer models did not suffer from deficiencies in soil moisture estimation that much. Rather the opposite: In summer, snow melt is non-existent (region 2) or at a constant rate (region 1), so the inherent exactitudes in snow modeling do not have that big of an impact. Lower accuracies in summer are also related to sample size: In all regions but region 1, only 20% of floods occurred in summer – having a stronger effect in the northernmost region where average spatial coverage was low to start with.

So overall, Machine-Learning modeling clearly outperformed the parametric approach, especially in the regions of higher complexity. As shown, non-linear relationships play a role in several factors and were fit accordingly by RF. In GLM, these can not be implemented. This proves RandomForest to be suited better for analysis of runoff extremes, thus giving an answer to one of the major research objectives of this study.

Model accuracy during calibration and validation, i.e. on training and test data, was similar. This indicates robust models that did not overfit to the training data. As displayed in residual plots (fig. 3.10), most models show a satisfying accuracy in low and medium values. Towards the upper end of the

value range, uncertainty bounds get wider and all models exhibit heteroscedascity that gets stronger the worse model accuracy is. In order to examine the causes for heteroscedascity, the gauging stations that exhibited the highest 2% of each models‘ residuals are displayed in figure 4.3. These strong residuals cluster in mountainous regions, e.g. the Alps, the Black Forest and the Thuringian Forest. Thus, it can be concluded that the flood-hydrograph of catchments in the areas of high altitude deviates from the ones of other catchments in the respective region. This, in turn, can not be accurately represented in the models. Therefore, an altered sampling scheme that clusters catchments by altitude could help to reduce heteroscedascity.

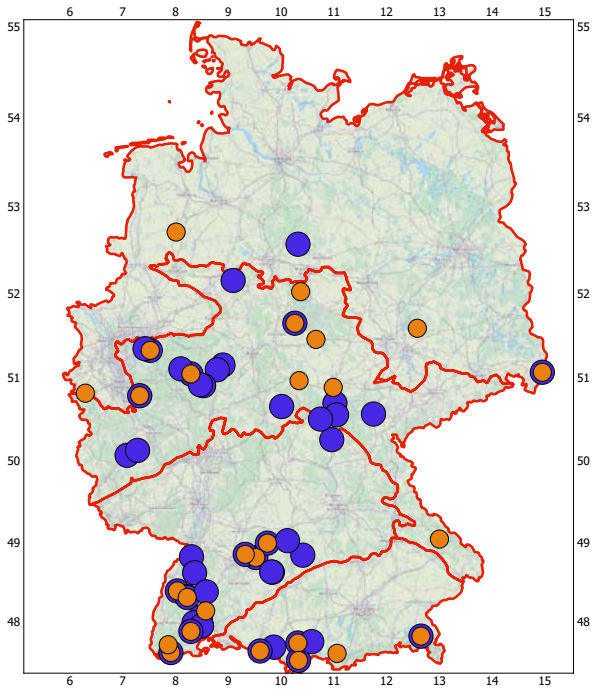


Figure 4.3: Map of stations that exhibited 98%-percentile residuals. Orange: summer, blue: winter.

Also, an effective implementation of soil moisture could help capturing events of high soil moisture and high precipitation intensity that lead to floods of high magnitude.
Disregarding constraints in predictors and sampling scheme, there are several methodological methods to account for heteroscedascity (Samaniego and Bárdossy, 2007): (1) Introduce variable transformations, (2) use a generalized model and (3) apply a weighted objective function. (1) was applied in this study by fitting the same models on data that was scaled to $[0, 1]$ in each catchment, separately. This did not improve model accuracy. As to (2), lognormal- and gamma-distributions were fit, which did not lead to higher accuracy or less heteroscedascity either. Another option would be to fit one of the commonly-known extreme value distributions, i.e. Gumbel or GEV. Using option (3), i.e. weighting the objective function by flood magnitude might lead to better accuracy as well. Both options were too laborious in terms of computational implementation to be applied within the scope of this study.

4.4 Methodological Considerations

Sampling Schemes

Regional Subsets:
In general, sampling was done according to the 4 regions according to the classification of "natural regions". This proved as a successful hands-on approach to group catchments that can be assumed to exhibit similar flood response characteristics. The fact that flood frequency varies between winter and summer in all regions but region 1 agrees with a study by Beurton and Thieken (2009), who identified a multi-modal pattern of both winter and summer floods in the south and a dominance of winter floods in the north. However, it showed that sampling size of the resulting datasets were considerably different due to variability in spatial coverage and average flood duration. Thus, sample size did not reflect the area covered by the respective regions. Differences in sample sizes were one cause of low model accuracies in

the northern regions and should be corrected for in future extensions of this study. Model residuals were highest in mountainous areas of most regions (fig. 4.3), indicating the presence of a different hydrograph. Therefore, a possible improvement of this study could be to cluster the data not by spatial location but by topographic characteristics. This would also balance out the low spatial coverage in the northernmost region as catchments of low topography from other regions would be added into the same cluster.

Sampling of flood events:

As to sampling of flood events, the 98%-quantile threshold as applied in this study is more conservative than the ones applied in other studies, e.g. by Samaniego-Eguiguren (2003) who used 95% as a threshold. Thus, a more rigid separation of flood events was applied to ensure that captured flood events do represent actual floods in reality. Still, determining the right threshold has been proven to be of influence on model results (Bezak et al., 2014), so further investigation should be carried out. For this, tools are at hand, e.g. the analysis of "mean exceedance" or hypothesis testing as to whether flood records follow a Poisson distribution (Lang et al., 1999).

Sampling of Preconditions:

Soil moisture is highly variable in space and, as shown by Nied et al. (2013), its spatial distribution exerts a strong control on flood generation. As temporal variability in soil moisture proved to be too inert to give a signal at different time intervals, a classification approach of spatial soil moisture patterns as in the above study could perform better at representing the variable component of soil moisture.

Geoprocessing: The geometric predictors DD and $ChSlope$ are subject to inexactitudes in stream network delineation. A fairly generalized approach was applied to all catchments in the same manner, extracting the three highest Strahler stream orders. The approach is limited in correctly delineating stream networks with two main streams: If these do not merge in proximity of the outlet, this is not corrected for in Strahler orders. The identified mainstream is of negligible length and total river length is underestimated. These uncertainties propagate into the estimates of DD and $ChSlope$. Possible improvements of this approach are presented by (Tarboton et al., 1991) that proposes a method to validate the stream network by analyzing the distributions of both stream orders and stream slopes.

Interactions

Due to limitations in the scope of this study, investigation of possible interactions could only be executed in a shallow manner, not giving any indication of interactions between soil moisture and precipitation. From hydrological perspective, however, there is no doubt that SM and P interact (Nied et al., 2017), so it should be expected that this interaction also reflects in the respective statistical model. While the structure of GLM as applied in this study did not include interaction terms, RF does internally fit these where appropriate. In the case of substantial interactions, PDPs have been shown to be biased. After all, interactions are only averaged out – If these are strong, they might still distort the functional relationship (Goldstein et al., 2015). In this study, there were several instances where variable importances agreed with hydrological reasoning but PDPs did not. Most importantly, this was the case in soil moisture but also in land cover estimates $Forest$ and $Impervious$. Thus, potential interactions should be investigated – A method to do this would be the so-called "ICEplots", which fragment the averaged values of PDPs into its components as to marginal distributions of predictors (Goldstein et al., 2015). This way, both the functional relationship and the presence of interactions can be detected. Another method was introduced and termed "H-statistics" by Friedman et al. (2008).

Predictors

The majority of those variables that were listed as relevant for flood generation mechanisms in the first chapter were included in the analysis. Still, there some constraints to these that will be listed in this section.

As model accuracy only improved slightly by including a estimate of snow melt, this led to the conclusion that snow melt estimation may be limited in preciseness. Therefore, it could be helpful to include snow cover not as a part of effective precipitation but as a separate predictor. The current approach inhibits the ability of RF algorithm to account for uncertainties in snow modeling by fitting interactions: If snow cover was included as a separate predictor, RF could model the three-way interaction of snow cover, precipitation and temperature. This way, rain-on-snow events could be reproduced in the model, which are not accounted for in P_{eff}. In the current approach. only a two-way interaction of P_{eff} and T can be used to correct snow melt estimation.

The soil moisture estimate that was used is an average over the whole soil column. For representing saturation processes, the use of averages within the upper soil layers, only, might lead to better results.

For most of the static variables, inherent inaccuracies were already mentioned. However, for land cover it is to be noted that a finer categorization could be beneficial. In the current approach, "Permeable" comprises all land cover types that are not forest or imperious, thus differences in vegetation cover are not accounted for. As explained, vegetation cover does have an influence on both runoff generation and concentration by altered interception and infiltration, so a finer classification could improve model accuracy. Also, changes in land cover over the time period of time-series were not considered. These were found to control flood magnitude in small catchments rather than in large ones, but with a stronger effect at high flood magnitudes (Blöschl et al., 2007; Kuraś et al., 2012). Thus, an influence has been observed and should be considered.
As mentioned, flood propagation could only be implemented to a limited extent, so this should be extended in future implementations of this study, as well. This should include temporal changes to river morphology.

