GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN IN PUBLICA COMMODA SEIT 1737

A COMPARISON OF SPACE-TIME MODELING APPROACHES FOR EXTREME PRECIPITATION IN GERMANY BETWEEN 1996 AND 2016

A thesis paper for the degree of Bachelor of Arts (B.A.)

Wintersemester 2022

By

Ekin Gülhan

Supervisor: Dr. Isa Marques
Enrolment number: 21675680
e-mail: ekin.guelhan(at)stud.uni-goettingen.de
Address: Hannoversche Straße 134, 37077 Göttingen
Date: 2023-03-20

Contents

1 Preface

Dear reader,

the author of this thesis paid special attention to the creation of a reproducible and dynamic reasearch paper. This .pdf version of the thesis has been created from a rmarkdown file[1] named `BachelorThesis.rmd` and this .rmd file is available in a GitHub repository[2]. It is available alongside all the R code, which is encapsulated in different thematically named R scripts. The repository on GitHub has been ordered in a way that allows easy and intuitive access and navigation: Input and Output, source code and the .rmd file are divided in seperate directories, which are then subdivided based on format of the data and the topic of the content. Whenever code has been written, great attention has been paid to readibility. To allow for such readibility, tidyverse syntax and functions have been used consistently. This implementation of the tidyverse does also reflect the authors understanding of an effective way to account for the data-flow centered nature of the programming language R for the merit of data science.

As a project, the thesis has been made to evolve around a file of type `R Project` to further enable reproducibility and mobility. The `BachelorThesis.R Project` file can be considered as the ultimate reference point within the toplevel of the project directory. When browsing through the directory or trying to reproduce a specific analytical step or even the whole thesis analysis, the reader should regard `BachelorThesis.R Project` as the entry point to be opened as an unconditional first step. Such restriction is neccessary, as it allows for the mobility of the thesis in the first place, when used in combination with the R packages `here` and `renv`. In this context, mobility is understood as a characteristic of the most top-level directory of the thesis, to be the sole reference point for the .rmd file and all R scripts. When knitting `BachelorThesis.rmd` to `BachelorThesis.pdf` or when executing any R script, there is no need to inconvenience oneself with paying attention to setting working directories and defining correct absolute paths.

While the reader can forego inconvenient absolute paths with the help of `BachelorThesis.R Project` and `here`, the combination of `BachelorThesis.R Project` and the R package `renv` allows for a secondary project-specific library. A `renv.lock` file snapshots the packages in the

---

[1]Note that this .pdf file was successfully knitted from the .rmd file on a windows PC with 8 gigabytes of RAM without any other programs running except for system related services.

[2]Repository link: https://github.com/ekingn/BachelorThesis.
Access to the repository on request. Reach out to me via ekin.guelhan@posteo.de

respective versions, in which they were used within the thesis project, in a descriptive list. By utlizing the `renv.lock` file through the package renv, any reader can easily install all the packages in the exact versions required for the reproduction of analytical steps and creating the present `BachelorThesis.pdf` file. The packages are assigned to a secondary project library, such that the primary user library remains unchainged.

In summary, if the readers intent goes beyond reading the `BachelorThesis.pdf`, he or she should do the following:

1. Open the file `BachelorThesis.R Project`

2. If not already done, install and require the R package `here` and `renv`

3. Execute `renv::restore()`.

4. Possibly repeat the last step multiple times.

The author also wants to point out, that that the contents of the project directory go beyond what is explicitly included in the `BachelorThesis.pdf` file. A plethora of data exploration results including various data visualizations can be found in the GitHub repository as well. I therefore invite the curious reader to browse its content in the hope, that he or she finds something of interest.

When reading equations and formulae, the reader should look out for symbols and strings in bold-font, which denote vectors.

I finally want to thank Dr. Isa Marques for her through-out available council and insightful as well as helpful feedback.

<div align="right">

Ekin Gülhan

2023-03-20

</div>

2   Introduction

While extreme precipitation rarely occurs, it can cause flooding comparable in scale to the floodings of 2002 and 2013 that impacted Europe and Germany specifically, resulting in loss of life and property (cf. Kuhlicke et al. (2016), p. 1)(cf. Kreibich et al. (2017), p. 2075).

Humans may not prevent the occurence of such detrimental natural incidents althogher, but they can and have installed a system of warnings and countermeasures to mitigate the damage to the best of their abilities. When extreme precipitation affected Germany in 2002 and 2013 respectively, the country already had its own specific flooding-damage mitigation system in work. An essential component of such a system is an effective flood warning system. Two parts of any such warning systems are

> "the detection of potentially hazarduous situations" and

> "rules on when, how and whom to warn in case of rising flood water levels and what to communicate in order to activate organization in charge of civil protection"

(cf. Kreibich et al. (2017), p. 2078). Research was conducted in the context of such forwarding of flood-warnings in the aftermath of the floodings of 2002, 2013 and 2021 in the form of surveys of the German residents in the affected areas (cf. Kreibich et al. (2017))(cf. Thieken et al. (2022)). This research reveals that most residents indeed successfully received a warning (at least during the flood of 2013[3]), but that a substantial group did not know *how to operationalize what they received*[4]. In other words: These citizens did not know what to do. Presently, there is no specific concrete research as to the reasons for this lack of decisiveness or decision paralysis. An interesting result of the same research however is, that out of those men and women, who were affected by the same flood, whoever went through a flooding before, was significantly less likely to experience decision paralysis. It seems therefore, that one way to improve the capacities of German residents of handling a flooding, is to let them experience it once.

Effective flooding management can obviously not rely on such a crude strategy. Against this background and given the lack of research, a plausible working hypothesis however might read

---

[3]cf. @Kreibich.2017, p.2075)

[4]In Germany, the federal states are primarily responsible for flood management (cf. @Kuhlicke.2016, p. 2).

as *"The residents experienced decision paralysis, because the forwarded flood-warnings were to technical and insufficiently practical and specific"*. Grossly speaking, there seems to be a case for this in the aforementioned difference between experienced and inexperienced flooding-victims, where the residents first experience a flooding, and then contemplate afterwards on the actions that helped and those that did not work. If people create categories of flooding severity and successful countermeasures on an individual basis, then it could be, that dispatched flood-warnings may be more effective in mitigating flood-damage, if they similarily provide more clear and unambiguous instructions.

On the basis of this hypothesis, the present thesis paper asserts, that research into the transformation of existent flood-warning systems may be warranted and it asserts furthermore, that such research could focus on the implementation of a categorical warning typology[5].

The idea of a categorical warning typology refers to a system, where distinct warnings are made up of two-components:

An assessement of the level of flooding severity for a given spatio-temporal location. Flooding severity is to be measured on a categorical scale.

A clear and unambiguous instruction for the residents, that is distinct for each assessed categorical level of flooding severity.

This thesis contributes to a small area within this research field. *It specifically contributes research to the ways of defining and measuring the level of dichotomously conceptualized flooding severity on the basis of predictions of extreme precipitation.* A simple sketch of a dichotomous concept of flooding severity on the basis of the predicted binary precipitation status and the forwarding of associated distinct warning and action prompts, reads as follows:

---

[5]At the current stage, this is but a working hypothesis, that requires efforts of social scientists towards validation or falsification

Category of predicted flooding severity	Conditions
C_0	$\widehat{EPE} = 0$
	(non-extreme Precipitation)
C_1	$\widehat{EPE} = 1$
	(extreme Precipitation)

Table 1: A dichotomous typology of flooding severity

and

$$f(C_i) = \begin{cases} \text{No action,} \quad \text{if} \quad i = 0 & (2.1) \\ \text{Follow evacuation orders and seek higher ground immediately,} \quad \text{if} \quad i = 1 & (2.2) \end{cases}$$

where

$f()$ is a function that maps the assessed flooding category C_i to the action, that residents are advised to perform.

C_i is the label or category of flooding severity assessed on the basis of the predicted status of precipitation (extreme/non-extreme)

$\widehat{EPE}$ is a binary prediction of the status of the precipitation (extreme/non-extreme) at a spatio-temporal location of interest.

Following this concept, a level of flooding severity from a dichotomuous scala is assigned to a spatio-temporal location on the basis of the predicted status of precipitation at that location.

A flood-warning forwarding alogorithm would hence first generate a prediction of the status of precipitation, on whose basis a label of flooding severity is established. This label or level of flooding severity would subsequently be translated into one of two distinct warnings and actions prompts, which are finally forwarded to the residents at the spatio-temporal location of interest.

2.1 Research Question

Within this context of research on enhancing the German flood-management system and more precisely on the prediction of extreme precipitation for the assignment of dichotomous levels flooding severity, the *research question* of this thesis paper reads as follows:

Research Question

How do two classifier models compare performance-wise in predicting the occurence of extreme precipitation in Germany between 1996 and 2016[a], if

1. *the classification logic in one classifier is defined on the basis of inverse distance weighting (IDW) based interpolation (cf. Wikle et al. (2019))*

2. *the classification logic in the other classifier is defined on the basis of a generalized additive mixed model (GAMM) with a logit link? (cf. Zumel and Mount (2014), chapter 9 | cf. Wood (2017), pp. 161 | cf. Wikle et al. (2019), pp. 169 and 189 | cf. Fahrmeir et al. (2021), pp. 53 and pp.575)*

---

[a]To be precise, what the classifiers are compared for, is only the performance of predicting for 1996 to 2012. For 2013 to 2016, it is the performance of their forecast with reference to what the classifiers are compared for. This is a result of the split of the data into a training data subset that ranges between 1996 and 2012 and a hold-out subset that ranges from 2013 to 2016. For the purpose of simplicity however, such distinction will not be made in the rest of the paper

---

An answer to this question is provided in three steps and these built the paper's main body. The first step is an act of framing. The prediction of the binary status of precipitation as either extreme or non-extreme is framed as the binary classification of precipitation events as either extreme or non-extreme precipitation events. Towards that end, precipitation and extreme precipitation events are defined first and afterwards the process of modelling a binary classifier model is described while focusing on the importance and definition of a classification logic.

The second step is a consideration of appropriate choices for the definition of the classification logic. The exposition of appropriate choices is developed on the basis of their fit to the principles, that are usually asserted to govern the phenomenon precipitation. Here, IDW based interpolation and a trained GAMM are argued as appropriate alternatives in defining the classification logic.

In the final and third step, a competetive evaluation of the classifiers is conducted with reference to their performance in the prediction of extreme precipitation.

2.2 Data

As a preliminary to this main body however, a short introduction to the analyzed data is due.

The training data used in this thesis is metereological data sampled by 438 German weather stations whose spatial locations are depicted on figure 1.

It has a panel format and is irregular, in the sense that the number of observations are not the same for all 438 weather stations, while the years of observation range between 1996 and 2012. This training data is complemented by a second set of metereological data, the hold-out data. The years, in which the metereological observations in the hold-out data were made, range between 2013 and 2016. This split of the data into a training-data subset and a hold-out data subset is a common data-scientific practice (cf. Zumel and Mount (2014)). The two classifier models are built (trained) on the basis of the above mentioned training data, for whom a small representative slice is presented in table 2 below. Afterwards the hold-out data is utilized as input to the trained model and the correspondingly generated classifications of precipitation evaluated with reference to model performance. These names for the two subsets

Figure 1: Map of the 438 German weather stations in the training data. Own representation.

of the analyzed data reflect the purpose of the corresponding data split: The training data is the basis for the estimation procedure, during which the classifier models *learn* a specific classification logic. Data scientists however regard the training data set as only partially appropriate to assess the performance of a built / trained classifier model out of concern for bias. This concern for bias is essentially related to the problem of overfitting, which occurs, when the classifier model adopts the variation in the training data to closely during the training procedure. The result of such overfitting might be poor performance once the trained model is given a sample of data as input, that is different from the training-data sample. A common reason for overfitting is, that
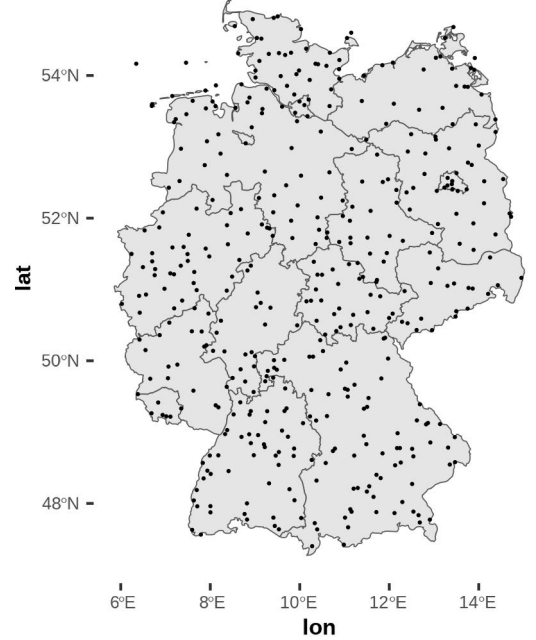
the classification logic is only valid for the data which the classifier model was trained upon, and not for the population data. The hold-out data thus is a dataset used instead of the training training data for model evaluation purposes; it is analyzed, how well the built models are "holding out" the data. This concludes a short introduction to the analyzed data. The following sections present the three aforementioned main-steps that lead up to the competetive evaluation of the two classifier models.

name	index	lon	lat	year	month	day	season	days_since_earliest_observation	rain	alt	mean_temperature	monthly_mean_precipitation	EPE
Aachen-Orsbach	1	6.0244	50.7983	2011	4	1	Spring	5569	0.0	230.99	13.3	41.15	0
Aachen-Orsbach	1	6.0244	50.7983	2011	4	2	Spring	5570	0.0	230.99	16.8	41.15	0
Aachen-Orsbach	1	6.0244	50.7983	2011	4	3	Spring	5571	9.9	230.99	10.8	41.15	0
Aachen-Orsbach	1	6.0244	50.7983	2011	4	4	Spring	5572	0.0	230.99	9.2	41.15	0
Aachen-Orsbach	1	6.0244	50.7983	2011	4	5	Spring	5573	0.9	230.99	8.8	41.15	0

[1] 'index' (=def.) indexnumber of the weather station
[2] 'mean_temperature' (=def.) daily mean temperature
[3] 'rain' (=def.) observed precipitation height in millimetre (mm)
[4] 'year','month' and 'day' (=def.) the year, month and day of an observation
[5] 'season' (=def.) the corresponding one among the annual seasons
[6] 'days_since_earliest_observation' (=def.) a type of julian date. The number of days, that have passed since the earliest observation in the whole training data.
[7] 'alt' (=def.) altitude level of weather station
[8] 'lon' and 'lat' (=def.) longitudinal and latitudinal degree of weather station
[9] 'name' (=def.) the city associated with the weather station
[10] 'EPE' (=def.) binary variable that indicates the true classification of the respective precipitation event. 1 := Extreme Precipitation Event. 0 := Non-Extreme Precipitation Event
[11] monthly_mean_precipitation (=def.) typical total precipitation of the respective month for the respective weather station

Figure 2: Rows 1-5 of the training data. Own representation

3 Framing the Prediction of Extreme Precipitation as a Binary Classification of Precipitation Events

The research question asks for the result of a competetive evaluation of the performance of the abovementioned classifier models in predicting extreme precipitation. What is meant with predicting extreme precipitation is the prediction of its' occurence and that in turn is a binary model response. The problem of modelling a binary response then is generally framed as a binary classification problem within (data) scientific literature (cf. Zumel and Mount (2014))(cf. Hosmer et al. (2013), p.18). In the next section, such framing will be conducted explicitly before introducing the two classifier models in detail.