Chapter 5

Conclusion

Floods can pose a large-scale hazard to human life and property, as has been seen during two strong flooding events in 2002 and 2013 that covered large parts of Central Europe. While hydrological research in Germany has advanced in understanding the link between climatic controls and floods, the interplay of precipitation and catchment preconditions as well as the role of static catchment attributes has not yet been investigated at a large scale. This study closes that gap by analyzing more than 29.000 flood events at 373 gauging stations across Germany. A combined estimate of both precipitation and snow melt, soil moisture, temperature and static catchment attributes were analyzed as to the control they exert on flood magnitude. The analysis was run on four regions of homogeneous physiographic conditions, separately for summer and winter, in order to detect the variation of flood generation mechanisms in space and time. Generally, precipitation and static catchment attributes proved to be the dominant controls of flood magnitude. The strong influence of soil moisture on flood magnitude that has been reported in other studies could not be reproduced, which is attributed to the spatial scale of the study at hand. Still, a seasonal and regional pattern in preconditions was detected: It showed that the control of both precipitation and soil moisture in flood magnitude is more pronounced in summer than in winter. This could be explained by a higher variability of the hydrological system in summer – That is, convective precipitation events of higher intensity and soil moisture that varies over a greater range of wetness states. Also, a spatial gradient in complexity of flood generation mechanisms was identified: In the southern regions of pronounced topography, flood magnitude is mainly controlled by precipitation, snow melt and topographic conditions. As a result, catchments exhibit a "flashy" response of comparatively short durations and a quick flood response to precipitation events within one to three days. Towards the north, rainfall-runoff transformation becomes more complex as topography is less pronounced and average flood response time is higher, at five to seven days. Therefore, multiple catchment characteristics exert control on flood magnitude, among these land-cover, drainage density and catchment area. The majority of functional relationships between flood magnitude and relevant factors that were detected by RF agreed to the ones that are listed in literature. Therefore, common hydrological knowledge could be validated at a large scale, which was one of this study's objectives.

This study represents a novel approach insofar as a Machine-Learning algorithm, the RF, was applied in flood magnitude analysis for the first time. A traditional GLM served as a baseline to evaluate RF's performance. RF clearly outperformed GLM, especially in regions and seasons of high complexity. By providing measures of variable importance in combination with the ability to fit and visualize non-linear relationships, RF proved to be suitable for analysis of hydrological extremes. However, both algorithms exhibited heteroscedastic behavior that has to be addressed in further extensions of this study. Residual analysis indicated that grouping the data by topography instead of spatial coherence could help in addressing this issue. For hydrological research, the results imply that Machine-Learning techniques should be considered as an appropriate alternative to traditional statistics more often when analyzing continuous data of hydrological extremes. Also, the aridity index AI proved to be a helpful measure of average catchment wetness that could be used in related statistical applications where average catchment conditions have to be accounted for.

A possible future extension of this study is to include changes of snow cover as a separate variable to gain further insight on the control of preconditions in snow cover on flood magnitude. Including land-cover changes and variables that represent flood routing mechanisms would complete the approach towards an all-encompassing representation of flood generation processes from a statistical perspective.

Bibliography

Bárdossy, A., and F. Filiz. 2005. Identification of flood producing atmospheric circulation patterns. Journal of hydrology **313**:48–57.

Beurton, S., and A. H. Thieken. 2009. Seasonality of floods in Germany. Hydrological Sciences Journal **54**:62–76.

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

BKG. 2010. Federal Agency for Cartography and Geodesy: Digital Elevation Model (DEM) .

Blöschl, G., S. Ardoin-Bardin, M. Bonell, M. Dorninger, D. Goodrich, D. Gutknecht, D. Matamoros, B. Merz, P. Shand, and J. Szolgay. 2007. At what scales do climate variability and land cover change impact on flooding and low flows? Hydrological Processes **21**:1241–1247.

Blöschl, G., R. Kirnbauer, and D. Gutknecht. 1990. Modelling snowmelt in a mountainous river basin on an event basis. Journal of Hydrology **113**:207–229.

Blöschl, G., M. Sivapalan, H. Savenije, T. Wagener, and A. Viglione. 2013. Runoff prediction in ungauged basins: synthesis across processes, places and scales. Cambridge University Press.

Booker, D., and T. Snelder. 2012. Comparing methods for estimating flow duration curves at ungauged sites. Journal of Hydrology **434**:78 – 94.

Breiman, L. 2001. Random forests. Machine learning **45**:5–32.

Bronstert, A., H. Bormann, G. Bürger, U. Haberlandt, F. Hattermann, M. Heistermann, S. Huang, V. Kolokotronis, Z. W. Kundzewicz, L. Menzel, et al., 2017. Hochwasser und Sturzfluten an Flüssen in Deutschland. Pages 87–101 *in* Klimawandel in Deutschland. Springer.

Castellarin, A., D. Burn, and A. Brath. 2001. Assessing the effectiveness of hydrological similarity measures for flood frequency analysis. Journal of Hydrology **241**:270–285.

Chifflard, P. 2006. Der Einfluss des Reliefs, der Hangsedimente und der Bodenvorfeuchte auf die Abflussbildung im Mittelgebirge: experimentelle Prozess-Studien im Sauerland. Geograph. Inst. der Ruhr-Univ.

DiCiccio, T. J., and B. Efron. 1996. Bootstrap confidence intervals. Statistical science pages 189–212.

Dormann, C. F. 2017. Parametrische Statistik: Verteilungen, maximum likelihood und GLM in R. Springer-Verlag.

Dunteman, G. H. 1989. Principal components analysis. 69, Sage.

DWD. 2015. Deutscher Wetterdienst: Climate station data .

Dyck, S., and G. Peschke. 1995. Grundlagen der Hydrologie. Verlag für Bauwesen, Berlin .

EEA. 2010. European Environmental Agency : CORINE Land Cover 2006. http://www.eea.europa.eu (last access: 1 July 2010) .

Efron, B., and R. J. Tibshirani. 1994. An introduction to the bootstrap. CRC press.

EWA. 2010. European Water Archive: Streamflow Records 1950-2010 .

Fohrer, N., H. Bormann, K. Miegel, and M. Casper. 2016. Hydrologie. UTB.

Freudiger, D., I. Kohn, K. Stahl, and M. Weiler. 2014. Large-scale analysis of changing frequencies of rain-on-snow events with flood-generation potential. Hydrology and Earth System Sciences **18**:2695.

Friedman, J. H., B. E. Popescu, et al. 2008. Predictive learning via rule ensembles. The Annals of Applied Statistics **2**:916–954.

Garvelmann, J., S. Pohl, and M. Weiler. 2013. From observation to the quantification of snow processes with a time-lapse camera network. Hydrology and Earth System Sciences **17**:1415–1429.

Goldstein, A., A. Kapelner, J. Bleich, and E. Pitkin. 2015. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. Journal of Computational and Graphical Statistics **24**:44–65.

GRDC. 2010. Global Runoff Data Centre: Streamflow records 1950-2010 .

Greenwell, B. M. 2017. pdp: An R Package for Constructing Partial Dependence Plots. The R Journal **9**:421–436.

Gudmundsson, L., and S. I. Seneviratne. 2015. Towards observation-based gridded runoff estimates for Europe. Hydrology and Earth System Sciences **19**:2859–2879.

Gutknecht, D., C. Reszler, and G. Blöschl. 2002. Das Katastrophenhochwasser vom 7. August 2002 am KampEine erste Einschätzung. e & i Elektrotechnik und Informationstechnik **119**:411–413.

Häckel, H. 2016. Meteorologie. UTB.

Hargreaves, G. H., and Z. A. Samani. 1982. Estimating potential evapotranspiration. Journal of the Irrigation and Drainage Division **108**:225–230.

Harrell, F. E. J., 2018. Hmisc: Harrell Miscellaneous - R package version 4.1.1.

Hastie, T., R. Tibshirani, and J. Friedman. 2009. The elements of statistical learning: data mining, inference and prediction. 2 edition. Springer.

He, Z., J. Parajka, F. Tian, and G. Blschl. 2014. Estimating degree-day factors from MODIS for snowmelt runoff modeling. Hydrology and Earth System Sciences **18**:4773–4789.

Herrera, M., L. Torgo, J. Izquierdo, and R. Prez-Garca. 2010. Predictive models for forecasting hourly urban water demand. Journal of Hydrology **387**:141 – 150.

Huza, J., A. J. Teuling, I. Braud, J. Grazioli, L. A. Melsen, G. Nord, T. H. Raupach, and R. Uijlenhoet. 2014. Precipitation, soil moisture and runoff variability in a small river catchment (Ardèche, France) during HyMeX Special Observation Period 1. Journal of hydrology **516**:330–342.

IHUK. 1999. Handbook of Flood Estimation. Institute of Hydrology Wallingford, UK .

Kuhn, M. 2008. Building predictive models in R using the caret package. Journal of Statistical Software **28**:1–26.

Kumar, R. 2010. Distributed Hydrologic Model Parameterization: Application in a Mesoscale River Basin. Dissertation, Helmholtz Centre for Environmental Research - UFZ.

Kumar, R., L. Samaniego, and S. Attinger. 2013. Implications of distributed hydrologic model parameterization on water fluxes at multiple scales and locations. Water Resources Research **49**:360–379.

Kundzewicz, Z. W., D. Graczyk, T. Maurer, I. Pi
'nskwar, M. Radziejewski, C. Svensson, and M. Szwed. 2005. Trend detection in river flow series: 1. Annual maximum flow/Détection de tendance dans des séries de débit fluvial: 1. Débit maximum annuel. Hydrological Sciences Journal **50**.

Kuraś, P. K., Y. Alila, and M. Weiler. 2012. Forest harvesting effects on the magnitude and frequency of peak flows can increase with return period. Water Resources Research **48**.

Lang, M., T. Ouarda, and B. Bobée. 1999. Towards operational guidelines for over-threshold modeling. Journal of hydrology **225**:103–117.

Lee, S., J.-C. Kim, H.-S. Jung, M. J. Lee, and S. Lee. 2017. Spatial prediction of flood susceptibility using random-forest and boosted-tree models in Seoul metropolitan city, Korea. Geomatics, Natural Hazards and Risk **8**:1185–1203.

Liaw, A., and M. Wiener. 2002. Classification and regression by randomForest. R news **2**:18–22.

Lima, A. R., A. J. Cannon, and W. W. Hsieh. 2015. Nonlinear regression in environmental sciences using extreme learning machines: a comparative evaluation. Environmental Modelling & Software **73**:175–188.

Linsley, R. K. 1943. A simple procedure for the day-to-day forecasting of runoff from snow-melt. Eos, Transactions American Geophysical Union **24**:62–67.