3.1 Defining Precipitation Events and Extreme Precipitation Events

Let Extreme Precipitation Event (EPE) be the class of interest and let precipitation events be the subjects, for whom the membership inside this class is to be modelled. When precipitation is named extreme, what is from now on meant thus, is that the associated precipitation event is an extreme precipitation event. Let a precipitation event $PE_{s_i,t}$ haven been observed at the spatio-temporal location $s_i, t \in ST$, where s_i denotes the spatial location of the observation and t denotes the specific day of the observation. And let ST denote the spatio-temporal domain of the training data set. Then a model is to be trained, that either classifies $PE_{s_i,t}$ as extreme or non-extreme.

The precipitation events referred to in this context are non-empirical, that is to say, they are not directly observable. They are an artifact and a generated classification of a precipitation event as extreme is at best correctly associated with the warning and action prompt, that is forwarded on its basis. The definition of precipitation events generally involve either one or a combination of three usually measured empirical characteristics of precipitation, namely precipitation height (PH), duration and intensity of precipitation (cf. Trömel and Schönwiese (2007))(cf. Keller and Atzl (2014))(cf. Thieken et al. (2016))(cf. Kreibich et al. (2017))(cf. Wagenaar et al. (2018))(cf. Gimeno et al. (2022)). For the purpose of this thesis, precipitation events are defined on the basis of precipitation height only. There is in fact a great variation of its spatio-temporal distribution and the most common observation is, that there is no precipitation at all, or $PH_{s_i,t} = 0$. Although rare, precipitation height can still realize in extreme forms. When it does, it is likely to our detriment. The definition of extreme precipitation events naturally follows this logic and involves precipitation height as well. Some researchers take the approach to define extreme precipitation on the basis of the lower or upper tail of the space-time distribution of precipitation height, which is to say, on the basis of a specific quantile (cf. Trömel and Schönwiese (2007))(cf. Gimeno et al. (2022)). A $PE_{s_i,t}$ would be considered extreme then, if $PH_{s_i,t}$ exceeded the respective α-quantile value. Whether a precipitation event that has been classified as extreme on the basis of such an approach would however warrant any action at all, is highly questionable: If extreme precipitation is for example defined utilizing the 95th percentile as the threshold value, which is the approach in Trömel (2007), then the ratio of extreme precipitation events to all precipitation events would be indeed 5%; the

absolute number of extreme precipitation events however were 86392. It contradicts common sense to assert, that the federal states of Germany had to warn their citizens in a total of 86392 cases between 1996 and 2012, because extreme precipitation is occurring at the corresponding spatio-temporal location, that might cause a dangerous flooding. This thesis took an alternative approach. The calculation of the threshold value and the subsequent classification of precipitation events as extreme read as follows:

First, let $PH_{s,y,m,d}$ be the daily total precipitation height observed at spatial location s of a German weather station and y, m and d denote a year, month and day of the month (compared to previous notation, $y, m, d := t$ and $s_i := s$). Also, let $PH_{s,y,m}$ be the total precipitation height of the month m in the year y at location s.

This monthly total precipitation height $PH_{s,y,m}$ then is calculated as:

$$PH_{s,y,m} = \sum_{d \in m} PH_{s,y,m,d} \tag{3.1}$$

For the same spatial location s, the average of all monthly total precipitation heights for the month m observed in the different years in the training data $y\ in\ Y_{training} = \{1996, 1997, \cdots, 2012\}$ can be calculated as:

$$\overline{PH}_{s,m} = \frac{1}{N_{s,m}} \sum_{y \in Y} PH_{s,y,m} \tag{3.2}$$

where $N_{s,m}$ is the number of years y, for whom precipitation height has been observed in the month m at the location s.

The equation 3.2 calculates the average monthly total precipitation height of month m at location s across all years y. It represents the typical total precipitation height for that month at that location.[6] The logic of classification of precipitation events as extreme or non-extreme finally utilizes this typical monthly precipitation height as

$$PE_{s_i,t} \quad \text{is classified as extreme,} \quad \text{if} \quad PH_{s,y,m,d} > \overline{PH}_{s,m} \tag{3.3}$$

$$PE_{s_i,t} \quad \text{is classified as non-extreme,} \quad \text{if} \quad \overline{PH}_{s,m} > PH_{s,y,m,d} \tag{3.4}$$

where $\overline{PH}_{s,m}$ serves as the threshold / cut-off value. This concludes the definition of precipitation

---

[6]Note that the monthly typical total precipitation height is very likely unique for each spatial location.

events, both as extreme and non-extreme. The next section puts this way of defining the binary status of a precipitation event in a specific context, that is related to the concept of classification.

3.2   A Standard of Assessing the Truth of a Binary Classification of Precipitation Events

(cf. Hosmer et al. (2013), pp. 169) It was earlier mentioned, that two classifier models are to be trained to generate classifications. And that flooding severity labels are assigned to a spatio-temporal location on the basis of the generated classifications, such that ultimately a distinct warning and action prompt is forwarded to the residents at that location. Against this background, the role of the particular classification logic above is to provide a socalled Standard of Truth (S.o.T.)[7]. A S.o.T. establishes an unconditional and unambiguous reference for assessing the validity or correctness of a generated classification. This means: If the generated classification alignes with the classification logic described above (the S.o.T.), then it is labeled as *true / correct*. Another way to express the necessity for a S.o.T. is, that it defines the classifications, that the researcher *wants* the trained classifiers to generate. In the context of this thesis, this could be because classifications that align with the S.o.T. have proved to significantly correlate with either non-dangerous or dangerous amounts of precipitation in past research[8]. The easiest way for using the S.o.T. as a reference of *truth* practically is to determine a set of classifications who align with the S.o.T. for $PE_{s_i,t} \, \forall i, t \in ST$. These *true classifications* are best procured in the very beginning of the analysis and each respective classification generated by a trained classifier for $PE_{s_i,t}$ can then later on be compared with the corresponding true classification for the same $PE_{s_i,t}$. Such a set of true classifications can be constructed as a binary variable in the training data and the process of the construction can be expressed as follows:

Let $EPE_{s_i,t}$ thus denote the classification of $PE_{s_i,t}$ which aligns with the above mentioned S.o.T..

---

[7]It is neccessary to establish a S.o.T., as model evaluation were impossible, if a specific classification logic were not labeled unconditionally *true* at some point. It may not be obvious, but a classification logic defined on the basis of IDW based interpolation or a trained GAMM do not necessarily generate classifications, that are *true* by the S.o.T..

[8]It is an important part of future research, to establish, whether extreme precipitation events as they are defined in this chapter indeed warrant any action. Given the limitations to this thesis however, it shall be assumed, that this is the case, notwithstanding the relevance of a critical analysis of this assumption. Compared to aforementioned approach of the 95th percentile as threshold-value, the S.o.T. utilized in this thesis at generates a significantly smaller number extreme precipitation events (102)

Then

$$EPE_{s_i,t} = EPE_{s,y,m,d} = \begin{cases} 1, & \text{if } PH_{s,y,m,d} > \overline{PH}_{s,m} \\ 0, & \text{if } \overline{PH}_{s,m} > PH_{s,y,m,d} \end{cases} \tag{3.5}$$

where

EPE is a binary variable that indicates the true classifications of precipitation events as either extreme or non-extreme

$EPE_{s_i,t}$ ($EPE_{s,y,m,d}$) is the value of that variable for $PE_{s_i,t}$

$EPE_{s_i,t}$ ($EPE_{s,y,m,d}$) evaluates to 1, if the true classification of $PE_{s_i,t}$ is the classification as extreme

$EPE_{s_i,t}$ ($EPE_{s,y,m,d}$) evaluates to 0, if the the true classification of $PE_{s_i,t}$ is the classification as non-extreme.

3.3 Modelling a Binary Classifier model: The Concept of Classification Logic

Thusfar the prediction of extreme precipitation has been framed as a binary classification problem, a S.o.T. has been established and a new variable has been defined, by the means of which generated classifications can be evaluated for their correctness. This section will introduce a concept of modelling a binary classification.

The purpose of classifier models is to take a specific input and return a classification as output. In the context of classifying precipitation events, the classifier model of interest needs to return a binary classification of as either extreme or non-extreme. After training the classifier is completed, it will have a classification logic of the form

$$\widehat{EPE}_{s_i,t} = \begin{cases} 1, & \text{condition A} \\ 0, & \text{condition B (otherwise)} \end{cases}$$

$$\tag{3.6}$$
$$\tag{3.7}$$

where

$\widehat{EPE}_{s_i,t}$ is the generated classification of $PE_{s_i,t}$

$\widehat{EPE}_{s_i,t}$ evaluates to 1, if condition A is true

$\widehat{EPE}_{s_i,t}$ evaluates to 0, if condition B is true

a value of 1 indicates the classification as extreme

a value of 0 indicates the classification as non-extreme

When combined, the conditions 3.6 and 3.7 make the individual classification logic of the respective classifier model. These conditions involve the spatio-temporal location of precipitation events as model input (predictor). Other than that however there is a myriad of possible ways to define condition A and condition B. One such way is by the means of a self-contained model. For such self-contained models, one broad distinction can be made between deterministic and statistical models, and between scoring models and probability estimation models. IDW based interpolation in fact is a deterministic scoring method (cf. Wikle et al. (2019), pp.78), while a trained GAMM is a statistical probability estimation model (cf. Wikle et al. (2019), pp.101).

The following chapter shortly characterizes deterministic scoring models and statistical probability estimation models. In doing so, it concludes the first step of answering the research question, which is the framing of the prediction of extreme precipitation as a classification decision. That chapter is followed by the second step of the main contribution of this thesis, which goes more into detail of how IDW based interpolation and a trained GAMM are appropriate alternatives for defining the classification of precipitation events as extreme and non-extreme.

3.4 Defining a Classification Logic Based on Deterministic and Statistical Models

Essentially, regression is a powerful tool, whenever the objective is to cover up for our lack of knowledge of the exact dynamics of the data-generating process responsible for the observations we make, through the utilization of statistical distributions (cf. Zumel and Mount (2014))(cf.

Wikle et al. (2019), p. 6 and pp. 78)(cf. Fahrmeir et al. (2021), pp. 23). Such is the case with precipitation, where even the most complex models fail at times. It does however not neccessarily inherently follow, that statistical models are fundamentally superior to deterministic modelling approaches. IDW based interpolation for example is among the simplest deterministic approaches for modelling precipitation, while still accounting for a key feature of its distribution. The choice of this paper, to provide a competitive evaluation of the performance of both IDW based interpolation as a representative of a deterministic modelling approach, and training a GAMM as a representative of a statistical (regression) approach, reflects this belief. There are strengths and weakness to both which correspond to their respective concept:

In contrast to a statistical model, a deterministic model does not include any random variation. Neither the response nor the predictors are assumed to be randomly distributed and error variance and the uncertainty of model predictions are non-measurable as the direct consequence. A deterministic scoring model then predicts a score, based on a specific dependency of the model response on either other predictors or on itself, which are also scores. To simplify the conceptualization of a scoring model, a score in this context is considered a measurable feature, that is not a probability[9]. A major advantage of deterministic models are, that researchers do not need to inconvenience themselves with statistics (the fact that deterministic models can become quite complex notwithstanding). A researcher may just infer a mechanism from his or her own observations and try to approximate the true process by error- and trial, as is common in the natural sciences. The major drawback is the lack of a way to obtain an intuition for the uncertainty of the predictions. Prediction error may be measured; it is not possible though to estimate such error for future predictions (forecasts).

(cf. Zumel and Mount (2014))(cf. Wikle et al. (2019), p. 6)(cf. Fahrmeir et al. (2021), pp. 23) The statistical model does not have these problems. Random variation is present in the form of model response and predictors, such that either some or all are assumed to follow random destributions. For these models, measuring error variance and prediction uncertainty is possible and even an important step towards inference. A statistical probability estimation model predicts a probability of the occurence of an event, and in the context of the intended use of the model for

---

[9]In that respect, the thesis differs from Zumel and Mount (2014), where probabilities are included in the definition of scores.

the definition of conditions A(see 3.6) and B(see 3.7), such an event can be seen as the occurence of the classification of a precipitation event as extreme. The estimated probability is therefore $p(PE_{s_i,t}$ is extreme) while $1 - p$ then describes he probability of the classification of a precipitation events as non-extreme. The model response, which is ultimately modelled while training the model to predict the probabilities of interest, consist either of the estimations for exactly these probabilities of $PE_{s_i,t}$ being extreme $\forall\, s, t \in ST$ in the training data, or of a binary variable, that directly indicates the classification of a precipitation event as extreme or non-extreme. These advantages of the probability estimation model come at a price, which is the statistical sophistication required for its application. On the basis of the idea to model the classification logic on the basis of a self-contained model, the binary classifier can now be expressed as

$$\widehat{\mathrm{EPE}}_{s,t} = \begin{cases} 1, & \text{if} \quad f(x_{s,t}) > \theta \\ 0, & \text{if} \quad f(x_{s,t}) \leq \theta \end{cases} \qquad (3.8)$$
$$(3.9)$$

where

$\widehat{EPE}_{s,t}$ is the estimated binary classification decision

a value of 1 indicates, that the estimated classification of $PE_{s_i,t}$ is a classification as extreme

a value of 0 indicates, that the estimated classification of $PE_{s_i,t}$ is a classification as non-extreme.

Compared to 3.6 and 3.7, in 3.8 and 3.9 "condition A" and "condition B" have been expressed explicitly with logical expressions that involve the output of a function $f()$. This function can either be the above mentioned deterministic scoring model, in which case $f(s, t) = \widehat{PH}_{s,t}$ and where $\widehat{PH}_{s,t}$ is the estimated $PH_{s,t}$. If the function $f()$ is the statistical probability estimation model, then $f(s, t) = \widehat{p}(PE_{s,t}$ is in class EPE), while $\widehat{p}(PE_{s,t}$ is extreme) is the estimated probability of a classification as extreme (precipitation event). The logical expression is a comparison of the respective model output with a threshold value θ and θ is a specific precipitation height in the case of the scoring model and it is a probability in case of the probability estimation model. The classifier

model can be described thus as a two layered machine: A core model $f()$ projects an observation associated with the spatio-temporal location s,t onto either $\widehat{PH}_{s,t}$ or onto $\widehat{p}(PE_{s,t}$ is extreme), which are then compared with the respective threshold value. The classification finally depends on which of the two logical expressions in 3.8 and 3.9 evaluates to *true* for the given $PE_{s,t}$.

This concludes the first of the three steps towards an answer to the research question. In the course of the past first step, the prediction of extreme precipitation has been framed as a binary classification of precipitation events as either extreme or non-extreme. For that, non-extreme and extreme precipitation events have been defined in the context of a Standard of Truth (S.o.T.) and it has been demonstrated, how self-contained models in the form of IDW based interpolation and a statistical probability estimation model can be used for the definition of a classifiers classification logic.

4 Appropriate Choices for the Classification Logic

The following section presents IDW based interpolation and a trained GAMM as concrete approaches to such definition of the classification logic. The general argument is, that IDW based interpolation is an appropriate approach for doing so, as it accounts for an essential feature of precipitation, i.e. the spatio-temporal autocorrelation of precipitation height. It is also argued, that a trained GAMM is an appropriate mean towards the same objective, as it not only accounts for the spatio-temporal autocorrelation as well, but also allows the implementation of wideheld beliefs of factors that explain precipitation. These explanations revolve around the water cycle and consider *altitude*, *mean_temperature* and *the spatio-temporal location* as the driving factors of precipitation.