Maniak, U. 2013. Hydrologie und Wasserwirtschaft: Eine Einführung für Ingenieure. Springer-Verlag.

Merz, B., and E. J. Plate. 1997. An analysis of the effects of spatial variability of soil and soil moisture on runoff. Water Resources Research **33**:2909–2922.

Merz, R., and G. Blöschl. 2003. A process typology of regional floods. Water Resources Research **39**.

Merz, R., and G. Blöschl. 2009*a*. Process controls on the statistical flood moments-a data based analysis. Hydrological processes **23**:675–696.

Merz, R., and G. Blöschl. 2009*b*. A regional analysis of event runoff coefficients with respect to climate and catchment characteristics in Austria. Water Resources Research **45**.

Meynen, E., J. Schmithüsen, B. für Landeskunde und Raumforschung (Germany), and Z. für Deutsche Landeskunde. 1953. Handbuch der naturräumlichen Gliederung Deutschlands: unter Mitwirkung des Zentralausschusses für deutsche Landeskunde. Number Bd. 1-2 in Handbuch der naturräumlichen Gliederung Deutschlands: unter Mitwirkung des Zentralausschusses für deutsche Landeskunde, Bundesanstalt für Landeskunde und Raumforschung.

Middleton, N. J., and D. S. Thomas. 1992. World atlas of desertification .

Morbidelli, R., C. Saltalippi, A. Flammini, and R. S. Govindaraju. 2018. Role of slope on infiltration: A review. Journal of Hydrology **557**:878 – 886.

Naghettini, M. 2016. Fundamentals of statistical hydrology. Springer.

Nied, M., Y. Hundecha, and B. Merz. 2013. Flood-initiating catchment conditions: a spatio-temporal analysis of large-scale soil moisture patterns in the Elbe River basin. Hydrology and Earth System Sciences **17**:1401–1414.

Nied, M., T. Pardowitz, K. Nissen, U. Ulbrich, Y. Hundecha, and B. Merz. 2014. On the relationship between hydro-meteorological patterns and flood types. Journal of hydrology **519**:3249–3262.

Nied, M., K. Schröter, S. Lüdtke, V. D. Nguyen, and B. Merz. 2017. What are the hydro-meteorological controls on flood characteristics? Journal of hydrology **545**:310–326.

Pallard, B., A. Castellarin, and A. Montanari. 2009. A look at the links between drainage density and flood statistics. Hydrology and Earth System Sciences **13**:1019–1029.

Paquet, E., F. Garavaglia, R. Garçon, and J. Gailhard. 2013. The SCHADEX method: A semi-continuous rainfall–runoff simulation for extreme flood estimation. Journal of Hydrology **495**:23–37.

Parr, T., K. Turgutlu, C. Csiszar, and J. Howard, 2018. Beware Default Random Forest Importances. Online ressource: http://parrt.cs.usfca.edu/doc/rf-importance/index.html (retrieved: 31.05.2018), University of San Francisco.

Patt, H., and R. Jüpner. 2001. Hochwasser-Handbuch. Auswirkungen und Schutz, Berlin, Heidelberg .

Peñas, F. J., J. Barquín, T. H. Snelder, D. J. Booker, and C. Álvarez. 2014. The influence of methodological procedures on hydrological classification performance. Hydrology and Earth System Sciences **18**:3393–3409.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. 2011. Scikit-learn Documentation: Machine learning in Python **12**:2825–2830.

Penna, D., H. Tromp-van Meerveld, A. Gobbi, M. Borga, and G. Dalla Fontana. 2011. The influence of soil moisture on threshold runoff generation processes in an alpine headwater catchment. Hydrology and Earth System Sciences **15**:689.

Petrow, T., and B. Merz. 2009. Trends in flood magnitude, frequency and seasonality in Germany in the period 19512002. Journal of Hydrology **371**:129 – 141.

Petrow, T., B. Merz, K.-E. Lindenschmidt, and A. H. Thieken. 2007. Aspects of seasonality and flood generating circulation patterns in a mountainous catchment in south-eastern Germany. Hydrology and Earth System Sciences **11**:1455–1468.

Petrow, T., J. Zimmer, and B. Merz. 2009. Changes in the flood hazard in Germany through changing frequency and persistence of circulation patterns. Natural Hazards and Earth System Sciences **9**:1409–1423.

Pfaundler, M. 2001. Adapting, analysing and evaluating a flexible Index Flood regionalisation approach for Switzerland. Federal Institute of Technology, Zurich .

R Core Team, 2018. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Rössler, O. K., P. A. Froidevaux, U. Börst, R. Rickli, O. Romppainen-Martius, and R. Weingartner. 2014. Retrospective analysis of a nonforecasted rain-on-snow flood in the Alps–a matter of model limitations or unpredictable nature? Hydrology and earth system sciences **18**:2265–2285.

Rossum, G., 1995. Python Reference Manual. Technical report, Amsterdam, The Netherlands.

Sadler, J., J. Goodall, M. Morsy, and K. Spencer. 2018. Modeling urban coastal flood severity from crowd-sourced flood reports using Poisson regression and Random Forest. Journal of Hydrology **559**:43 – 55.

Samaniego, L., and A. Bárdossy. 2007. Relating macroclimatic circulation patterns with characteristics of floods and droughts at the mesoscale. Journal of Hydrology **335**:109–123.

Samaniego, L., R. Kumar, and S. Attinger. 2010. Multiscale parameter regionalization of a grid-based hydrologic model at the mesoscale. Water Resources Research **46**.

Samaniego, L., R. Kumar, and M. Zink. 2013. Implications of parameter uncertainty on soil moisture drought analysis in Germany. Journal of Hydrometeorology **14**:47–68.

Samaniego-Eguiguren, L. E., 2003. Hydrological consequences of land use/land cover and climatic changes in mesoscale catchments. Ph.D. thesis, Institut fr Wasser- und Umweltsystemmodellierung, Universitt Stuttgart.

Schröter, K., M. Kunz, F. Elmer, B. Mühr, and B. Merz. 2015. What made the June 2013 flood in Germany an exceptional event? A hydro-meteorological evaluation. Hydrology and Earth System Sciences **19**:309–327.

Shortridge, J. E., S. D. Guikema, and B. F. Zaitchik. 2016. Machine learning methods for empirical streamflow simulation: a comparison of model accuracy, interpretability, and uncertainty in seasonal watersheds. Hydrology and Earth System Sciences **20**:2611–2628.

Solomatine, D. P., and A. Ostfeld. 2008. Data-driven modelling: some past experiences and new approaches. Journal of Hydroinformatics **10**:3–22.

Strahler, A. N. 1957. Quantitative analysis of watershed geomorphology. Eos, Transactions American Geophysical Union **38**:913–920.

Strobl, C., A.-L. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis. 2008. Conditional variable importance for random forests. BMC bioinformatics **9**:307.

Strobl, C., A.-L. Boulesteix, A. Zeileis, and T. Hothorn. 2007. Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC bioinformatics **8**:25.

Sui, J., and G. Koehler. 2001. Rain-on-snow induced flood events in Southern Germany. Journal of Hydrology **252**:205–220.

Tarboton, D. G., R. L. Bras, and I. Rodriguez-Iturbe. 1991. On the extraction of channel networks from digital elevation data. Hydrological processes **5**:81–100.

Terti, G., I. Ruin, J. J. Gourley, P. Kirstetter, Z. Flamig, J. Blanchet, A. Arthur, and S. Anquetin. 2017. Toward Probabilistic Prediction of Flash Flood Human Impacts. Risk Analysis .

Uhlemann, S., A. Thieken, and B. Merz. 2010. A consistent set of trans-basin floods in Germany between 1952–2002. Hydrology and Earth System Sciences **14**:1277–1295.

Uhlenbrook, S., A. Steinbrich, D. Tetzlaff, and C. Leibundgut. 2002. Regional analysis of the generation of extreme floods. International Association of Hydrological Sciences, Publication pages 243–249.

Wang, Z., C. Lai, X. Chen, B. Yang, S. Zhao, and X. Bai. 2015. Flood hazard risk assessment model based on random forest. Journal of Hydrology **527**:1130–1141.

Wharton, G. 1994. Progress in the use of drainage network indices for rainfall-runoff modelling and runoff prediction. Progress in Physical Geography: Earth and Environment **18**:539–557.

Wood, S. N. 2006. Generalized additive models: an introduction with R. Chapman and Hall/CRC.

Zhao, G., B. Pang, Z. Xu, J. Yue, and T. Tu. 2018. Mapping flood susceptibility in mountainous areas on a national scale in China. Science of The Total Environment **615**:1133 – 1142.

Zink, M., R. Kumar, M. Cuntz, and L. Samaniego. 2017. A high-resolution dataset of water fluxes and states for Germany accounting for parametric uncertainty. Hydrology and Earth System Sciences **21**:1769–1790.

Zrinji, Z., and D. H. Burn. 1994. Flood frequency analysis for ungauged sites using a region of influence approach. Journal of hydrology **153**:1–21.