4.1 Principles That Govern Patterns of Precipitation

The spatio-temporal precipitation height is an empirical feature of the metereological phenomenon of precipitation and the scientific community usually discusses the following principles as governing

the data-generating process[10]:

<div align="center">

The Wather Cyle   Atmospheric Conditions

Topography   Geographic Location

</div>

The water cycle is currently the most fundamental frame for explaining precipitation. It describes the movement of water between the earth's surface and the atmosphere in the form of a cyle, where water first evaporates, condenses into clouds, and when droplets within the clouds finally reach a weight sufficiently great, the droplets fall back onto the earths surface to return the water. Atmospheric conditions, topogrophy and the geographic location are then tied into this cyclic pattern to explain the individual mechanisms:

> While the evaporated water condenses inside higher layers of the sky, air temperature is asserted to positively correlate with the amount of water, that the air can hold. The higher the temperature, the greater the precipitation height. Yet air temperature is not only a volatile atmospheric condition; researchers have established, that the geographical variation between locations that are closer to and further away from the equator matters systematically with reference to air temperature. As the intensity of the sunlight increases towards the equator, air temperature increases as well.

> Topography is in general concerned with the shape of the earths surface. Metereologists widely assume, that mountains intervene in the movement of air. The idea is, that wind causes air to rise alongside the mountains and hence to reach higher layers of the atmosphere and that this results in increased precipitation at the windward side of the mountain and relatively less precipitation on the leeward side of the mountains. This outline of the principles that govern the phenomonen of precipitation is neccessarily rudimentary. It shall serve as a gross theoretical background for the later on following explanation of the benefits of training a GAMM to predict the probability of a precipitation event being extreme.

---

[10]It should be noted, that the idea, that there is a spatio-temporal point-process responsible for virtually any observable data, including precipitation data, is extensively covered by Diggle (see @Diggle.2023). The statistics to act upon that notion however exceeds the current level of statistical education of the author of the thesis.

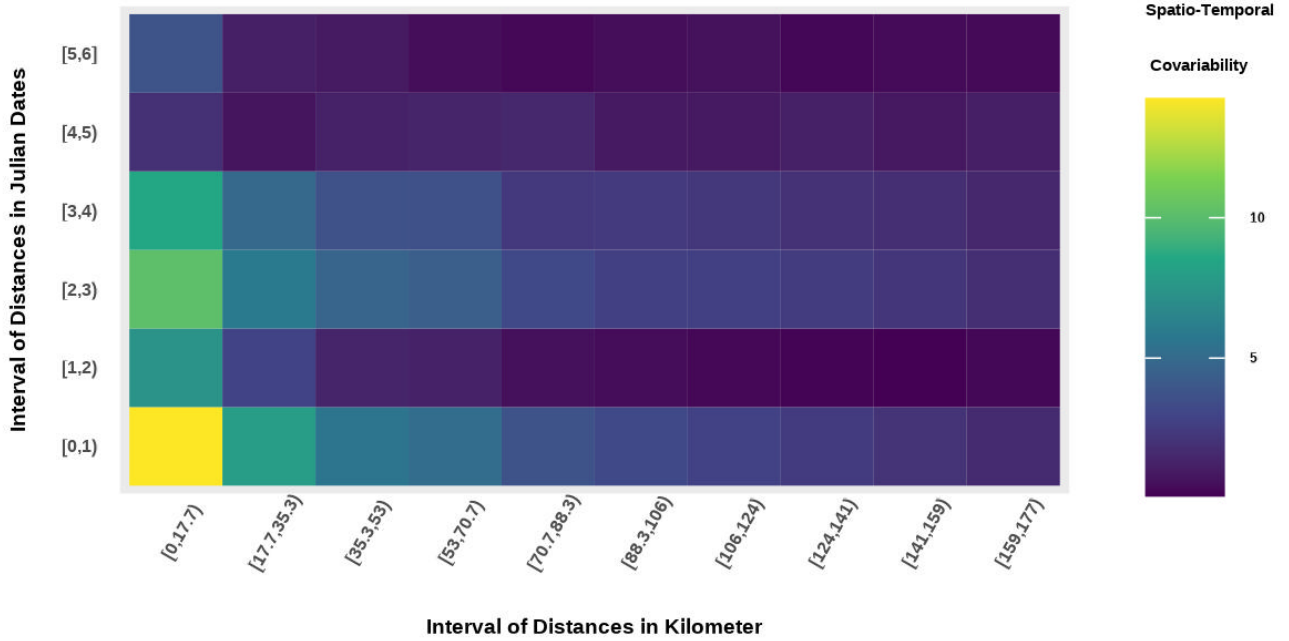## 4.2 Spatio-Temporal Autocorrelation of Precipitation Heights

Figure 3: Variation in the Spatio-Temporal Covariance.

Another key characteristic alongside these commonly asserted patterns of precipitation is the strong autocorrelation of precipitation heights. Such autocorrelation fundamentally refers to the striking similarity of observations of precipitation heights at closeby locations. A different way to phrase this is, that precipitation heights cluster spatio-temporally. In real life, these clusters express as something that can be observed anywhere in the world: On a rainy day, it will rain or snow in similar amounts up to several kilometers from any spatial location. And the same logic applies temporally as well. If it rains for instance now, it is more likely that it rained some minutes or hours ago than yesterday, or in some minutes or hours than tomorrow. One can even argue that a human could see a manifestation of these clusters in reality, if he or see were to hover in exactly the right distance above the clouds to see were precipitation happens and for how long it rained/snowed. This observation of the spatio-temporal autocorrelation has a profound implication for any attempt to predict extreme precipitation: If a model, that is not capable of accounting for this clustered nature of precipitation, is trained to predict extreme precipitation, it will inevitably generate biased results. They will be biased in the sense, that the prediction error for precipitation in such clusters will systematically be too large, wether the precipitation

was over- or underpredicted. While the next sections go on to show how IDW based interpolation and GAMMs are capable of handling and identifiying such autocorrelation, its' existence in the training data will be demonstrated as a preliminary.

The chosen way to demonstrate the spatio-temporal autocorrelation of precipitation heights in the training data is an analysis of the variation of spatio-temporal covariances. The corresponding mathematical equation is provided in the appdendix (cf. equation 7.19). This variation is visualized[11] in the figure 3.

(cf. Auer and Rottmann (2020), pp. 90)(cf. Fahrmeir et al. (2021), p.692) Covariances generally indicate, whether and how two variables co-behave. Behaviour in this context is to be understood in relation to the variables' mean value. For each observation, the difference to the mean of its respective variable is calculated. For a pair of observations of the two variables, the cross-product of these differences are calculated next. If both observations exceed their respective mean, the cross product is positive. And the positive cross-product grows, as the difference of both observations to their respective mean grows. Such a cross-product is negative, if one of both observations exceed, while the other deceed their respective mean. The covariance finally is the sum of those cross-products, where a positive value hence indicates, that pairs of observations of both variables tended to mutually exceed or deceed their respective mean, or behave similiarily in other words. And where a negative value indicates, that pairs of observations behaved in exactly the opposite way.

As a concept, spatio-temporal covariance is an extension of this broader concept of measuring similarity in the behaviour of two variables. Instead of assessing how two variables co-behave, it measures, how observations of the same variable co-behave, if they were made at different spatio-temporal locations. It is in this context paramount, that a variety of such covariances are calculated for different groups of observed precipitation heights. These groups can be based on intervals of spatial and temporal distance. All pairs of spatio-temporal locations then for whom the spatial and temporal distances fall inside a pair of spatial and temporal interval would be grouped. Such grouping is necessary, as a single covariance calculated upon the entirety of all pairs of observations couldn't be reasonably interpreted in terms of spatio-temporal autocorrelation: If

---

[11]Visualization was created by the author after calculating the corresponding data in R by means of tidyverse data-manipulation functions. The visualization itself was created by the author applying functions of the R package `ggplot2`. The associated R script is available in the GitHub repository.

the covariance for instance is highly positive, it could be, because observations at either close or distant spatio-temporal locations co-behave. If it were highly negative, it could be again, because these paired observations behave opposingly either for close or distant spatio-temporal locations. There would be no way to tell, whether pairs at close or distant locations are the cause. The only way to utilize covariance appropriately therefore is to group the spatio-temporal observations first. Such grouping was conducted on the base of intervals of spatial and temporal distances. The procedure for this reads as follows: Spatio-temporal observations of precipitation height are first paired based on all combinations of spatio-temporal locations. Based on the calculated spatial[12] and temporal[13] distances between the paired observations, they are then assigned to intervals of spatial distance and temporal distances. After determining all combinations of intervals, the covariance is calculated only for those pairs of observations within the same combination of intervals of spatial and temporal distance. Any such group of pairs for whom the spatio-temporal distance is within the same duo of distance intervals essentially consitutes a spatio-temporal cluster. An analysis of the variation of the covariance of these clusters finally enables an assessment of the existence of spatio-temporal autocorrelation. If spatio-temporal autocorrelation exists for the distribution of precipitation heights in the training data, then the covariance is expected to decrease

for the same interval of temporal distance with expanding intervals of spatial distance

for the same interval of spatial distance with expanding intervals of temporal distance

along a simultaneous expansion of both intervals of spatial and temporal distance.[14]

Considering the visualization of the variation of covariance between the respective interval-based

---

[12]The distance between two spatial locations can be measured in different ways. A common way is to calculate the euclidean distance. A drawback of the euclidean distance is, that it assumes a flat space. The earths surface is characterized by curvature however. Hence a better measure was identified in the Haversine Formula. Using the Haversine Formula, the spatial distance between two spatial locations was measured in kilometer while accounting for the curvature of the earths surface. For the implementation of the Haversine Formula, the author wrote different R functions, whose source code can be found in the Scripts of type `Haversine-*.R`

[13]The temporal distance between two observations is measured as the difference between their respective julian dates.

[14]It should be noted, that it does not matter whether the covariance is negative or positive. Both reflect a pattern of behaviour and the higher the absolute value, the greater the similarity of behaviour or cohesiveness of the cluster in other words. What really matters is the difference between clusters of close locations and clusters of distant locations. The absolute values have thus been visualized instead of the original positive or negative values.

clusters (cf. figure 3), all expectations for spatio-temporal autocorrelation are met. The absolut value of the covariance is the greatest for observations that are no more than 17.7 kilometers and no days apart. Aligning with the above defined expectations, the covariance indeed decreases, if either the spatial or temporal interval expand while the other remains the same. The empirical covariance also decreases, if both intervals expand simultaneously.

The existence of spatio-temporal autocorrelation is therefore safe to assume and the following sections show how it can be accounted for when defining the classification logic.

4.3   IDW based Interpolation as an Appropriate Choice for the Classification Logic

Everything is related to everything else, but near things are more related than distant things.

- Waldo Tobler (1970)

In contrast to regression, Inverse Distance Weighting based interpolation (cf. equation 7.15) does not model the conditional expected value of a variable like precipitation height, nor the conditional probability for a specific observation to be a member of a certain class. (cf. Bivand et al. (2013) , pp. 215)(cf. Wikle et al. (2019) , pp. 78) It models the variable based on a characteristic of its own univariate distribution: its neighborhood in terms of spatial and temporal distance. For any spatio-temporal location, the existing available observations are evaluated against their relative spatial and temporal distance to the spatio-temporal location[15], for whom the precipitation height is to be interpolated. The model prediction of the variable for that respective location is then determined as a weighted average of all available observed values minus any existing observations at the location in question.

---

[15]As opposed to the chapter above, that is concerned with spatio-temporal autocorrelation, spatial and temporal distances are not measured seperately for IDW-based interpolation , but jointly on the basis of euclidean distance.

Inverse Distance Weighting based Interpolation

$$\widehat{rain}_{s_i,t_i} := \sum_{s_j,t_j \in ST} \left(w_{(s_i,t_i),(s_j,t_j)} \cdot rain_{s_j,t_j} \right) \tag{4.1}$$

Inverse Distance Weighting based interpolation is a representative of a deterministic scoring model. It assumes, that an observation of a score like precipitation height at any spatio-temporal location can be calculated as an average of all other observations, but such, that observations that are relatively distant have less influence on the predicted value than close observations (cf. equation 4.1)[16]. The quote cited above is a statement of Waldo Tobler, widely known as *Tobler's first law of geography* and Inverse Distance Weighting based interpolation fundamentally can be considered a practical implication of that notion. Tobler's law is one way to approximate the spatio-temporal autocorrelation of precipitation height, which was extensively demonstrated above. Hence, within limitation, IDW based interpolation can be appropriately involved in the definition of the classification logic[17]. The classifier model based IWD based interpolation[18] finally reads as

$$\widehat{EPE}_{s,t} = \begin{cases} 1, & \text{if } \widehat{rain}_{s_i,t_i} > \overline{PH}_{s,m} \\ 0, & \text{if } \overline{PH}_{s,m} > \widehat{rain}_{s_i,t_i} \end{cases} \tag{4.2}$$

where

$\widehat{rain}_{s_i,t_i} > \overline{PH}_{s,m}$ is a logical expression, which evaluates to *true* for a given spatio-temporal location, if the predicted precipitation height exceeds the typical monthly precipitation height (cf. equation 3.2). And if the logical expression evaluates to true, than the associated precipitation event is classified as extreme.

$\overline{PH}_{s,m} > \widehat{rain}_{s_i,t_i}$ is a logical expression, which evaluates to *true* for a given spatio-temporal location, if the typical monthly precipitation height exceeds the predicted precipitation

---

[16]For a complete breakdown of the equation, confer equations 7.15,7.16 and 7.17

[17]Future research may test this assumption more in-depth. In the context of this thesis, this reasoning of the appropriateness of IDW based interpolation as a classifier core shall suffice.

[18]The R script for the calculation of the confusion matrix and the associated performance indicators is named `Inverse-Distance-Weighting-Calculation` and can be found at .\BachelorThesis\Scripts\Data-Manipulation\Training-Data\Inverse-Distance-Weighting-Calculation.R"

height. And if this logical expression evaluates to true, than the associated precipitation event is classified as non-extreme.

4.4 A Trained GAMM as an Appropriate Choice for the Classification Logic

The Generalized Additive Mixed Model with a Logit Link function relies on an algorithm to identify the dependence of precipitation height on mean temperature, altitude, spatial and temporal location. This dependence is broadly speaking expressed as the estimated coefficients and respective smoothing terms (cf. Wood (2017)}|*Fahrmeir et al. (2021), pp. 53 and pp.575)*. This section argues the appropriateness of a trained GAMM for the definition of a logic of the classification of precipitation as extreme or non-extreme. It does so progressively, by demonstrating how interventions in the regression model structure for the purpose of creating a suitable regression model, eventually lead to the GAMM. The starting point is classical linear regression, which is generally a good start for statistical modelling (cf. Zumel and Mount (2014), chapter 7.).

4.4.1 Classical Linear Regression

(cf. Fahrmeir et al. (2021), pp. 24) The classical linear regression involves modeling the model response as the conditional expected value of a variable, with this expected value being dependent on one or more covariates in a linear fashion, plus an error term.[19] The random error terms are assumed to be identically and independently Gaussian-distributed. The covariates form a linear predictor such that, based on the modeled dependence of the response on both the random errors and the linear predictor, the response is conceptualized as a random variable as well, whose distribution, beyond adopting the i.i.d. features, is additionally influenced by the linear predictor. Importantly, as the error terms conceptually account for random, unsystematic differences between the response and the linear predictor, as well as measurement errors, they are assumed

---

[19]For the model equation, see equation 7.1

to be independent not only of themselves but also of the linear predictor. These characteristics collectively render linear regression fundamentally inadequate for the present purpose, which is to define a logic of the classification of precipitation as extreme or non-extreme. Evidence has been provided above for the existence of spatial, temporal, and spatio-temporal dependencies within the distribution of precipitation heights and this clashes with the assumed independence among the responses and among the error terms. If the responses are non-independent (autocorrelation), then these dependencies need to be captured through the linear predictor or the error terms. As the linear predictor, however, is inadequate for modeling any other than linear effects of the covariates on the responses, these dependencies would conceptually be assumed to be reflected in the error terms. The dependencies would therefore reflect in the error terms, particularly as autocorrelation of the error terms, which contradicts the assumption of i.i.d. error terms mentioned above. Linear regression hence is by concept inadequate to meet the requirements of contributing to the definition of a logic of the classification of precipitation, in contrast to IDW based interpolation as demonstrated above. To this inadequacy finally also adds the incapability of modelling non-linear dependencies[20] (as the name already suggests).