List of Symbols/Acronyms

Symbol	Unit	Description
Δt	d	Time interval of preconditions
1S - 4S	-	Regional subsets/models of summer
1W - 4W	-	Regional subsets/models of winter
AI	fraction	Aridity Index
AIC	-	Aikake's Information Criterion
$Altitude$	m a.s.l.	Mean catchment altitude
$Area$	km^2	Catchment area
BIC	-	Bayesian Information Criterion
$ChSlope$	°	Mean channel slope
CV	-	Coefficient of Variation
DD	km^{-1}	Drainage density
DEM	-	Digital elevation model
$FLCV$	-	Coefficient of Variation of flow path length
$FLMax$	km	Maximum flow path length
$Forest$	%	Share of "Forest" in catchment land cover
GLM	-	Generalized Linear Model
$Impervious$	%	Share of "Impervious" in catchment land cover
l	km	Sum of channel lengths in a catchment
MAF	-	Maximum Annual Flood
mHm	-	Mesoscale Hydrologic Model
ML	-	Machine-Learning
MSE	mm^2	Mean Squared Error
n_{PC}	-	Number of principal components
P	mm	Precipitation
P_ann	mm	Mean annual precipitation
P_{eff}	mm	Effective precipitation
$P_{eff0...7}$	mm	Mean effective precipitation of Δt=0 - 7 d
PCA	-	Principal Component Analysis
PDP	-	Partial Dependence Plot
$Permeable$	%	Share of "Permeable" in catchment land cover
PET_{ann}	mm	Mean annual potential evapotranspiration
POT	-	Peak-Over-Threshold method
Q	mm	Daily mean streamflow
Q_f	mm	Flood magnitude
r^2	-	Pearman's Squared Correlation Coefficient
R^2	fraction	Coefficient of Determination
RF		RandomForest algorithm
rfe	-	Random Feature Selection algorithm
$RMSE$	mm	Root Mean Squared Error
RoS	-	Rain-on-Snow event
s	%	Explained variance by PCA-analysis
SM	fraction	Soil moisture
$SM_{0...7}$	fraction	Mean soil moisture of Δt=0 - 7 d
$Slope$	°	Mean catchment slope
stepBIC	-	StepBIC feature selection algorithm
SVM	-	Support Vector Machine algorithm
T	°C	Air temperature
$T_{eff0...7}$	°C	Mean air temperature of Δt=0 - 7 d
X	%	Percentile threshold

Analysis of Collinearity

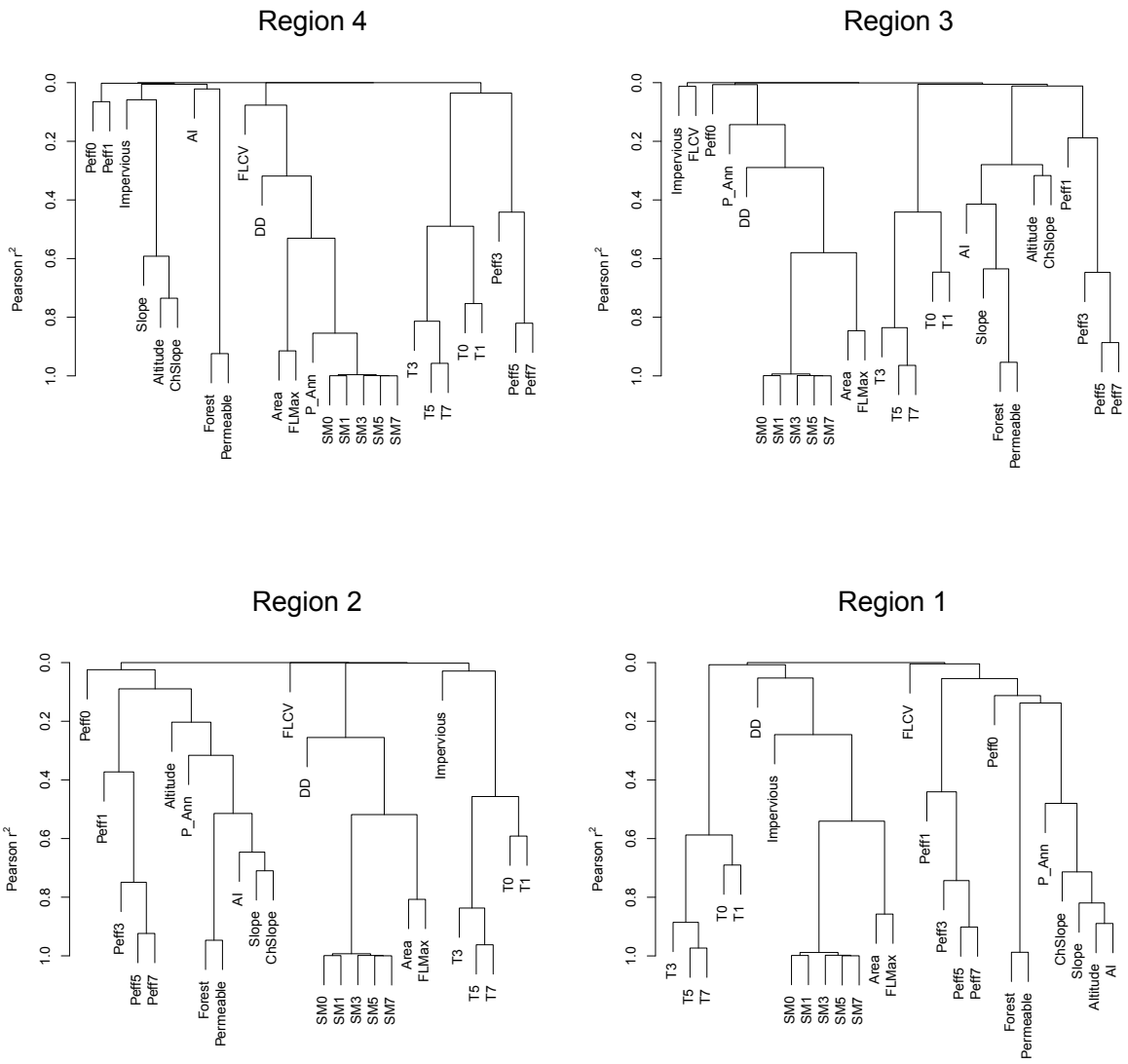


Figure A.1: Results of Cluster Analysis of each regional dataset

Model Validation

Accuracy Statistics

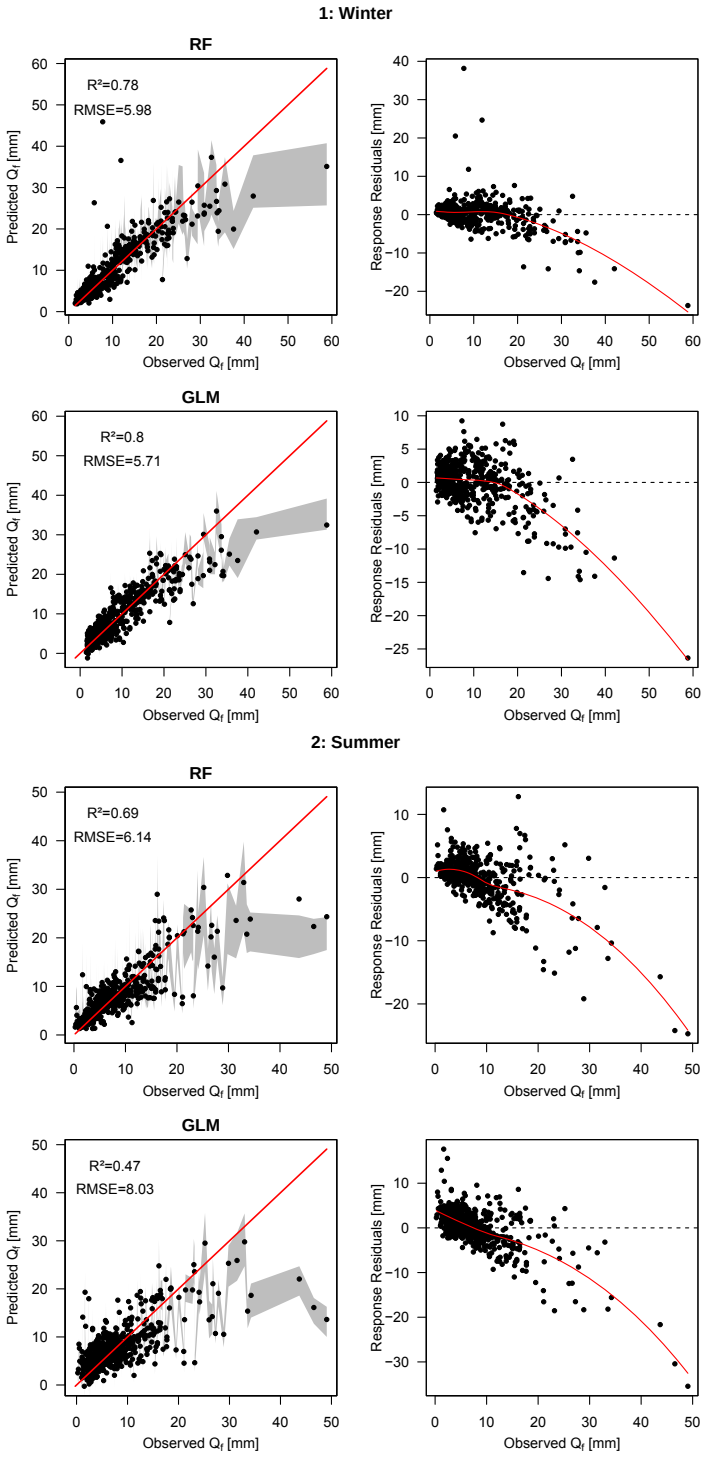
Table A.1: Model accuracy of RF and GLM by R^2 on training and test data across regions and seasons.

Region/Season	**1s**	**1w**	**2s**	**2w**	**3s**	**3w**	**4s**	**4w**	**Mean**	
RF	0.85	0.75	0.72	0.78	0.58	0.74	0.54	0.7	0.71	Training
GLM	0.79	0.72	0.56	0.59	0.47	0.57	0.33	0.44	0.56	
RF	0.86	0.78	0.69	0.78	0.49	0.73	0.58	0.75	0.71	Test
GLM	0.74	0.8	0.47	0.59	0.38	0.56	0.40	0.44	0.55	

Table A.2: Model accuracy of RF by R^2 on training data, including either P or P_{eff}.

	Model	1S	1W	2S	2W	3S	3W	4S	4W	Mean
Model	RF_P_{eff}	0.85	0.75	0.72	0.78	0.58	0.74	0.54	0.70	0.71
Model	RF_P	0.84	0.73	0.72	0.76	0.59	0.69	0.54	0.67	0.69

Analysis Plots



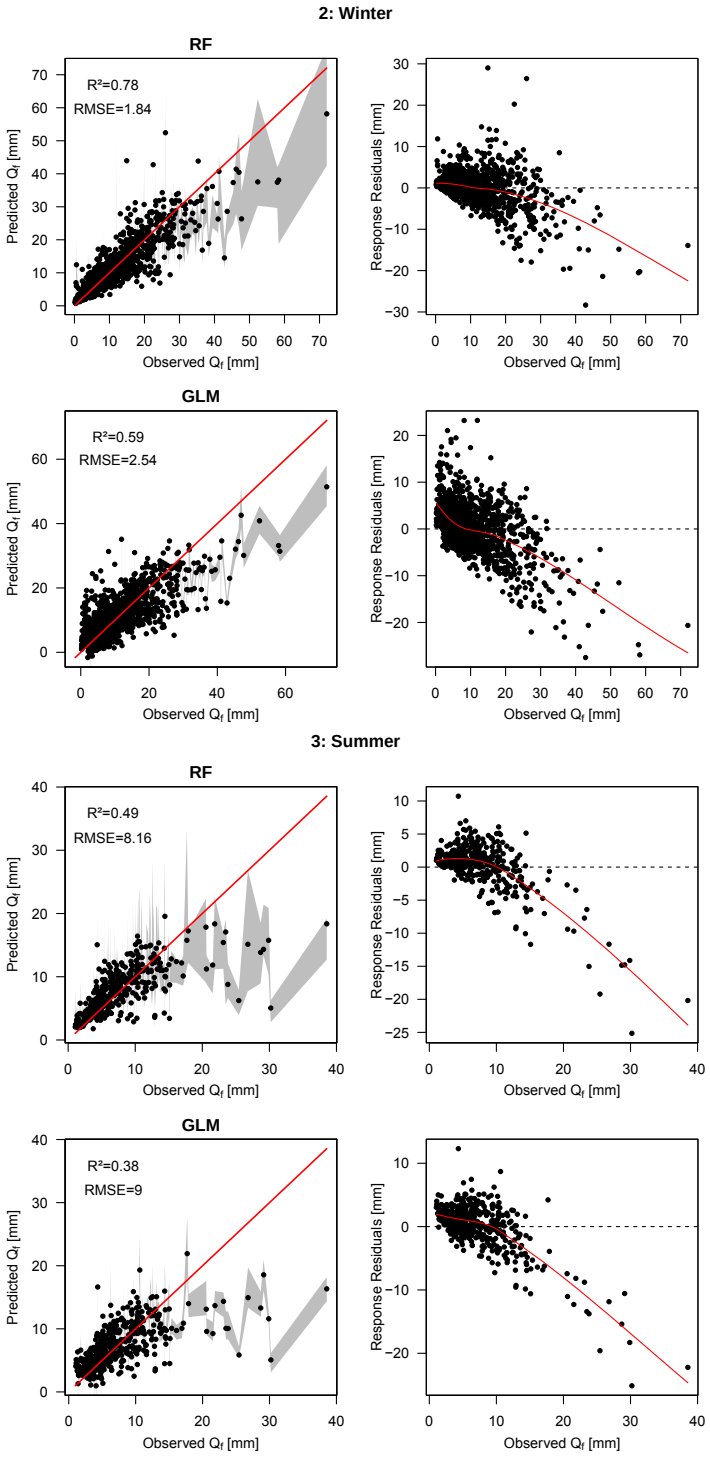


Figure A.2: Analysis plots of RF and GLM on test data across all regions and seasons. Left: Predicted vs observed Q_f. Red line depicts ideal 1:1 fit. Grey areas depict the lower and upper prediction bounds of a 100-fold bootstrap procedure. Right: Residuals on response scale. Red line depicts a LOESS-regression on residuals to visualize goodness-of-fit in lower values of Q_f.

Model Interpretation

Variable Importances RF by Variable

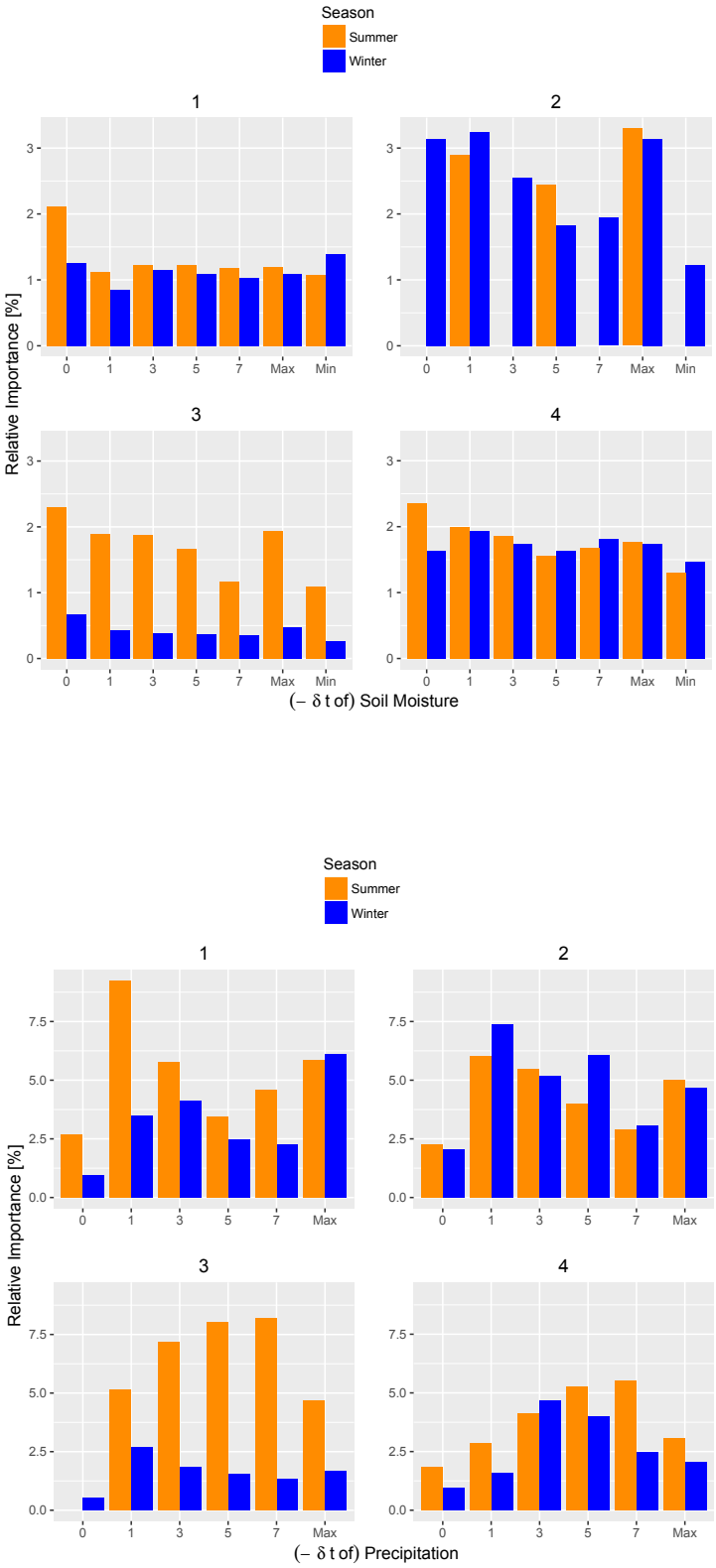
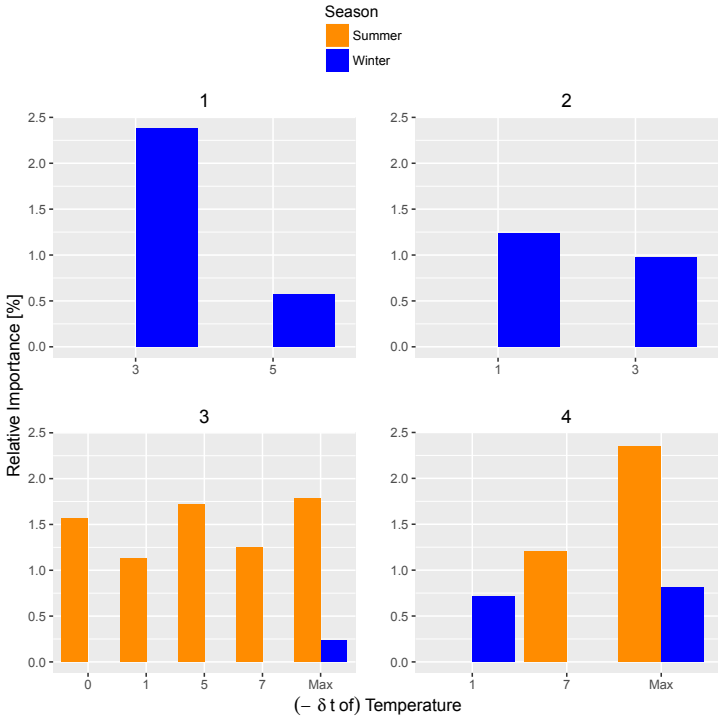
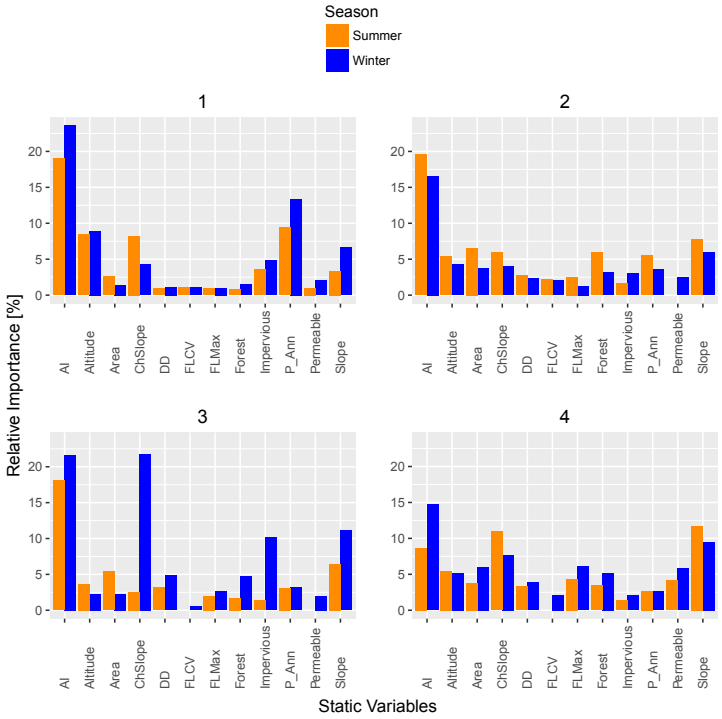