4.4.2 Logistic Regression Models as *Generalized Linear Models*

Albeit these defeciencies, a trained classical linear regression model can still generate predictions of precipitation height for a spatio-temporal location of a precipitation height, and is thus a good starting point. And with changes to the model structure, it can be transformed into a probability estimation model, the logistic regression model. Deriving a logistic regression model from the linear regression model itself does not resolve the above mentioned shortcoming. As a probability estimation model however, it provides a basic framework such as the estimation of the model coefficients per *maximum likelihood* (cf. Hosmer et al. (2013), pp.8)(cf. Wood (2017), pp. 74) and *iteratively re-weighted least squares* (cf. Hosmer et al. (2013), pp. 20)(cf. Wood (2017), pp. 249) in particular, which applies to the GAMM as well.

(cf. Hosmer et al. (2013)) A trained logistic regression model can generate an estimation of the probability, that a precipitation event is extreme (an EPE). With the model equation of linear

---

[20]These dependencies are usually considerer non-linear; for non-linear dependency of precipitation on elevation cf. Chu (2012), p.3177

regression (confer equation 7.1) as the starting point, it can be obtained through a *generalization.* For that purpose, firstly the error term is dropped from the right-hand side, so that compared to linear regression, the model response is no longer modelled as dependent on the linear predictor plus an error term, but only on the linear predictor.

The core acts of the generalization then, which effectively allow to model probabilities $p(PE_{s,t}\,is\,extreme) = p_{s,t}$ as dependent on a linear predictor, involve

- observations of these probabilities in the data

- a logit-function $logit(p_{s,t}) = \log\left(\frac{p_{s,t}}{1-p_{s,t}}\right)$, which projects the observed probabilities onto log-odds (also called logit) and

- a sigmoid function S of the form $S\left(logit(p_{s,t})\right) = \frac{1}{1+e^{logit(p_{s,t})}} = p_{s,t}$, that is the inverse of the logit-function and returns probabilities with log-odds as input

and the concept is as follows:

The logit-function projects the probability values on a continuous scale, such that the resulting continuous log-odds can be modeled as dependent on a linear predictor. Since the log-odds are not the variable of interest however, the logit's inverse function, S, is applied on both sides of the model equation. As a consequence, within the logistic regression model, the probabilities $p_{s,t}$ as the variable of interest are modeled as dependent on the inverse logit function S with the logg-odds as the model argument. And since the log-odds were modeled as dependent on a linear predictor in the first place, the model equation[21] presents itself as

Logistic Regression Model Equation

---

[21]For a in-depth breakdown of the model equation and the concept of logistic regression, see the Appendix at 7.2

$$p(PE_{s,t} \text{ is extreme}) = \mathbf{S}\left(\log\left(\frac{p_{s,t}}{1 - p_{s,t}}\right)\right)$$

$$= \mathbf{S}\left(\begin{array}{l} \beta_0 + \beta_1 \text{Mean Temperature}_{s,t} + \beta_2 \text{Altitude}_{s,t} \\[4pt] + \beta_3 \text{Longitude}_{s,t} + \beta_4 \text{Latitude}_{s,t} \\[4pt] + \beta_5 \text{Julian Date}_{s,t} \end{array}\right) \qquad (4.3)$$

Above, the probabilities $p(PE_{s,t} \text{ is extreme}) \; \forall \, (s_i, t) \in ST$ are thus modelled to depend on a linear predictor. In contrast to linear regression however, the coefficients can no longer be estimated per *least squares estimation*. *Maximum Likelihood Esitmation* is applied instead, which is less a particular estimation method than an idea or type of how to estimate model coefficients (cf. Hosmer et al. (2013), pp. 8). Maximum likelihood estimation revolves around the concepts

parameters of population distribution   sample distribution

Single Data-point Likelihood   Likelihood of entire dataset

Maximization of the Likelihood Function   Alogrithm for Maximization

Closed-Form Solution and Partial Derivatives   Optimizing iteratizing algorithm

Iteratively Re-weighted Least Squares

4.4.2.1   Maximum Likelihood Estimation   (cf. Hosmer et al. (2013), pp. 8)(cf. Wood (2017), pp. 74) On multiple occasion thusfar, the term *data-generating process* has been mentioned. It is used on the notion, that there is a distinct and identfiable point-process somehow responsible for the distribution of precipitation. This relates to the fundamentally different approaches of estimation that least squares estimation and maximum likelihood estimation represent: Least squares estimates the model coefficients on the basis of sample data and assumptions of parameters of the distribution of erros as input. In other words: *The direction of the estimation process is from sample to population.* The sample coefficients are considered unbiased and efficient estimates of the population parameters, such that hypotheses of the model coefficients concerning the process

that generated the population data can be falsified (and accepted preliminarily) (at least from the frequentistic point of statistics (cf. Efron (1978))). While Maximum Likelihood Estimation and Least Squares Estimation have that frequentistic perspective of statistics in common, the direction of inference is reversed. In Maximum Likelihood Estimation, *the direction of the estimation process is from population to sample.* The estimation process assummes different values for the population parameters (model coefficients). And it does so, until the algorithm found a set of parameters, such that it is *maximum likely*, that the sample has been drawn from the population that is characterized by these specific parameter values. In other words: Instead of deciding upon population parameters, that make sense given the sample parameters, maximum likelihood sets population parameters, for whom the sample makes sense.

In theory, there are two distinct approaches for the application of Maximum Likelihood Estimation for the task of determining model coefficients, such as the coefficients of the logistic regression model above. One approach and the more obvious of both approaches is an attempt at a closed-form solution, that involves solving a system of partial derivatives of the likelihood function. Likelihood refers to a measurement of the above mentioned probability to draw or observe the sample given the population data as described by setting the coefficients of the logistic regression model. There are two types of likelihood: The likelihood of observing a single point of the sample data given the coefficients and the likelihood of observing the entire data, which is the product of all individual likelihood values and ultimately is to be maximized. The formula for the likelihood of an individual sample data point distinctly depends on the respective regressoin model model, whose coefficients are to be determined. For the above mentioned logistic regression model, the likelihood function reads as:

Likelihood-Function

$$L(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5) = \prod_{s,t \in SP} P(EPE_{s,t}|\mathbf{x}_{s,t}, \beta) \tag{4.4}$$

$$= \prod_{s,t \in SP} p_{s,t}^{EPE_{s,t}} (1 - p_{s,t})^{1 - EPE_{s,t}} \tag{4.5}$$

where

$L(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5)$ is the likelihood of observing the entire sample data given the

coefficients $\beta_0, \ldots, \beta_5$. Maximum likelihood is set out to maximize that likelihood.

$P(EPE_{s,t}|\mathbf{x}_{s,t}, \beta)$ is the likelihood of observing the true classification of the precipitation event at spatio-temporal location s, t given the observed values of the five predictors at that location and the determined coefficients. From a mathematical perspective, it is equivalent to a conditional probability.

$p_{s,t}^{EPE_{s,t}}$ is the probability of observing a true classification of the precipitation event as extreme at spatio-temporal location s, t.

$(1 - p_{s,t})^{1 - EPE_{s,t}}$ is the probability of observing a true classification of the precipitation event as not-extreme at the same location

$p_{s,t}^{EPE_{s,t}} \times (1 - p_{s,t})^{1 - EPE_{s,t}}$ is the probability of observing the actual true classification of the precipitation event at that point, whether it is a classification at extreme or non-extreme

Again, for the maximization of that Likelihood function, the closed form solution would be the initial obvious choice and it consists of solving a system of partial derivatives of the same function to the coefficients, which reads as

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\beta}} = \mathbf{X}^{\top}(\mathbf{EPE_{s,t}} - \mathbf{p}) \overset{!}{=} 0 \tag{4.6}$$

where

$\boldsymbol{\beta}$ is the vector of coefficients $\beta_0, \ldots, \beta_5$

$\frac{\partial \mathcal{L}}{\partial \beta}$ the gradient of the Log-Likelihood function with respect to the coefficients $\boldsymbol{\beta}$

$\frac{\partial \mathcal{L}}{\partial \beta} \overset{!}{=} 0$ is a system of first-order optimization conditions

$\mathbf{X}^{\top}$ is the transposed design matrix

$\mathbf{EPE} - \mathbf{p}$ is the difference between two vectors, where $\mathbf{EPE}$ is the vector of true classifications of precipitation events $PE_{s,t} \; \forall \, (s,t) \in ST$ and $\mathbf{p}$ is the vector of predicted probabilities $p_{s,t}$. These are called *likelihood equations*

(cf. Hosmer et al. (2013), p. 8-10). The problem is, that these predicted probabilities $p_{s,t}$ are not available. The logistic regression model thus is only one of many cases, where the obvious choice of maximizing the Likelihood function per solving a system of first-order optimization conditions or a closed form solution in other words, is not applicable (cf. \cite{Hosmer.2013).

The second approach abovementioned approach is more a family of approaches, which all have in common, that they set the coefficients along the execution of some type of iterative optimizing algorithm. *Iteratively Re-Weighted Least Squares* is such an algorithm. It usually is the go-to algorithm applied for maximum likelihood estimation of model coefficients in logistic regression and involves iteratively solving a least squares problem (cf. Hosmer et al. (2013), p.20). (cf. Wood (2017), pp. 105-107) Towards the objective of estimating the parameters $\boldsymbol{\beta}$ of the above described logistic regression model and given the Likelihood function $L(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5)$ of that model, Iteratively Re-Weighted Least-Squares proceeds both iteratively and cyclic. (cf. Wood (2017), p. 207) It begins (the iteration) with an initial set of model coefficients $\beta_0, \ldots, \beta_5$ (iteration count $c = 0$) and calculates the weights $w_{(s,t),c}$ for each data point (s,t). It uses these weights for an update of the previous (initial) coefficient values, which reads as

Update of Logistic Regression Model Coefficients during iteration c+1

$$\boldsymbol{\beta}_{c+1} = \boldsymbol{\beta}_c + (\boldsymbol{X}^\top \boldsymbol{W}_c \boldsymbol{X})^{-1} \boldsymbol{X}^\top (\boldsymbol{EPE} - \boldsymbol{p}_c) \tag{4.7}$$

where

$\boldsymbol{\beta}$ is the vector of model coefficients at iteration c

X is the design matrix. It encompasses the values of the predictor variables of the logistic regression model, i.e. *mean temperature, altitude, longitude, latitude, julian date.*

W_c is a diagonal matrix containing the weights $w_{(s,t),c}$ for observations assigned to the different spatio-temporal locations at the c-th iteration. The weight for each data point is given by $w_{(s,t),c} = p_{s,t}(1 - p_{s,t})$, where $p_{s,t}$ is the probability of EPE $= 1$ for spatio-temporal location (s,t) at iteration c.

$\boldsymbol{EPE}$ is the vector of true classifications of the precipitation events for all spatio-temporal locations

p_c is the vector of estimated probabilities to observe the true classifications EPE

The equation above describes the process of a single update of the model coefficients. After the update, the newly update coefficients are used to procure a linear predictor $n_{(s,t)} = x_{s,t}^\top \ldots$, where $x_{s,t}$ is a vector of the predictor values at spatio-temporal location (s,t) from the design matrix. This linear predictor used to calculate new estimations of the probabilities $\boldsymbol{p}$ using the logistic function, such that $p_{s,t} = 1/(1 + exp(-\eta_{s,t}))$. The weights then are updated per $w_{(s,t)} = p_{s,t}(1 - p_{s,t})$ and a new diagonal matrix W_c is obtained. At this point, a new iteration of the update of the model coefficients of the form equation 4.7, begins. Iteratively Re-Weighted Least Squares hence is a cycle of updating the model coefficients, using the updated model coefficients to procure new weights and updating the model coefficients *given* the new weights once again. The number of updates depends on convergence, or in other words, the number of iterations, after which a specific scoring function, that has been monitoring this iterative process, reaches a specific value (cf. Wood (2017), p. 107).

Once again, it is important to stress the essential difference between this iterative optimizing algorithm and the closed-form solution per solving the system of first-order optimization conditions illustrated above: For the closed-form solution, it would have been necessary to have estimations for the probabilities $p_{s,t} \; \forall \; (s,t) \in \; ST$ right from the start. In logistic regression however, these are conceptualized to depend on the model coefficients in the form per $p_{s,t} = 1/(1 + exp(-\eta_{s,t}))$; in other words: maximizing the the likelihood function per closed-form solution requires, that the result of the maximizing process - the model coefficients - are available in the first place. Iteratively Re-Weighted Least Squares however does not have this problem, since it starts the iterative process with a set of coefficient values, a *prior*, if one were to use Bayesian terminology.

4.4.3 Generalized Additive Mixed Model with a Logit-Link Function

The previous logistic regression subsection discussed maximum likelihood estimations and the inapplicability of a closed-form solution, which is replaced by Iteratively Re-Weighted Least Squares. The aim of this section is to show that a trained GAMM is a suitable choice for defining classification logic. Logistic regression however, like linear regression, has limitations in classifying precipitation events, as it fails to consider spatio-temporal dependencies and non-linear

relationships with predictor variables. The subsequent subsection highlights the commonalities between the GAMM and logistic regression models, which include their roles as probability estimation models, their reliance on a linear predictor for the response, and their shared use of maximum likelihood estimation through Iteratively Re-Weighted Least Squares as the estimation method. Maybe more importantly however, it stresses the distinct differences between the GAMM and the logistic regression model, that allow the GAMM to account for the spatio-temporal autocorrelation of precipitation as well as the non-linear relationships between the model response and the predictor variables.
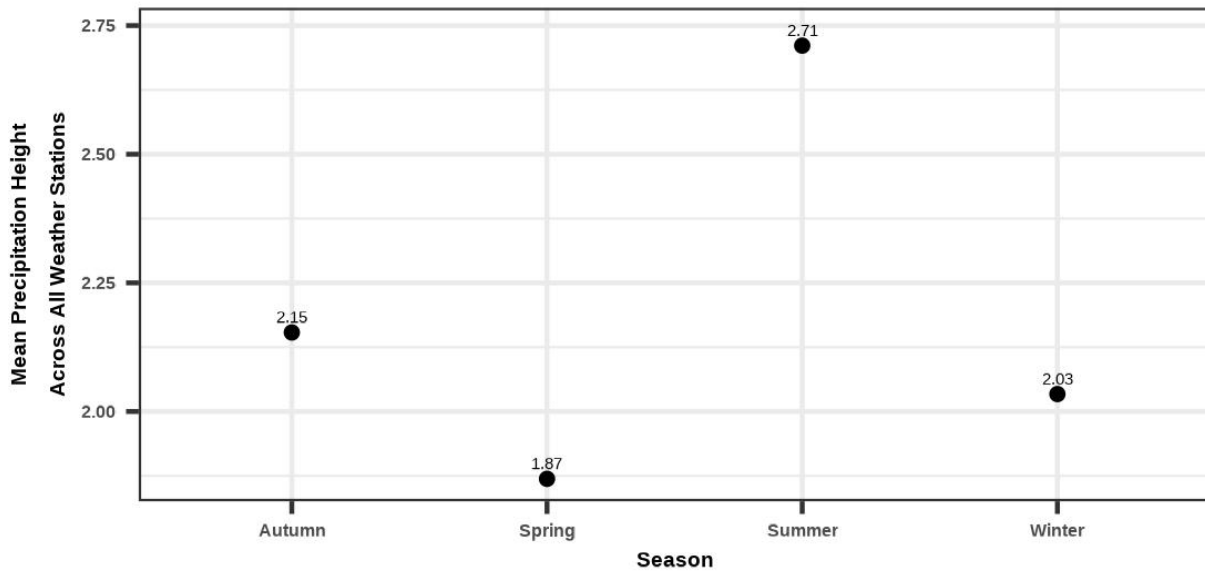

Figure 4: Seasonal Differences of Mean Precipitation

As should be the case with any regression model, an attempt was made to structure the definition of the gross generalized additive mixed model structure on reasonable prior assumptions on the dependencies of model response on the predictors. Various acts of exploratory data analysis have been conducted[22]. Alongside the spatio-temporal autocorrelation as demonstrated above (confer figure 3), the most conclusive results are an apparent difference between the mean of precipitation

---

[22]For the corresponding thematically named R Scripts, confer the directories `Data-Manipulation` and `Plotting-(visualization)` and then the subdirectory `Training-Data`(for the Data-Manipulation directory) and the subdirectory `Entirety-of-data-(no-subsetting)`(for the directory of r scripts about visualizations)

across all years and weather stations between the four annual seasons, which is presented in figure 4.