Figure A.3: Variable importance of RF of all predictors by region, season and time interval


Season
Summer
Winter
1
2
3
4
Relative Importance [%]
(– δ t of) Temperature


Season
Summer
Winter
1
2
3
4
Relative Importance [%]
AI
Altitude
Area
ChSlope
DD
FLCV
FLMax
Forest
Impervious
P_Ann
Permeable
Slope
Static Variables

Variable Importances GLM by Variable

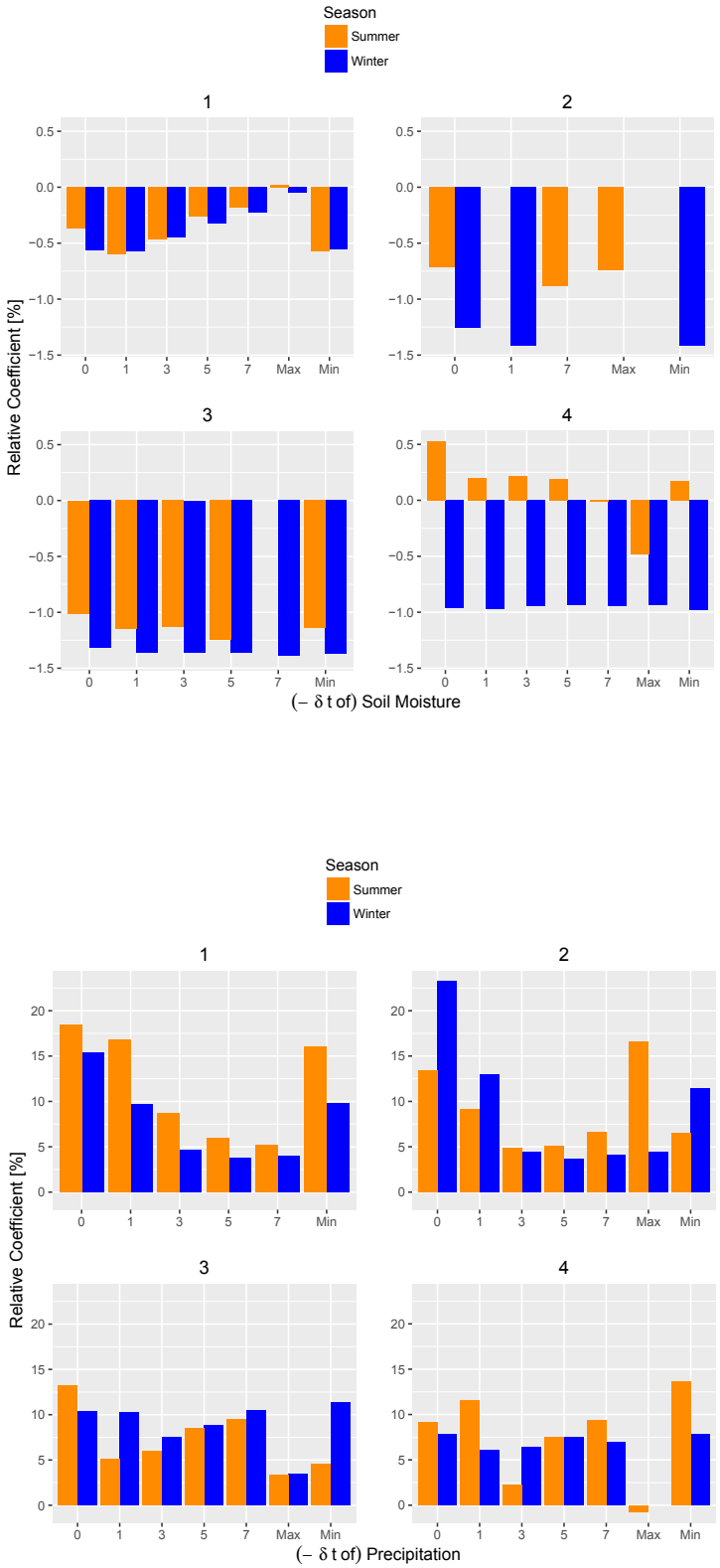
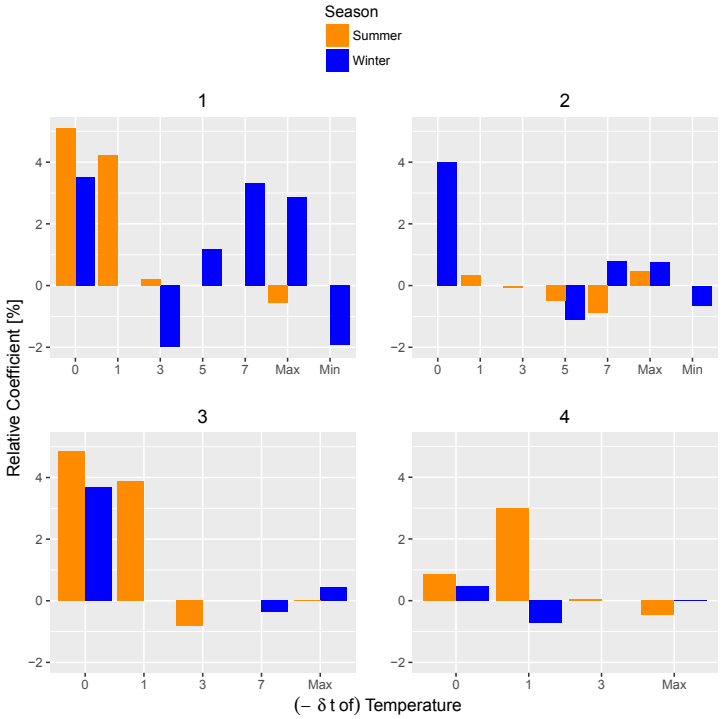
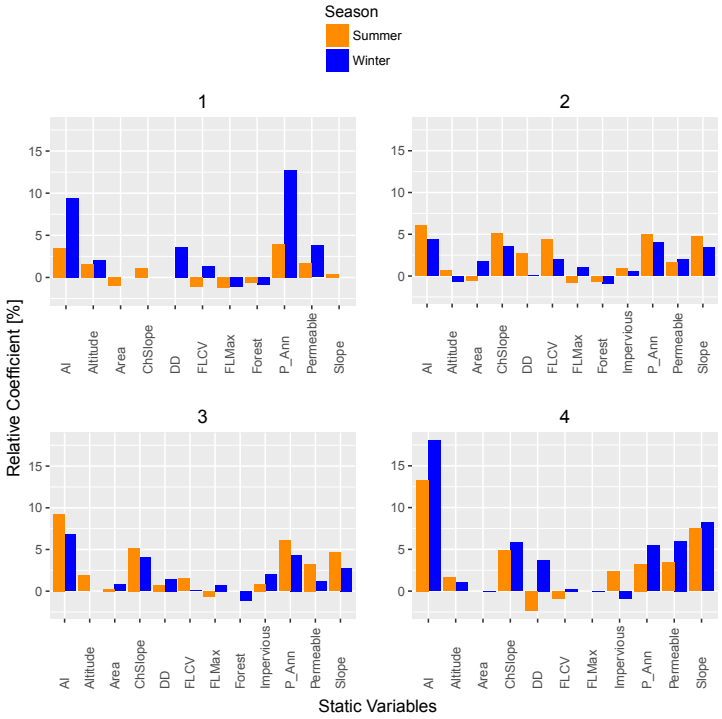