Based on the findings from the exploratory data analysis, the GAMM was designed to account for the observed seasonal variability in precipitation patterns. By incorporating random effects and random slopes based on the categorical variable "season," the model can capture season-specific deviations in both intercepts and slopes, ensuring that the influence of seasonal fluctuations on precipitation patterns is appropriately considered in the analysis. This choice of model structure not only improves the accuracy of the model in representing the underlying precipitation patterns but also provides a comprehensive understanding of the factors contributing to extreme precipitation events and their seasonal variations.

The implementation of smooth terms each for the spatial location and the temporal location is meant as well to capture spatio-temporal autocorrelation and to capture non-linear dependencies as well as clusters.

In summary, the GAMM structure was informed by the prior assumptions and the results of the exploratory data analysis, which highlighted the importance of accounting for spatio-temporal autocorrelation and seasonal variability in precipitation patterns. Incorporating random effects and random slopes based on the "season" variable allows for a more robust and nuanced analysis of the factors influencing extreme precipitation events and their seasonal variations, while smooth terms capture non-linear dependencies and changes of dependecies across spatio-temporal clusters. The model equation that results from those implementations is described in the following subsection.

4.4.3.1 GAMM: Model Equation and Model Concept

The GAMM that was trained in this thesis study reads as follows:

Generalized Additive Mixed Model - Model Equation

$$\text{logit}(p(PE_{s,t} \, is \, extreme)) = \beta_0 + f_{alt}(\text{Altitude}_{s,t}) + f_{lon}(\text{Longitude}_{s,t}) + f_{lat}(\text{Latitude}_{s,t})$$

$$+ f_{jul}(\text{Julian Date}_{s,t}) + f_{temp}(\text{Mean Temperature}_{s,t})$$

$$+ u(\text{Season}_{s,t}) + f_{random \, slope \, alt}(\text{Altitude}_{s,t}|\text{Season}_{s,t})$$

$$+ f_{random \, slope \, jul}(\text{Julian Date}_{s,t}|\text{Season}_{s,t}) + \epsilon$$

$$= \beta_0 + \sum_{i=1}^{5} \beta_i f_i(\boldsymbol{x}) + \sum_{j=1}^{4} u_j b_j(\boldsymbol{x}) + \sum_{k=1}^{M} v_k c_k(\boldsymbol{x}) \tag{4.8}$$

(cf. Zumel and Mount (2014), chapter 9 | cf. Wood (2017), pp. 161 | cf. Wikle et al. (2019), pp. 169 and 189)

This model equation can be broken down in the form of

$$\textbf{Linear Predictor} \qquad \beta_0 + \sum_{i=1}^{5} \beta_i f_i(\boldsymbol{x}) + \sum_{j=1}^{4} u_j b_j(\boldsymbol{x}) + \sum_{k=1}^{M} v_k c_k(\boldsymbol{x}) \tag{4.9}$$

$$\textbf{Thin Plate Regression Spline} \qquad f_i(\boldsymbol{x}) = \sum_{(s,t) \in ST} w_{i,(s,t)} U(\|\boldsymbol{x} - \boldsymbol{x}_{i,(s,t)}\|) + \sum_{k=1}^{m} \alpha_{ik} g_{ik}(\boldsymbol{x})$$

$$\tag{4.10}$$

$$\textbf{Radial basis function} \qquad U(r) = \begin{cases} r^2 \log(r) & \text{for } d = 2 \\ r^{(d-2)} & \text{for } d \geq 3 \text{ and even (number)} \end{cases} \tag{4.11}$$

$$\textbf{Polynomial basis functions} \qquad g_{ik}(\boldsymbol{x}) = x_1^{a_{i1}} x_2^{a_{i2}} \cdots x_d^{a_{id}} \qquad \text{subject to } \sum_{i=1}^{d} a_{ij} \leq k, \, a_{ij} \geq 0$$

$$\tag{4.12}$$

where

- $\text{logit}(p(PE_{s,t}\,is\,extreme))$ are the logarithmized odds (log-odds), that a precipitation event is extreme. These log-odds are calculated as $\text{logit}(p(PE_{s,t}\,is\,extreme)) = \ln\left(\frac{p(PE_{s,t})}{1-PE_{s,t}}\right)$. Another way to describe these log-odds, is that they are the result of applying the logit-link function to the probabilities $p(PE_{s,t}) \,\forall\,(s,t)\,\in\,ST$, where the logit-link function calculates the odds of $PE_{s,t}$ being extreme and then obtains the natural logarithm of that odds. The log-odds are obtained in the first place, to project to probability values onto a continuous scale.

- β_0 is the model intercept, which represents the log-odds for the case that all predictor variables are set to zero.

- $\boldsymbol{x}$ is a vector of the values of the five predictors at spatio-temporal location $(s,t) \in ST$

- $x_{i,(s,t)}$ is the observation of the i-th predictor at spatio-temporal location (s,t)

- $r = \|\boldsymbol{x}-\boldsymbol{x}_{i,(s,t)}\|$. r in other words is the euclidean distance between the vector of observations and an element of that specific vector for the same spatio-temporal location (s,t)

- the **Linear Predictor** is a linear combination of functions of the five predictor variables altitude, latitude, longitude, mean temperature and the julian date. The functions or smooth terms respectively are thin plate regression splines.

- β_i is coefficient of the i-th **thin plate regression spline**. This coefficient is not easily interpretable, when for instance compared to linear regression or logistic regression. What these coefficients do is to determine the influence of an individual smooth term on the overall linear combination / the log-odds. In that respect, they serve as weights.

- m is the number of polynomial basis functions that construct the linear combination of poylnomial terms in each **thin plate regression spline**

- u_j is a random intercept, which corresponds to one of four factors of the grouping variable season

- $b_j(\boldsymbol{x})$ is a basis function that is associated with the random intercept u_j

- v_k is a random slope; one random slope for each grouping factor (one of the four seasons) and the chosen continuous predictor. In the case of this model, the seasons are assumed to effect only the dependency of the log-odds on the predictors altitude and julian date.

- $c_k(\boldsymbol{x})$ is the basis function associated with the random slope term.

- **thin plate regression spline** refers to a smooth term / function. A thin plate regression spline is a complex basis function, which itself is the sum of a linear combination of radial basis functions $U(r)$ and of a polynomial basis function g_{ik}.

- **radial basis function** is a type of basis function. When combined with other radial basis functions to a linear predictor, they make part of the complex basis function referred to as **thin plate regression spline** above. The radial basis function $U(r)$ measures the distance between the vector of predictor variables $\mathbf{x}$ and the single datapoint $x_{i,(s,t)}$ by transforming the euclidean distance $\|\boldsymbol{x} - \boldsymbol{x}_{i,(s,t)}\|$ in way, that accounts for a determined dimensionality of the input data.

- **polynomial basis function** is a simple polynomial term. In the **thin plate regression spline**, the linear combination of polynomial basis functions / polynomial terms is added to improve the flexibility of the model.

The model response above are log-odds. By applying the logit-link function, which projects the log-odds onto probabilities, these probabilities can be obtained as

$$\widehat{p}_{s,t}^{-1} = \left(1 + \exp(-(\beta_0 + \sum_{i=1}^{5} \beta_i f_i(\boldsymbol{x}) + \sum_{j=1}^{4} u_j b_j(\boldsymbol{x}) + \sum_{k=1}^{M} v_k c_k(\boldsymbol{x}))\right) \tag{4.13}$$

where

$\widehat{p}_{s,t}$ is the predicted probability, that the precipitation event at spatio-temporal location (s,t) is extreme

$\left(1+\exp(-(\beta_0+\sum_{i=1}^{5}\beta_i f_i(\boldsymbol{x})+\sum_{j=1}^{4}u_j b_j(\boldsymbol{x})+\sum_{k=1}^{M}v_k c_k(\boldsymbol{x}))\right)$ is the reciprocal of the sigmoid function (the inverse of the logit-link function) applied to the log-odds (*logit*).

To convert this GAMM into a classifier model, an threshold layer is added, which classifies events as extreme, if their associated predicted probability of being extreme exceeds the threshold of 0.01 (1%%): The classifier model based on the above specified GAMM can finally be expressed as:

$$\widehat{EPE}_{s,t} = \begin{cases} 1, & \text{if } \widehat{p}_{s,t} > 0.01 \\ 0, & \text{if } 0.01 > \widehat{p}_{s,t} \end{cases} \tag{4.14}$$

where $\widehat{EPE}$ is the predicted class label for event $PE_{s,t}$, and 0.01 is the chosen threshold value.

4.4.3.2 How This GAMM Accounts for Non-Linearity and Autocorrelation

The motivation for structuring the generalized additive mixed model as above has already been touched upon in a previous subsection. This subsection is dedicated to extend the depth of these deliberations, after which the results of the estimation of the GAMM are evaluated. This then concludes the second of three steps towards answering the research question (confer 2.1 for the research question).

Modelling the spatio-temporal prediction of probabilities that $p(PE_{s,t})$ is extreme is appropriate for the definition of a logic of the classification of precipitation on the basis of a concept of a generalized additive mixed model as delineated above, for several reasons:

The GAMM above provides flexibility: Through the utilization of thin plate regression splines the GAMM achieves great flexibility in modelling the non-linear relationships between the log-odds of the probability of a precipitation event being extreme as the model response and the predictor variables (altitude etc.). This flexibility also allows the GAMM to account for spatio-temporal clusters / spatio-temporal autocorrelation while being trained.

The GAMM above accounts for autocorrelation:

– As already mentioned, estimating the weights and smoothing parameter of thin plate spline regression terms enables accounting for spatio-temporal autocorrelation. This is the case, as the slope can adjust freely to differences between spatio-temporal clusters through the concept of modelling thin plate regression splines. As a more immediate

means of accounting for the spatio-temporal autocorrelation however, random intercepts and random slopes have been implemented on the basis of the categorical grouping variable seasons. By including random intercepts, that model can account for variability in the probability of a precipitation event being extreme that is not explained by the fixed predictor variables. In contrast, the implementation of random slopes is motivated by the belief, that the way that mean temperature and altitude effect the log-odds of a precipitation event being extreme is influenced by the current metereologic season.

– As the result of incorporating smooth terms for spatial and temporal variables, the GAMM is able to capture complex spatial patterns and temporal trends in the probability of extreme precipitation events. This is particularly important when dealing with climatic data such as precipitation data, as precipitation data exhibits strong seasonality and spatial variability, as has been demonstrated above (cf. figure 3).
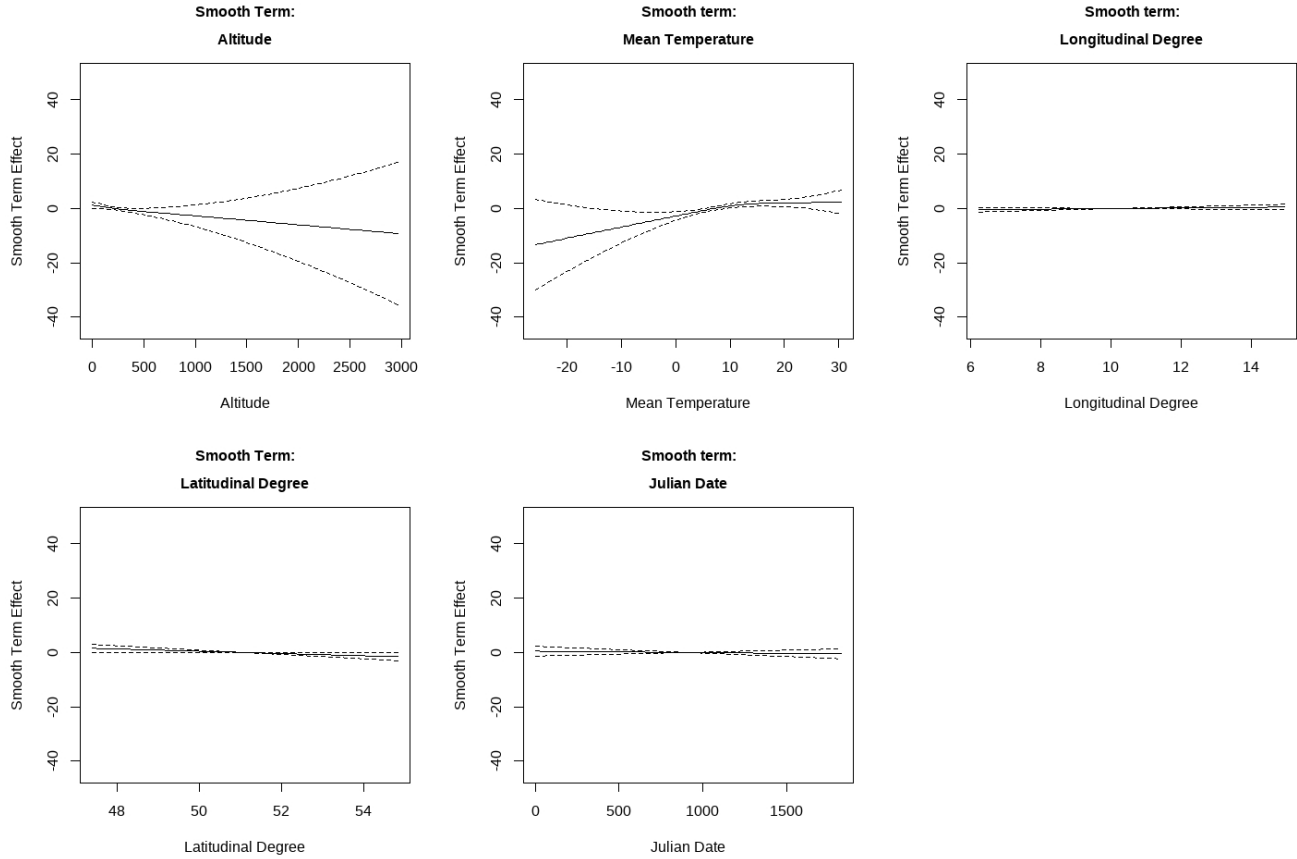
Figure 5: Smooth Terms of the predictors

4.4.3.3 GAMM: Estimation Results

The GAMM as outlined above was estimated through the R package `mgcv` and the corresponding R script named `Training-a-Generalized-Additive-Mixed-Model.R` can be found in the `Regression` subdirectory of the `Scripts` directory. The estimation was challenging in a practical sense, as training the GAMM for more than 1,744,299 observations proved computationally impossible for the current computational set-up of the author, particularly with reference to the working memory, which is limited to 8GB. The author gradually reduced the sample size until it eventually only included observations between the years 1996 and 2000, temporally reducing the sample size to 529,600 observations.

With regard to the interpretation of the estimation results, the smooth terms or the estimated thin plate regression splines are of primary interest. The interpretation of thin plate regression splines however differs from the linear regression and logistic regression model, as the estimated

36

coefficients are not of primary interest. These coefficients represent the weight, that each regression spline has within the corresponding linear combination of regression splines. They are however in contrast to linear and logistic regression coefficients not directly interpretable. Data scientists are usually interested in analyzing the plotted smooth term functions. These visually demonstrate the non-linear dependency between the response (log-odds) and the respective predictor, that the GAMM has identified while being trained.

Figure 5 displays the graphs of smooth term functions for the five predictor variables. A solid line illustrates the non-linear association between the log-odds and the predictor variable picked up by the GAMM while beign trained. Two interrupted lines encircling it indicate the 95 percent confidence interval. The graphs demonstrate that the non-linear relationship is most evident for altitude and mean temperature. For longitudinal degree and Julian date, there appears to be no clear relationship. Examining the smooth term function related to the latitudinal degree, a definitive interpretation is not easily discernible. However, there seems to be a decline in log-odds as the latitudinal degree increases, which ultimately results in a reduced probability of extreme precipitation events as the latitudinal degree rises. This decline is less noticeable compared to the decrease in log-odds with increasing altitude. A Hovmöllerplot, that has been created in the process of exploratory data analysis, seemingly supports the assertion, that precipitation height and hence the probability exteme precipitation increases, as the latitudinal decreases (cf. **??**). Upon observing the plot of the smooth term linked to altitude, simultaneously a significant widening of the 95 percent confidence interval becomes apparent. This likely reflects a characteristic of the sample data, where the majority of weather stations are situated at lower altitudes, and the number of weather stations in the training data typically diminishes as the altitude level rises (cf. figure 8). A comparable observation can be made by examining the smooth term for mean temperature. While it appears, that the log-odds increase, as the mean temperature increases, the confidence interval substantially narrows as well. The reason for this observation of a narrowing confidence interval is likely related to the structure of the sample data as well, where simply less observations are made at lower temperatures. The relationship relationship between log-odds and mean-temperature however appears non-linear and a decreasing marginal effect seems to be prevalent: With an increase in mean temperature, the additional log-odds appear to decline. In this scenario, there seems to be a notable difference between mean

temperatures below or above approximately 10 degrees.

Overall, given the analysis and interpretation of the smooth terms, the prior assumptions on the patterns of precipitation appear partially supported. As theorized above, there seems to be pronounced dependence of the log-odds / the probabilities of extreme precipitation on mean temperature and altitude, while the respective relationship appears to be linear for altitude against prior expectations. For the preidctors julian date and longitudinal degree the smooth term functions appear perfectly flat, signifying, that there is no significant relationship between the model response and these predictors. Finally, with reference to the interpretation of the smooth term function associated with the latitudinal degree, a relationship can not be made out unambiguously.

This concludes the second step towards an answer to the research question, which revolved around IDW based interpolation and a trained GAMM being appropriate choices for the definition of a classification logic. The following sections provide the thrid and last step, which is a comparison of the performance of the two trained classifier models in predicting extreme precipitation between 1996 and 2016.

5   A Competetive Evaluation of the Performance of the Classifier Models

Up to this point, the prediction of extreme precipitation was framed as the classification of precipitation events as either extreme or non-extreme. It has been established, that patterns of precipitation are governed by spatio-temporal autocorrelation and the dependency of precipitation height on altitude, mean temperature and the spatio-temporal location. And IDW based interpolation and a trained GAMM have finally been asserted as appropriate choices when attempting to define the logic behind the classification of precipitation events, precisely because they account for these assumed patterns.

All of these considerations eventually prepare what is to come now, that is, the essence of the answer to the research question, *how the classifier models compare performance-wise.* (cf. 2.1 for the research question). It revolves around measurements of confusion-matrix based performance

indicators. The following three subsections are structured as follows:

Firstly, the concept of confusion-matrix based performance indicators is explained in the context of the above mentioned Standard of Truth. Then, the measured values for the performance-indicators are presented for both classifier models individually. Finally, a comparison of the measurements is conducted and interpreted.

5.1 How Confusion-Matrix Based Performance Indicators Quantify Classifier Performance

In an earlier section, the relevance of the Standard of Truth (S.o.T.) for evaluation purposes was established. In this context, determining a set of true classifications was proposed and the procedure was described (see equation 3.5). These true classifications were deemed a necessary prerequisite for practically assessing the validity of a generated classification by comparing it to the true classification. According to a general rule for classifier model evaluation, a classifier is considered better if it produces more true classifications (Hosmer et al. (2013), pp. 169)(Zumel and Mount (2014)). A more nuanced understanding of the classifier model's performance can be achieved by counting the number of true and false classifications while distinguishing between classifications of $PE_{s_i,t}$ as extreme or non-extreme for all $(s_i, t) \in ST$. For improved readability and to align with common practices in the data science literature, a classification of $PE_{s_i,t}$ as extreme is referred to as a *positive*, while a classification of $PE_{s_i,t}$ as non-extreme is referred to as a *negative.* By counting the number of true negatives (TN), false positives (FP), false negatives (FN), and true positives (TP) and arranging them in a 2x2 matrix, the standard 2x2 confusion matrix is obtained, such that

$$
\textbf{Confusion Matrix} := \begin{array}{c|c|c|} & \text{Predicted Negative} & \text{Predicted Positive} \\ \hline \text{Negative by S.o.T.} & \text{True Negatives} & \text{False Positives} \\ \hline \text{Positive by S.o.T.} & \text{False Negatives} & \text{True Positives} \\ \end{array} \quad (5.1)
$$

Now, let Q be the complete set of true classifications of $PE_{s_i,t} \; \forall \; (s_i, t) \in ST$ and let P be the complete set of generated classifications of $PE_{s_i,t} \; \forall \; (s_i, t) \in ST$. Also let q be a subset of Q and p be a subset of P.

Then:

[23]True Positives :=

$|\{p \times q \mid q \in Q,\, p \in P,\, q = p = \text{positive}\}|$

False Positives :=

$|\{p \times q \mid q \in Q,\, p \in P,\, q = \text{negative}, p = \text{positive}\}|$

$$(5.2)$$

False Negatives :=

$|\{p \times q \mid q \in Q,\, p \in P,\, q = \text{positive}, p = \text{negative}\}|$

True Negatives :=

$|\{p \times q \mid a \in A,\, p \in P,\, q = p = \text{negative}\}|$

$$(5.3)$$

As mentioned earlier, these count values or values in the cells of the confusion matrix respectively, are used to acquire a more differentiated view of the performance of the classifier model. Specifically they are the basis for the calculation of a set of performance indicators, called *Precision, Sensitivity, Specificity, False Positive Rate* and *False Negative Rate*.[24] Their respective formulae read as

Classifier Performance Indicators

$$\text{Precision} := \frac{TP}{TP + FP} \qquad \text{Sensitivity} := \frac{TP}{TP + FN} \qquad \text{Specificity} := \frac{TN}{TN + FP}$$

$$\text{False Positive Rate} := \frac{FP}{TN + FP} \qquad \text{False Negative Rate} := \frac{FN}{TP + FN}$$

$$(5.4)$$

Each of these measure a particular aspect of the classifier performance:

---

[24]Another indicator called *Accuracy* is not considered. The problem with Accuracy, which describes the ratio of predicted true classifications to the number of predicted classifications, is a strong bias in the case of asymmetrical distributions, as is the case with Precipitation events, where most of the precipitation events are non-extreme.

The **Precision** of a classifier is the fraction of the generated positives, for which the generated classification as positive is correct. This fraction gives an answer to the question *how often a generated classification of $PE_{s_i,t}$ as extreme is true.*

The **Sensitivity** of a classifier is the fraction of the true generated classifications of $PE_{s_i,t}$ as extreme to the overall number of true positives. It gives an answer to *how many times the classifier correctly identified extreme precipitation events.*

The **Specificity** of a classifier is the fraction of the true classifications of precipitation events as non-extreme to the number of true classifications of PE as non-extreme in the training data. It gives an answer to *how many times the classifier correctly identified non-extreme precipitation (or precipitation events as non-extreme).*

The **False Positive Rate** is the performance indicator complementary to the True Negative Rate, as it describes the fraction of true classifications of precipitation events as non-extreme in the training data, which the model didn't generate.

The **False Negative Rate** is the indicator of model performance, which is complementary to the true Positive Rate, as it describes the fraction of true classifications of precipitation events as extreme in the training data, which the model didn't generate.

This concludes the theory of the quantification of classifier model performance based on the standard 2×2 confusion matrix and associated performance indicators. The following two sections display the measured the confusion matrix and performance indicators for the two distinct classifiers. These two sections will then be followed by a comparison of the assessed performance and a conclusion of this thesis paper with reference to the research question.

5.2   Confusion-Matrix and Performance Indicators for the IDW Based Classifier

The 2×2 standard confusion matrix for the classifier, that utilizes the predictions of precipitation height generated by Inverse Distance Weighting based interpolation for defining the classification logic, presents itself[25]} as

	Predicted Negative	Predicted Positive
Negative by S.o.T.	1,744,197	0
Positive by S.o.T.	0	102

Figure 6: 2×2 standard confusion matrix for IDW based classifier. Own representation

The performance indicators *Precision, Sensitivity, Specificity, False Positive Rate and False Negative Rate* were measured as

Measurements for the classifier performance indicators

$$\text{Precision} = 1 \qquad \text{Sensitivity} = 1 \qquad \text{Specificity} = 1$$

$$\text{False Positive Rate} = 0 \qquad \text{False Negative Rate} = 0$$

$$(5.5)$$

The examination of the confusion matrix and the associated measurements of the performance indicators reveals, that this binary classifier generated correct classifications only. It correctly identified all precipitation events that are extreme based on the S.o.T. as extreme and correctly identified all precipitation events that are non-extreme based on the S.o.T. as non-extreme. Within the confusion matrix, this is reflected in the absence of any false positives and false negatives. For the given task of generating binary classifications of the precipitation events in the training data, the IDW based classifier thus achieved perfect performance. This perfect performance is also

---

[25]The R script for the calculation of the confusion matrix and the associated performance indicators is named `Inverse-Distance-Weighting-Calculation` and can be found at .\BachelorThesis\Scripts\Data-Manipulation\Training-Data\Inverse-Distance-Weighting-Calculation.R"

reflected in the measurements of the performance indicators, that are associated with the confusion matrix cells:

- Precision was measured as 1 (100%). This measured value translates into the statement "100% *of the classifications of precipitation events as extreme are true*".

- Sensitivity was measured as 1 (100%). This measured value translates into the statement "100% *of true classifications of precipitation events as extreme were generated by the classifier model*". In other words: The classifier correctly identified 100% of the true extreme precipitation events as such.

- Specificity was measured as 1 (100%). This measured value translates into the statement "100% *of true classifications of precipitation events as non-extreme were generated by the classifier model*". In other words: The classifier correctly identified 100% of the true non-extreme precipitation events as such.

- The False Positive Rate was measured as 0 (0%). This measured value translates into the statement "0% *of the true negatives have not been identified by the classifier*".

- The False Negative Rate was measured as 0 (0%). This measured value translates into the statement "0% *of the true positives have not been identified by the classifier*".

5.3 Confusion-Matrix and Performance Indicators for the GAMM Based Classifier

The 2×2 standard confusion matrix for the classifier that utilizes the predictions of probabilities of a precipitation event being extreme generated by a trained GAMM for defining the classification logic, presents itself[26]} as

	Predicted_Negative	Predicted_Positive
Negative by S.o.T.	529,577	0
Positive by S.o.T.	23	0

Figure 7: 2×2 standard confusion matrix for classifier based on a trained GAMM. Own representation

The performance indicators *Precision, Sensitivity, Specificity, False Positive Rate and False Negative Rate* were measured as

Measurements for the classifier performance indicators

$$\text{Precision} = NaN \qquad \text{Sensitivity} = 0 \qquad \text{Specificity} = 1$$

$$\text{False Positive Rate} = 0 \qquad \text{False Negative Rate} = 1$$

$$(5.6)$$

When examining the GAMM based classifier model, the first thing to notice is that no precipitation event was classified as being extreme at all. For the subset of the training data which was used to train the GAMM due to computational limitations, 23 true classifications of precipitation events as extreme precipitation events could have been made by the GAMM based classifier model, and the GAMM apparently missed all of them. Since it classified all precipitation events as non-extreme, it follows naturally however, that all possible true classifications of precipitation events as non-extreme were made by the the classifier. It is likely, that this specific feature of the classifier

---

[26]The R script for the calculation of the confusion matrix and the associated performance indicators is named `GAMM-Confusion-Matrix.R` and can be found at `.\BachelorThesis\Scripts\Data-Manipulation\Training-Data\GAMM-Confusion-Matrix.R"`

performance is associated with the choice of 0.01 as the probability threshold which a predicted probability had to succeed in order for the corresponding precipitation event to be classified as extreme. This specific feature of the model performance is dinstinctly reflected in the measured performance indicator values:

- Precision was measured as NaN (NA). In the case of this particular classifier, Precision can not be measured, because the classifier never classified a precipitation event to be extreme.
- Sensitivity was measured as 0 (0%). This measured value translates into the statement "0% *of true classifications of precipitation events as extreme were generated by the classifier model*". In other words: The classifier correctly identified 0% of the true extreme precipitation events as such.
- Specificity was measured as 1 (100%). This measured value translates into the statement "100% *of true classifications of precipitation events as non-extreme were generated by the classifier model*". In other words: The classifier correctly identified 100% of the true non-extreme precipitation events as such.
- The False Positive Rate was measured as 0 (0%). This measured value translates into the statement "0% *of the true negatives have not been identified by the classifier*".
- The False Negative Rate was measured as 1 (100%). This measured value translates into the statement "100% *of the true positives have not been identified by the classifier*".

5.4   A Comparison of the Performance Indicators of Both Classifiers

This chapter marks the end of the third and final step of the three steps in this thesis towards an answer to the research question (for the research question confer 2.1). The first step had been *framing the prediction of extreme precipitation as a binary classification problem.* In that context, precipitation and extreme precipitation events had been defined, a Standard of (the) Truth of generated classifications had been established and the general concept of binary classifier models had been delineated. The second step had been demonstrating spatio-temporal autocorrelation on the basis of spatio-temporal covariances and arguing the case of IDW based interpolation and a trained GAMM as appropriate choices for the definition of the logic of the binary classification of precipitation as extreme and non-extreme. The third and final step so far had been an explanation of the general structure of a confusion-matrix and a description of performance indicators that are

calculated on the basis of the confusion-matrices's cells.

This chapter concludes the third step and thus the process of answering the research question, by providing an explicit comparison of the performance of the two classifier models with respect to the prediction of extreme precipitation between 1996 and 2016.