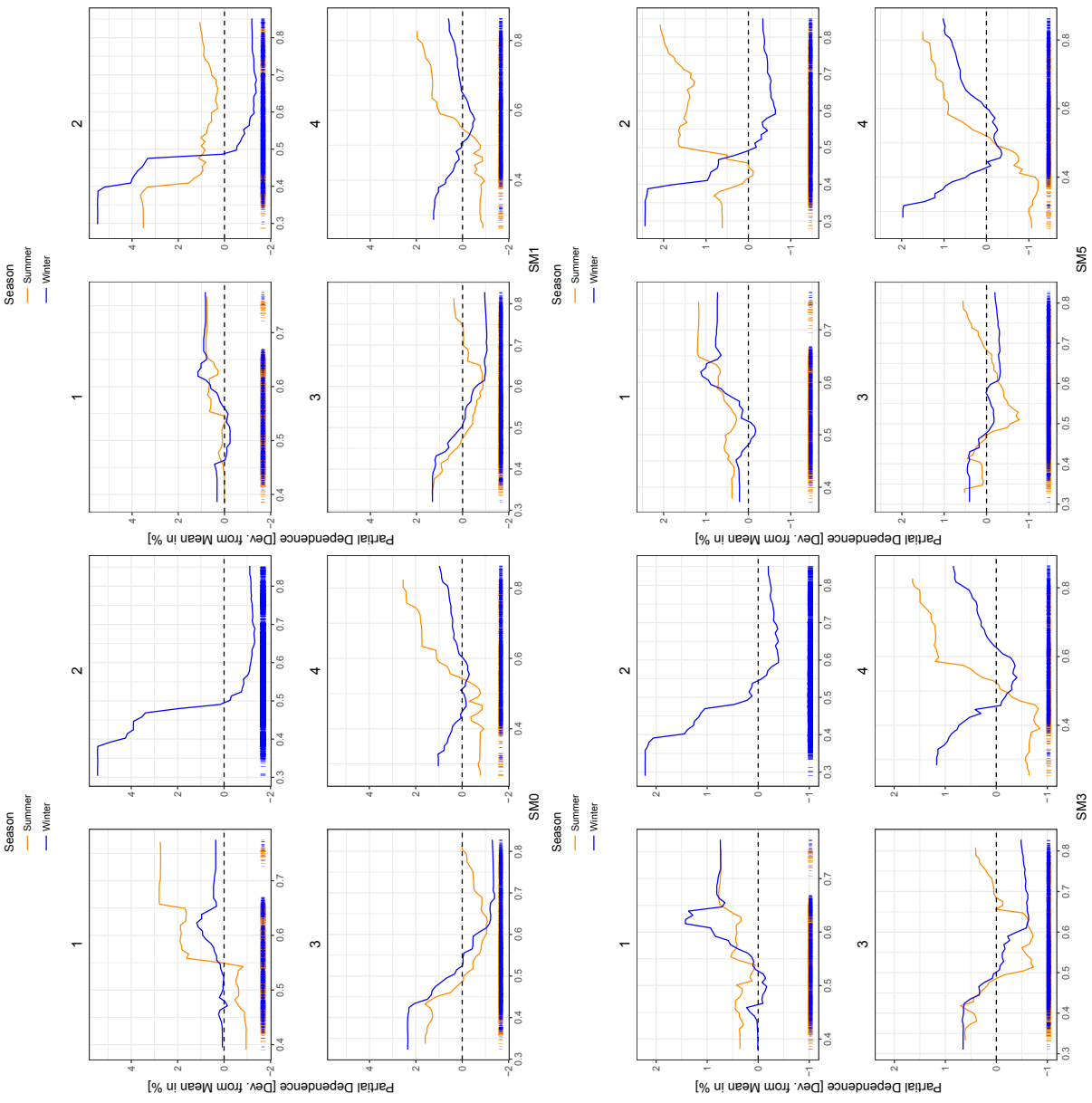
Figure A.4: Variable importance of GLM of all predictors by region, season and time interval

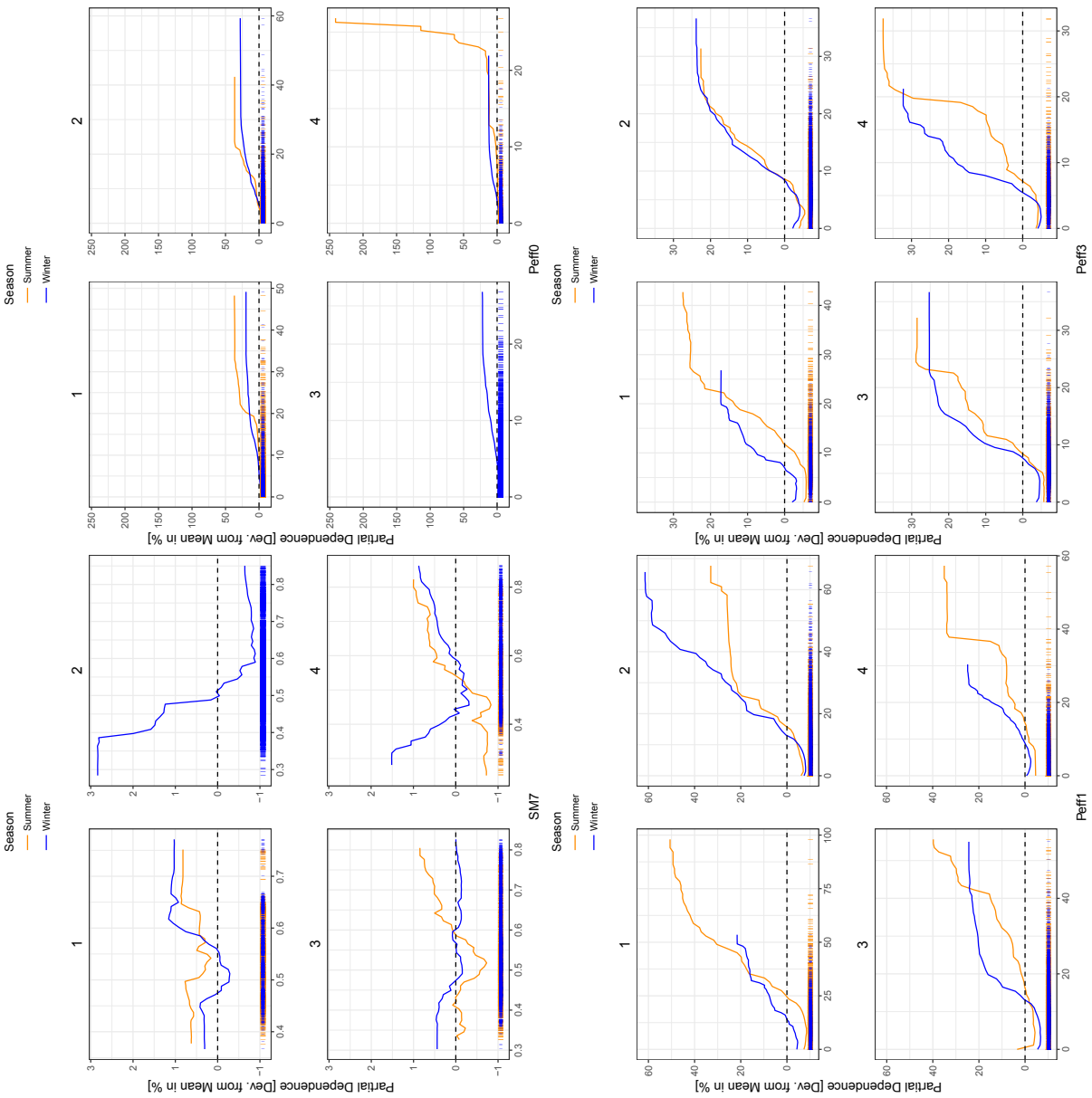

Season
Summer
Winter
1
2
3
4
Relative Coefficient [%]
(– δ t of) Temperature
0
1
3
5
7
Max
Min

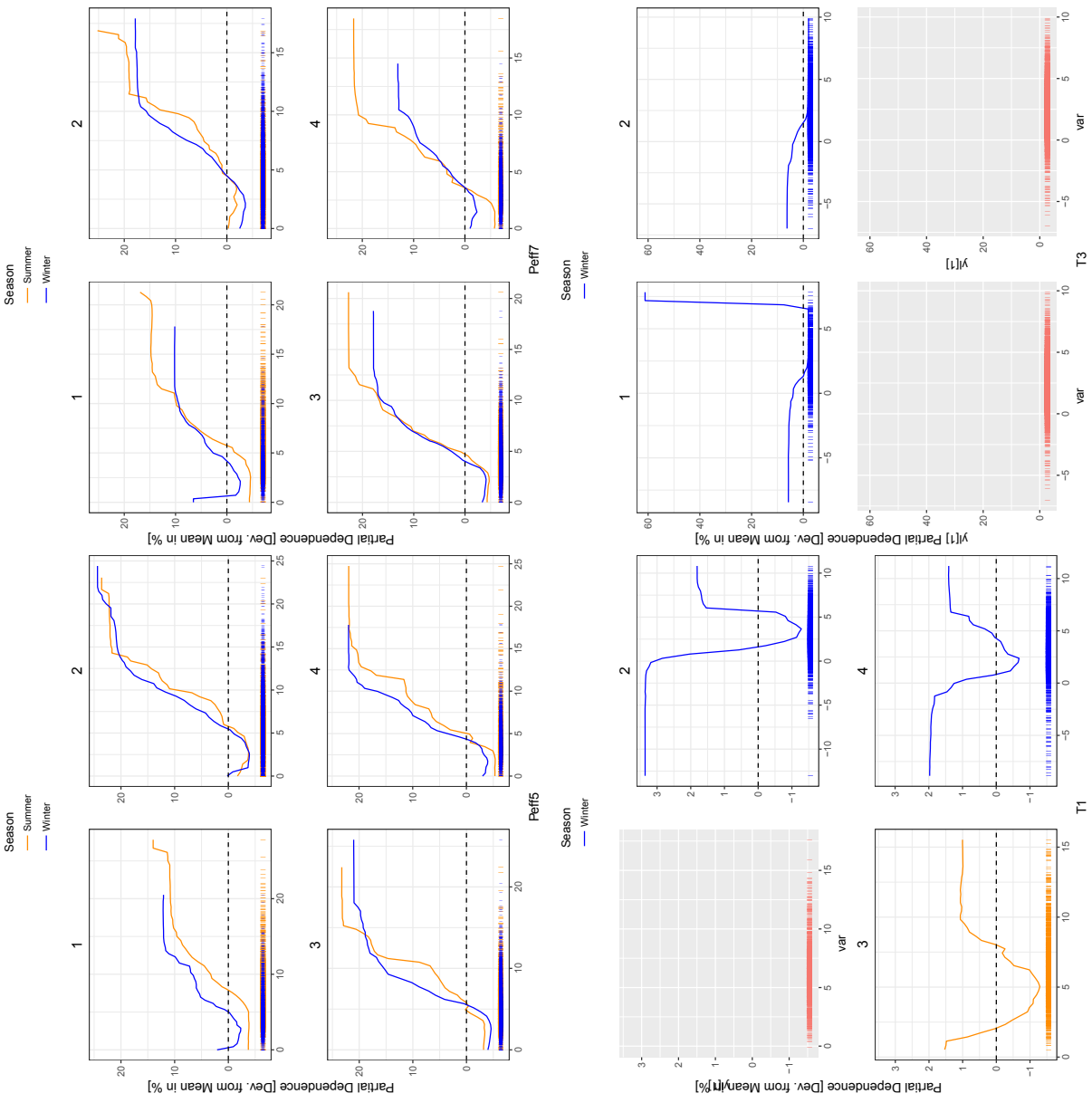

Season
Summer
Winter
1
2
3
4
Relative Coefficient [%]
Static Variables
AI
Altitude
Area
ChSlope
DD
FLCV
FLMax
Forest
Impervious
P_Ann
Permeable
Slope