Against this background, it is most noticeable when comparing the two classifier models, that the IDW based classifier distinctly outperforms the GAMM based classifier. While the IDW based classifier achieves a perfect performance for the training data, as reflected in the measured perfect Precision, Sensitivity and Specificity, the GAMM based classifier has failed in all 23 instances to correctly predict precipitation events as being extreme. It needs to be noted, that the GAMM was trained on a subset of the training data only, due to computational restraints. The IDW based classifier without doubt still overperformed the GAMM based classifier; there are however three reasons, why this particular difference of performance should not neccessarily be interpreted as a sign for the superiority of the IDW based classifier. Firstly, IDW based interpolation is a scoring model that generates predictions of precipitation height. And observed precipitation height in turn is involved in the Standard of Truth of classifications of precipitation events. This is relevant, since for the given task of the classifier to generate a binary classification decision, the only thing that matters in terms of the correctness of a classification is whether both the actually observed precipitation height and the predicted precipitation height exceed the classification threshold or both deceed it. And there is only one threshold for binary classifications. This implies, that a generated classification decision would be true even, if the predicted precipitation *just* exceeds the classification threshold, while the actual observed precipitation height exceeds it *by a much greater margin.* It could also be the other way around; what is important, is that a classification logic that is based on predicted scores fundamentally differs from a classification logic that is based on predicted probabilities. Secondly, the classifier was constructed based on the trained GAMM by utilizing a probability value of 0.01 (or 1%) as the chosen probability threshold. It is very likely the case, that the GAMMs performance varies depending on the chosen probability. Thirdly and finally, it should also be noted, that a planned cross-validation of the two classifier models on the basis of the prepared hold-out data could not be conducted for the reason of time restraints. It can thus not be ruled out, that the two classifier models could have compared differently on the basis of the hold-out (outer-sample data).

6 Conclusion

Based on these results of the performance-wise comparison of the IDW based classifier and the GAMM based classifier, this thesis paper concludes with the following answer to the research question:

Answer to the Research Question

The IDW based classifier distinctly outperforms the GAMM based classifier for the given training data. This observation however should not be generalized into an assertion of the superiority of the IDW based classifier for these three reasons:

1. While IDW based interpolation is a scoring method that generates estimations of precipitation height, the trained GAMM predicts probabilities of precipitation events being extreme. The two distinct classifier models therefore differ fundamentally with reference to the nature of their classification logic. This in turn complicates a comparison of their performance.

2. For the GAMM based classifier, the probability threshold was set as 1%. Different threshold values however likely result in different confusion tables and hence in different measured performance.

3. Only performance based on the training data was compared. Outer-sample data or cross-validation respectively could not be utilized due to time restraints. Both classifiers could have performed differently for outer-sample data.

7 Appendix

The appendix serves as a succinct reference to the mathematical structure of both models, that were described in the context of defining the classification logic and also of measurements of autocorrelation. Other than that, it contains graphics that were referenced in thesis's main part.

7.1 Defining the Classification Logic

In this subsection of the appendix, models are presented mathematically that were previously referenced in the context of defining the classification logic.

7.1.1 Classical Linear Regression

$$
\begin{aligned}
\textbf{Precipitation Height} &= \beta_0 + \beta_1 \textbf{Mean Temperature} + \beta_2 \textbf{Altitude} \\
&+ \beta_3 \textbf{Longitude} + \beta_4 \textbf{Latitude} + \beta_5 \textbf{Julian Date} + \epsilon
\end{aligned}
\tag{7.1}
$$

with assumptions

Linearity:

$$y_{s,t} = \beta_0 + \sum_{j=1}^{p} \beta_j x_{j,s,t} + \epsilon_{s,t}$$

Independence:

$$E(\epsilon_{s,t}) = 0, \quad Var(\epsilon_{s,t}) = \sigma^2, \quad Cov(\epsilon_{s_i,t_i}, \epsilon_{(s_j,t_j)}) = 0 \quad \forall \, (s_i, t_i) \neq (s_j, t_j)$$

Homoscedasticity:

$$Var(\epsilon_{s,t}) = \sigma^2 \; \forall \, s, t \in \, ST$$

Normality:

$$\epsilon_{s,t} \sim N(0, \sigma^2)$$

No perfect multicollinearity:

$$\mathrm{rank}(X) = p$$

No autocorrelation:

$$Cov(\epsilon_{s_i,t_i}, \epsilon_{(s_j,t_j)}) = 0 \quad \forall \quad \begin{pmatrix} s_i \\ t_i \end{pmatrix} \neq \begin{pmatrix} s_j \\ t_j \end{pmatrix}$$

Declaration of Symbols and Equation Components

Symbols related to equation 7.1

Precipitation Height: Response

β_0: Intercept parameter.

β_1: Coefficient for $MeanTemperature$.

β_2: Coefficient for $Altitude$.

β_3: Coefficient for $Longitude$.

β_4: Coefficient for $Latitude$.

β_5: Coefficient for $JulianDate$.

ϵ: Error term.

Symbols related to the assumptions

Linearity Assumption:

$y_{s,t}$: Response.

β_j: Coefficient for the j-th predictor variable.

$x_{j,s,t}$: Value of j-th predictor variable at space-time location s, t.

Independence Assumption:

$E(\epsilon_{s,t})$: Expected value of error term for precipitation height observed at space-time location.

$Var(\epsilon_{s,t})$: Variance of error term at space-time location.

$Cov(\epsilon_{s_i,t_i}, \epsilon_{(s_j,t_j)})$: Covariance of error terms between two space-time locations.

Homoscedasticity Assumption:

σ^2: Constant variance of error term.

$s, t \in: ST$: Space-time locations as elements of the space-time domain derived from the combination of the two-dimensional spatial domain (longitude,latitude) and the temporal domain made up of julian dates.

Assumption of normality:

$\epsilon_{s,t}$: Error term.

$N(0, \sigma^2)$: Normal distribution with mean 0 and variance σ^2.

Assumption of no perfect multicollinearity :

rank(X): Rank of the design matrix X.

p: Number of predictor variables.

Assumption of absend autocorrelation:

$Cov(\epsilon_{s_i,t_i}, \epsilon_{s_j,t_j})$: Covariance of error terms.

7.1.2 Logistic Regression

$$p(PE_{s,t} \text{ in class EPE}) = \mathbf{S}\left(\log\left(\frac{p_{s,t}}{1-p_{s,t}}\right)\right)$$

$$= \mathbf{S}\begin{pmatrix} \beta_0 + \beta_1\text{Mean Temperature}_{s,t} + \beta_2\text{Altitude}_{s,t} \\ + \beta_3\text{Longitude}_{s,t} + \beta_4\text{Latitude}_{s,t} \\ + \beta_5\text{Julian Date}_{s,t} \end{pmatrix} \quad (7.2)$$

---

where

Coefficients

$$\beta_0, \beta_1, \beta_2, \beta_3, \text{and } \beta_4 \tag{7.3}$$

are *estimated* per *Maximum Likelihood Estimation.*

Predicted Probability

$$p_{s,t} = \left(1 + e^{-(\beta_0 + \beta_1 MeanTemperature_{s,t} + \beta_2 Altitude_{s,t} + \beta_3 Longitude_{s,t})}\right.$$
$$\left. \times e^{-(\beta_4 Latitude_{s,t} + \beta_5 JulianDate_{s,t})}\right)^{-1} \tag{7.4}$$

is the model prediction of the probability that the PE at spatio-temporal location s, t is an EPE.

Conditional Probability

$$P(EPE_{s,t}|\mathbf{x}_{s,t}, \beta) = p_{s,t}^{EPE_{s,t}}(1 - p_{s,t})^{1-EPE_{s,t}} \tag{7.5}$$

is the conditional probability of observing an EPE at the given location and time, given the predicted probability $p_{s,t}$. The term $p_{s,t}^{EPE_{s,t}}$ represents the probability of observing that an extreme precipitation event actually occurs at the spatio-temporal location, and $(1 - p_{s,t})^{1-EPE_{s,t}}$ represents the probability of not observing an EPE at the spatio-temporal location.

Likelihood

$$L(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5) = \prod_{s,t \in SP} P(EPE_{s,t}|\mathbf{x}_{s,t}, \beta) \tag{7.6}$$

$$= \prod_{s,t \in SP} p_{s,t}^{EPE_{s,t}}(1 - p_{s,t})^{1-EPE_{s,t}} \tag{7.7}$$

is the likelihood function or the likelihood of the entire training data, respectively.

7.1.3 Generalized Additive Mixed Model

Generalized Additive Mixed Model - Model Equation

$$
\begin{aligned}
\text{logit}(p(PE_{s,t}\,is\,extreme)) = {}& \beta_0 + f_{alt}(\text{Altitude}_{s,t}) + f_{lon}(\text{Longitude}_{s,t}) + f_{lat}(\text{Latitude}_{s,t}) \\
& + f_{jul}(\text{Julian Date}_{s,t}) + f_{temp}(\text{Mean Temperature}_{s,t}) \\
& + u(\text{Season}_{s,t}) + f_{random\ slope\ alt}(\text{Altitude}_{s,t}|\text{Season}_{s,t}) \\
& + f_{random\ slope\ jul}(\text{Julian Date}_{s,t}|\text{Season}_{s,t}) + \epsilon \\
= {}& \beta_0 + \sum_{i=1}^{5}\beta_i f_i(\boldsymbol{x}) + \sum_{j=1}^{4} u_j b_j(\boldsymbol{x}) + \sum_{k=1}^{M} v_k c_k(\boldsymbol{x})
\end{aligned}
\tag{7.8}
$$

This model equation can be broken down in the form of

Linear Predictor
$$
\beta_0 + \sum_{i=1}^{5}\beta_i f_i(\boldsymbol{x}) + \sum_{j=1}^{4} u_j b_j(\boldsymbol{x}) + \sum_{k=1}^{M} v_k c_k(\boldsymbol{x})
\tag{7.9}
$$

Thin Plate Regression Spline
$$
f_i(\boldsymbol{x}) = \sum_{(s,t)\in ST} w_{i,(s,t)} U(\|\boldsymbol{x} - \boldsymbol{x}_{i,(s,t)}\|) + \sum_{k=1}^{m} \alpha_{ik} g_{ik}(\boldsymbol{x})
\tag{7.10}
$$

Radial basis function
$$
U(r) = \begin{cases} r^2 \log(r) & \text{for } d = 2 \\[2mm] r^{(d-2)} & \text{for } d \geq 3 \text{ and even (number)} \end{cases}
\tag{7.11}
$$

Polynomial basis functions
$$
g_{ik}(\boldsymbol{x}) = x_1^{a_{i1}} x_2^{a_{i2}} \cdots x_d^{a_{id}} \qquad \text{subject to } \sum_{i=1}^{d} a_{ij} \leq k,\ a_{ij} \geq 0
\tag{7.12}
$$

where

$\text{logit}(p(PE_{s,t}\,is\,extreme))$ are the logarithmized odds (log-odds), that a precipitation event is extreme. These log-odds are calculated as $\text{logit}(p(PE_{s,t}\,is\,extreme)) = \ln\left(\frac{p(PE_{s,t})}{1-PE_{s,t}}\right)$. Another way to describe these log-odds, is that they are the result of applying the logit-link function to the probabilities $p(PE_{s,t}) \,\forall\, (s,t) \in ST$, where the logit-link function calculates the odds

of $PE_{s,t}$ being extreme and then obtains the natural logarithm of that odds. The log-odds are obtained in the first place, to project to probability values onto a continuous scale.

β_0 is the model intercept, which represents the log-odds for the case that all predictor variables are set to zero.

$\boldsymbol{x}$ is a vector of the values of the five predictors at spatio-temporal location $(s,t) \in ST$

$x_{i,(s,t)}$ is the observation of the i-th predictor at spatio-temporal location (s,t)

$r = \|\boldsymbol{x} - \boldsymbol{x}_{i,(s,t)}\|$. r in other words is the euclidean distance between the vector of observations and an element of that specific vector for the same spatio-temporal location (s,t)

the **Linear Predictor** is a linear combination of functions of the five predictor variables altitude, latitude, longitude, mean temperature and the julian date. The functions or smooth terms respectively are thin plate regression splines.

β_i is coefficient of the i-th **thin plate regression spline**. This coefficient is not easily interpretable, when for instance compared to linear regression or logistic regression. What these coefficients do is to determine the influence of an individual smooth term on the overall linear combination / the log-odds. In that respect, they serve as weights.

m is the number of polynomial basis functions that construct the linear combination of poylnomial terms in each **thin plate regression spline**

u_j is a random intercept, which corresponds to one of four factors of the grouping variable season

$b_j(\boldsymbol{x})$ is a basis function that is associated with the random intercept u_j

v_k is a random slope; one random slope for each grouping factor (one of the four seasons) and the chosen continuous predictor. In the case of this model, the seasons are assumed to effect only the dependency of the log-odds on the predictors altitude and julian date.

$c_k(\boldsymbol{x})$ is the basis function associated with the random slope term.

thin plate regression spline refers to a smooth term / function. A thin plate regression spline is a complex basis function, which itself is the sum of a linear combination of radial basis functions $U(r)$ and of a polynomial basis function g_{ik}.

radial basis function is a type of basis function. When combined with other radial basis functions to a linear predictor, they make part of the complex basis function referred to as **thin plate regression spline** above. The radial basis function $U(r)$ measures the distance between the vector of predictor variables $\mathbf{x}$ and the single datapoint $x_{i,(s,t)}$ by transforming the euclidean distance $\|\boldsymbol{x} - \boldsymbol{x}_{i,(s,t)}\|$ in way, that accounts for a determined dimensionality of the input data.

polynomial basis function is a simple polynomial term. In the **thin plate regression spline**, the linear combination of polynomial basis functions / polynomial terms is added to improve the flexibility of the model.

Predicted Probability

$$\widehat{p}_{s,t}^{-1} = \left(1 + \exp(-(\beta_0 + \sum_{i=1}^{5} \beta_i f_i(\boldsymbol{x}) + \sum_{j=1}^{4} u_j b_j(\boldsymbol{x}) + \sum_{k=1}^{M} v_k c_k(\boldsymbol{x}))\right) \tag{7.13}$$

where

$\widehat{p}_{s,t}$ is the predicted probability, that the precipitation event at spatio-temporal location (s,t) is extreme

$\left(1 + \exp(-(\beta_0 + \sum_{i=1}^{5} \beta_i f_i(\boldsymbol{x}) + \sum_{j=1}^{4} u_j b_j(\boldsymbol{x}) + \sum_{k=1}^{M} v_k c_k(\boldsymbol{x}))\right)$ is the reciprocal of the sigmoid function (the inverse of the logit-link function) applied to the log-odds (*logit*).

Classifier based on the GAMM (confer equation 7.8)

$$\widehat{EPE}_{s,t} = \begin{cases} 1, & \text{if } \widehat{p}_{s,t} > 0.01 \\ 0, & \text{if} 0.01 > \widehat{p}_{s,t} \end{cases} \tag{7.14}$$

where $\widehat{EPE}$ is the predicted class label for event $PE_{s,t}$, and 0.01 is the chosen threshold value.

X

7.1.4 Inverse Distance Weighting Based Interpolation

$$\widehat{rain}_{s_i,t_i} := \sum_{s_j,t_j \in ST} \left(w_{(s_i,t_i),(s_j,t_j)} \cdot rain_{s_j,t_j} \right) \tag{7.15}$$

(cf. Wikle et al. (2019) , pp. 78)(cf. Bivand et al. (2013) , pp. 215)

---

where

$$d^{\alpha}_{(s_i,t_i,(s_j,t_j))} := \left(\sqrt{(lon_{s_i} - lon_{s_j})^2 + (lat_{s_i} - lat_{s_j})^2 + C(t_i - t_j)^2} \right)^{\alpha} \tag{7.16}$$

and

$$w_{(s_i,t_i),(s_j,t_j)} := \frac{d^{-\alpha}_{(s_i,t_i),(s_j,t_j)}}{\sum\limits_{s_j,t_j \in ST} d^{-\alpha}_{(s_i,t_i),(s_j,t_j)}} \tag{7.17}$$

Declaration of Symbols and Equation Components

$\widehat{\mathbf{rain}}_{\mathbf{s,t}}$:

The estimated precipitation height at a specific location and time, denoted by longitude lon_i, latitude lat_r, and Julian day t.