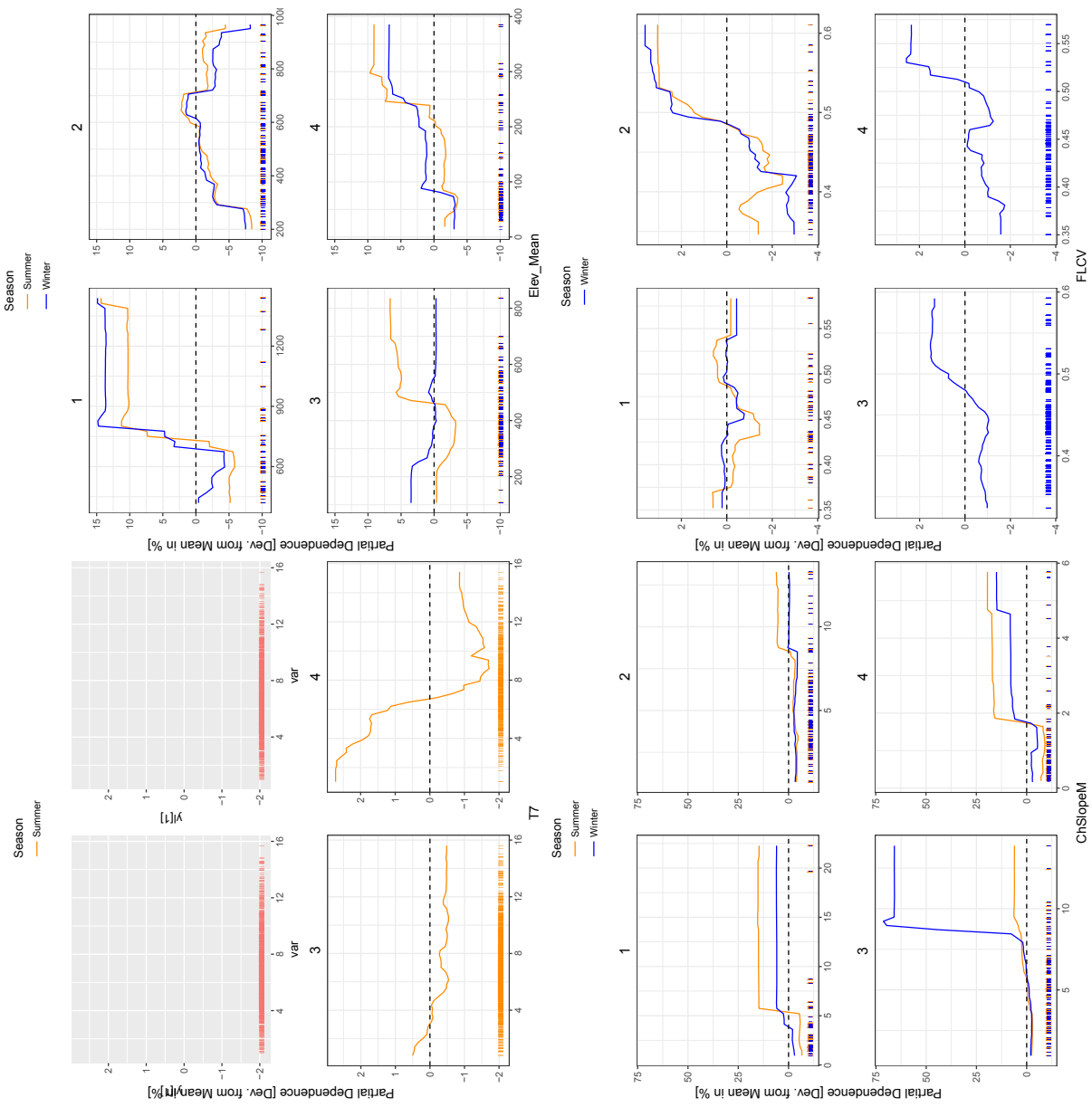
Partial Dependence Plots of RF

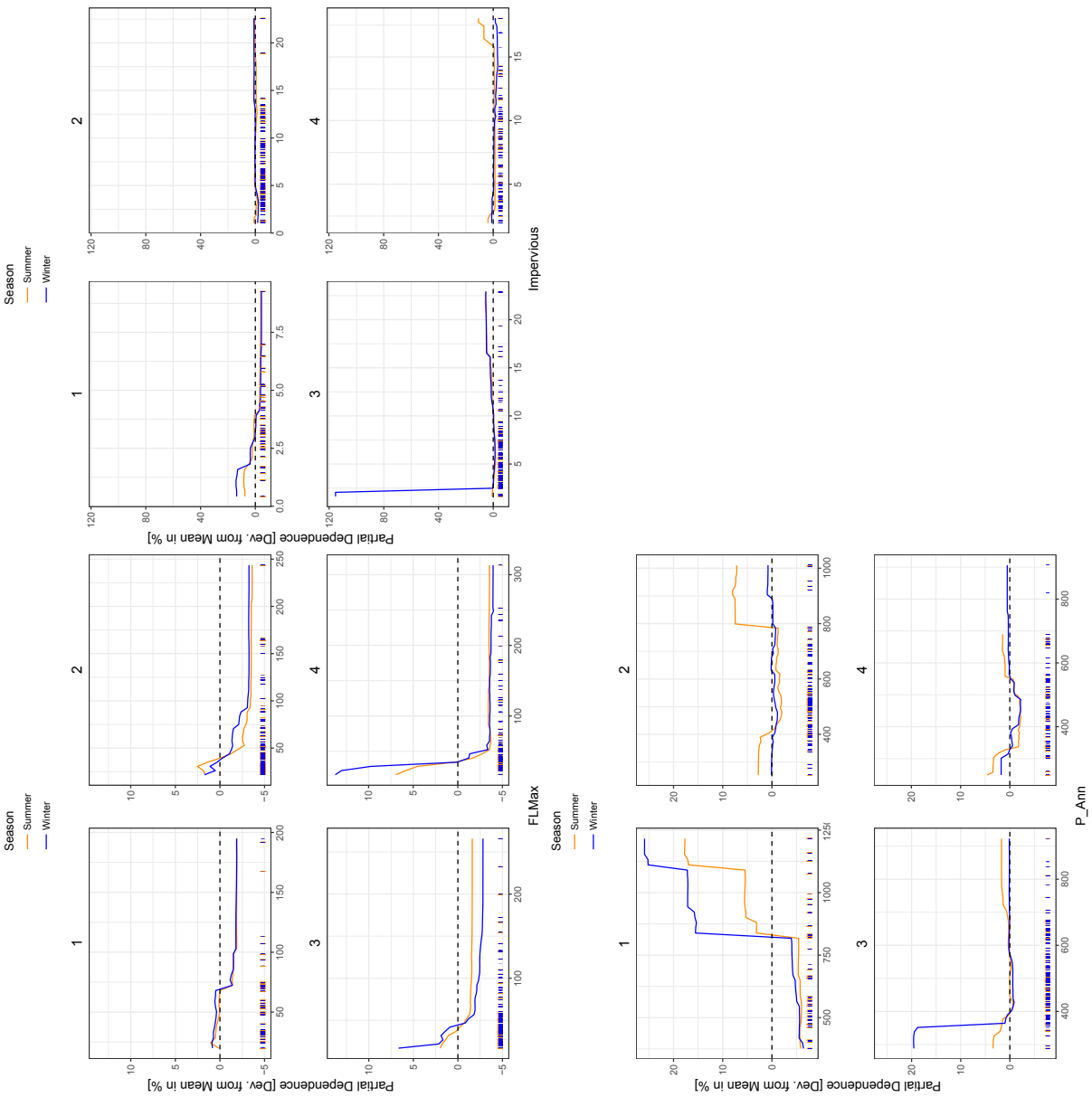
Figure A.5: Relative partial dependence plots of $\hat{Q}_f$ are displayed for all dynamic predictors. For static predictors, the ones already shown in the main text are excluded.


Season
Summer
Winter
SM0
SM1
SM3
SM5
Partial Dependence [Dev. from Mean in %]
1
2
3
4


Season
Summer
Winter
Partial Dependence [Dev. from Mean in %]
Peff0
Peff1
Peff3
SM7


Season
Summer
Winter
Partial Dependence [Dev. from Mean in %]
Peff7
Peff5
T1
T3
var
yl[1]
Partial Dependence [Dev. from Mean yl[1] %]


Season
Summer
Winter
Partial Dependence [Dev. from Mean in %]
Elev_Mean
T7
ChSlopeM
FLCV
var
y[t]
1
2
3
4


Season
Summer
Winter
1
2
3
4
Partial Dependence [Dev. from Mean in %]
FLMax
Impervious
P_Ann

Eidesstattliche Erklärung

Hiermit versichere ich, Lennart Schmidt, die vorliegende Arbeit selbstständig und unter ausschließlicher Verwendung der angegebenen Literatur und Hilfsmittel erstellt zu haben. Alle Stellen, die wörtlich oder sinngemäß veröffentlichtem oder unveröffentlichtem Schrifttum entnommen sind, habe ich als solche kenntlich gemacht. Die Arbeit wurde bisher in gleicher oder ähnlicher Form keiner anderen Prüfungsbehörde vorgelegt und auch nicht veröffentlicht. Die elektronische Version dieser Masterarbeit stimmt in Inhalt und Formatierung mit den auf Papier ausgedruckten Exemplaren überein.

Freiburg, den 19.06.2018