$\sum\limits_{\mathbf{s_j,t_j \in ST}}$:

A summation over all the locations and times in the spatio-temporal domain $\mathbb{R}^2 \times \mathbb{R}$ emerging from the conjunction of the two-dimensional spatial domain, within which longitudinal and latitudinal degree each establish one dimension, and the one-dimensional spatial domain, consisting of julian dates as real numbers.

$$\left(\mathbf{w}_{(s_i,t_i),(s_j,t_j)} \cdot \mathbf{rain}_{(s_j,t_j)} \right):$$

A product of two terms. The first term, $w_{((s_j,t),s,t)}$, is a weight assigned to each location and time in ST based on its distance from s_i, t_i, the location and time of interest. The second term, $rain_{s_j,t_j}$, is the available observed precipitation height at the spatio-temporal location s_j, t_j.

7.2 Measurements of Spatial, Temporal and Spatio-Temporal Autocorrelation

This subsection of the appendix mathematically presents measurements of autocorrelation for reference.

7.2.1 Haversine Formula

$$d := 2r \arcsin\left(\sqrt{\sin^2\left(\frac{\phi_2 - \phi_1}{2}\right) + \cos(\phi_1)\cos(\phi_2)\sin^2\left(\frac{\lambda_2 - \lambda_1}{2}\right)}\right) \tag{7.18}$$

Dauni et al. (2019)

Declaration of Symbols and Equation Components

d:

Distance between two spatial points in kilometer

r:

Radius of the world in kilometer (6371km)

ϕ_1 and ϕ_2:

Latitudinal Degrees of the two points measured in radians. ϕ is calculated from a decima latidunial degree per $\phi = lat\,\frac{\pi}{180}$, where lat is a latitudinal degree.

$\lambda_\mathbf{1}$ and $\lambda_\mathbf{2}$:

Longitudinal Degrees of the two points measured in radians. λ is calculated from a decimal latitudinal degree per $\lambda = lon \, \frac{\pi}{180}$

sin and **cos**:

Trigonometric functions. For a given angle measured in radians, both functions return either the sine or the cosine.

arcsin:

$arcsin(x)$ is the inverse of the sine function. For a given sine x, it returns the sine, whose angle measured (measured in radians) is then x

7.2.2 Empirical Spatio-Temporal Covariance

$$\widehat{C}_{rain}(d_{spatial}, d_{julian}) := \frac{1}{|N(d_{spatial})|} \frac{1}{|N(d_{julian})|} \times \left(\sum_{s_i,s_j \,\in\, N(d_{spatial})} \sum_{t_i,t_j \,\in\, N(d_{julian})} \right.$$
$$\left. (rain_{s_i,t_i} - \widehat{\mu}_{t_i}) \left(rain_{s_j,t_j} - \widehat{\mu}_{t_j} \right) \right) \tag{7.19}$$

Cressie and Wikle (2011) | Wikle et al. (2019) | Auer and Rottmann (2020)

Declaration of Symbols and Equation Components

$\widehat{\mathbf{C}}_{\mathbf{rain}}(\mathbf{d_{spatial}}, \mathbf{d_{julian}})$:

The empirical covariance of the precipitation height data for a given spatial distance in kilometer $d_{spatial}$ and a given temporal distance d_{julian}.

$\mathbf{d_{spatial}}$:

The interval of spatial distances between two weather stations denoted each as a pair of angular coordinates, within which the precipitation heights are considered to be correlated.

$\mathbf{d_{julian}}$:

The interval of temporal distances between t and t, within which the precipitation heights are considered to be correlated

$\mathbf{t_i}$ and $\mathbf{t_j}$:

Julian days or the days elapsed since the earliest observation within the precipitation data, respectively.

N($d_{spatial}$):

A set of pairs of weather stations, for whom the spatial distance within the pair evaluates to a value within the interval $d_{spatial}$.

N(d_{julian}):

A set of pairs of weather stations, for whom the temporal distance within the pair evaluates to a value within the interval d_{julian}.

|N($d_{spatial}$)|:

The number of pairs weather stations within the set $d_{spatial}$.

|N(d_{julian})|:

The number of pairs of temporal locations of the weather stations that are within the set d_{julian}).

$$\sum_{(s_i,s_j)\in N(d_{spatial})} \sum_{(t_i,t_j)\in N(d_{julian})} :$$

A double summation over pairs of observations of the precipitation height made at a pair of weather stations spatio-temporally located, such that the pair belongs to $d_{spatial}$ spatially and d_{julian} temporally.

rain_{s_i,t_i} and rain_{s_j,t_j}:

The precipitation height value at spatio-temporal location s_i, t_i and s_j, t_j

$\widehat{\mu}_{\mathbf{rain,temporal}}(\mathbf{t_i})$ and $\widehat{\mu}_{\mathbf{rain,temporal}}(\mathbf{t_j})$:

The spatial mean of the precipitation height data at time t_i and t_j.

$(\mathbf{rain_{s_i,t_i}} - \widehat{\mu}_{\mathbf{rain,temporal}}(\mathbf{t_i}))$:

The deviation of the precipitation height data from their temporal mean at spatio-temporal location s_i, t_i

$(\mathbf{rain_{s_j,t_j}} - \widehat{\mu}_{\mathbf{rain,temporal}}(\mathbf{t_j}))$:

The deviation of the precipitation height data from their temporal mean at spatio-temporal location s_j, t_j.

$(\mathbf{rain_{s_i,t_i}} - \widehat{\mu}_{\mathbf{rain,temporal}}(\mathbf{t_i}))(\mathbf{rain_{s_j,t_j}} - \widehat{\mu}_{\mathbf{rain,temporal}}(\mathbf{t_j}))$:

The product of the two deviations

$\frac{1}{\mathbf{2|N(d_{spatial})|}} \frac{1}{\mathbf{2|N(d_{julian})|}}$:

A normalization factor that scales the covariance to account for the number of pairs of weather stations, for whom the distance in kilometer evaluates to a value within $d_{spatial}$, while their temporal distance evaluates to value within the interval $d_{temporal}$ being used in the calculation.

7.2.3   Empirical Spatial Covariance

$$\widehat{C}_{rain,\left(d_{julian}\right)}(s_i, s_j) := \frac{1}{T - d_{julian}} \times \left(\sum_{t=1+d_{julian}}^{T} \left(rain_{s_i,t} - \widehat{\mu}_{rain,temporal}(s_i)\right) \right.$$

$$\left. \left(rain_{s_j,t-d_{julian}} - \widehat{\mu}_{rain,temporal}(s_j)\right) \right)$$

$$(7.20)$$

Cressie and Wikle (2011) | Wikle et al. (2019) | Auer and Rottmann (2020)

Declaration of Symbols and Equation Components

$\widehat{\mathbf{C}}_{\mathbf{rain,d_{julian}}}$:

> The estimated spatial covariance of the precipitation data, such that a given temporal distance d_{julian} applies to all pairs of observations, for whom the cross-product is calculated.

$\mathbf{s_i}$ and $\mathbf{s_j}$:

> Two spatial locations of weather stations.

$\mathbf{T}$:

> The number of julian dates or temporal locations, respectively.

$\mathbf{d_{julian}}$:

> A specific temporal distance between two julian dates.

$\mathbf{rain_{s_i,t}}$ and $\mathbf{rain_{s_j,t}}$:

Observations of precipitation height at the spatio-temporal locations s_i, t and s_j, t.

$\widehat{\mu}_{\mathbf{rain,spatial}}(\mathbf{s_i})$ and $\widehat{\mu}_{\mathbf{rain,spatial}}(\mathbf{s_j})$:

The estimated means of the precipitation data at locations $\mathbf{s_i}$ and $\mathbf{s_j}$, respectively.

7.2.4   Empirical Temporal Covariance

$$\widehat{C}_{rain}^{(d_{spatial})}(t,t) := \frac{1}{|N(d_{spatial})|} \times \left(\sum_{(lon_r,lat),(lon_s,lat_k) \,\in\, N(d_{spatial})} (rain_{lon_r,lat,t} - \widehat{\mu}_{rain,spatial}(t)) \right.$$

$$\left. (rain_{lon_s,lat_k,t} - \widehat{\mu}_{rain,spatial}(t)) \right) \tag{7.21}$$

Cressie and Wikle (2011) | Wikle et al. (2019) | Auer and Rottmann (2020)

Declaration of Symbols and Equation Components

$\widehat{C}\mathbf{rain}^{(\mathbf{dspatial})}$:

> The estimated covariance of the precipitation data for a given spatial distance $d_{spatial}$.

t and t:

> Julian days that are a certain temporal distance apart.

$\mathbf{N(d_{spatial})}$:

> A set of pairs of spatial locations, for whom the spatial distance between the paired locations is within a certain interval of spatial distance $d_{spatial}$.

$\mathbf{|N(d_{spatial})|}$:

> The number of pairs spatial locations / weather stations, for whom the spatial distance lies within the interval of spatial distance $d_{spatial}$.

$\mathbf{(lon_r, lat)}$:

> A pair of longitude and latitude coordinates.

$\mathbf{rain_{lon_r,lat,t}}$:

The precipitation height value at the location (lon_r, lat) and time t.

$\widehat{\mu}_{\mathbf{rain,temporal}}(\mathbf{t})$ and $\widehat{\mu}_{\mathbf{rain,temporal}}(\mathbf{t})$:

The estimated means of the precipitation height data for times t and t, respectively.

7.2.5 Empirical Spatial Semivariogram

$$\gamma(d_{spatial}, t) := \frac{1}{2|N(d_{spatial})|} \times \left(\sum_{s_i, s_j \in N(d_{spatial})} \left(rain_{s_i,t} - rain_{s_j,t} \right)^2 \right) \tag{7.22}$$

Cressie and Wikle (2011) | Wikle et al. (2019)

Declaration of Symbols and Equation Components

$\gamma(\mathbf{d_{spatial}}, \mathbf{t})$:

The semivariogram function for the precipitation height data over a given interval of spatia distance in kilometer $d_{spatial}$ on a given day t.

$\mathbf{d_{spatial}}$:

An interval of spatial distance between weather stations measured in kilometer within which the precipitation height data are considered to be correlated.

$\mathbf{N(d_{spatial})}$:

The set of pairs of weather stations, for whom the distance between them lies within the interval $d_{spatial}$.

$|\mathbf{N(d_{spatial})}|$:

The number of pairs of weather stations in the set $N(d_{spatial})$.

$\mathbf{s_i}$ and $\mathbf{s_j}$:

Two spatial locations of a weather stations

$\mathbf{rain_{s_i,t}}$ and $\mathbf{rain_{s_j,t}}$:

Observations of precipitation height at the spatio-temporal locations s_i and s_j for the Julian day t.

7.2.6   Empirical Temporal Semivariogram

$$\gamma(d_{julian}, s) := \frac{1}{2|N(d_{julian})|} \times \left(\sum_{(t_i, t_j) \,\in\, N(d_{julian})} \left(rain_{s,t_i} - rain_{s,t_j} \right)^2 \right) \tag{7.23}$$

Cressie and Wikle (2011) | Wikle et al. (2019) | Auer and Rottmann (2020)

Declaration of Symbols and Equation Components

$\gamma(\mathbf{d_{julian}}, \mathbf{s})$:

> The temporal semivariogram function for the precipitation height data at the location s over a given interval of temporal distances d_{julian}.

$\mathbf{d_{julian}}$:

> An interval of temporal distances within which the precipitation data are considered to be correlated.

$\mathbf{N(d_{julian})}$:

> The set of pairs of julian days, for whom their distance evaluates to a value within a certain interval d_{julian}.

$\mathbf{|N(d_{julian})|}$:

> The number of pairs of julian days within the set $N(d_{julian})$.

$(\mathbf{t_i}, \mathbf{t_j})$:

A pair of two julian days

$\mathbf{rain_{s,t_i}}$ and $\mathbf{rain_{s,t_j}}$:

Observations of precipitation height at the location s and the julian days t_i and t_j, respectively.

7.3 Graphics

This subsection of the appendix encompasses graphics, which were referenced in the main part above. As is the case with any graphic in this thesis paper, these graphics were created by the author as well.

7.3.1   Dependence of Precipitation on Altitude

In the subsection concerning the results of the estimation of the GAMM, it was argued, that the confidence interval encircling the smooth term function is expanding for increasing altitude levels because there are substantially less weather stations at higher altitudes and thus less observations of precipitation. The following graphic is a histogram, that relates the count of observations to intervals of altitude level[27].
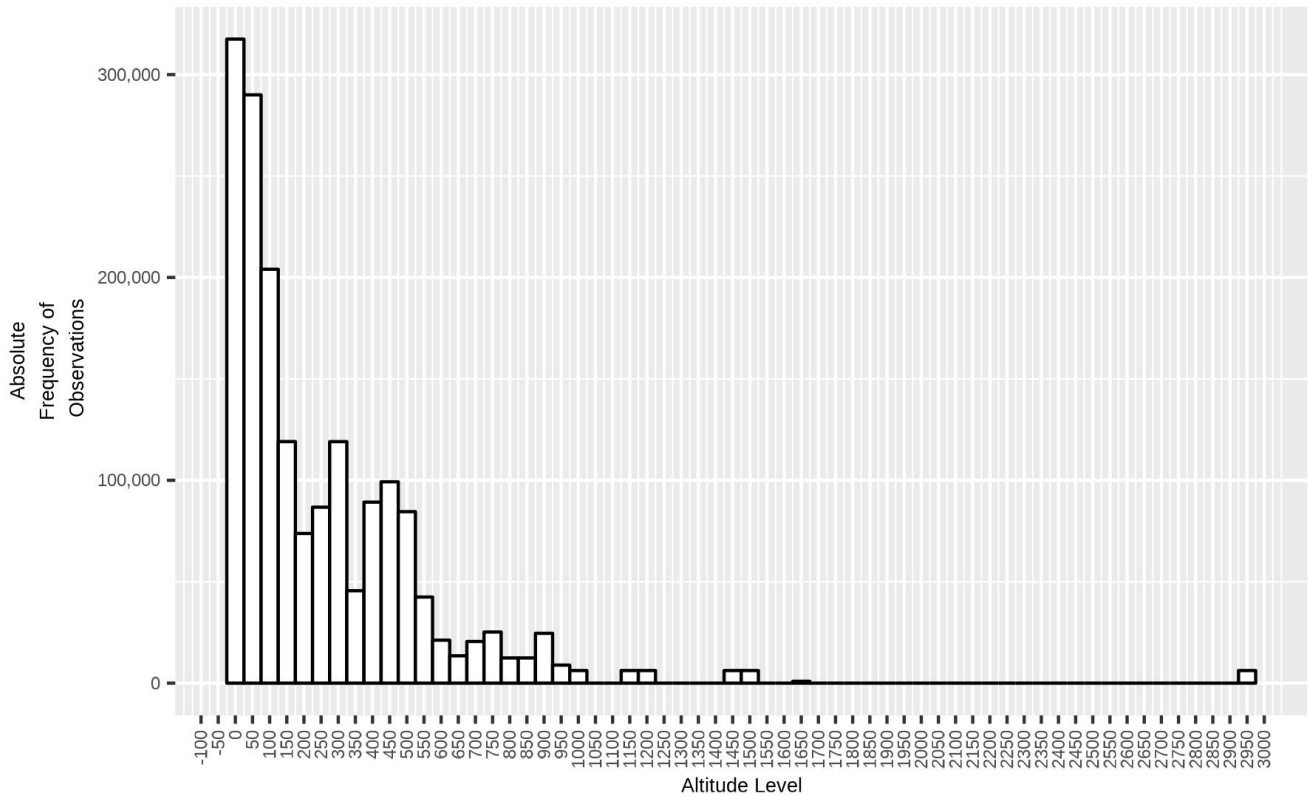

Figure 8: Altitude Distribution: Frequency of Observations Across Altitude Intervals. Own representation

---

[27].\Scripts\Plotting-(visualization)\Entirety-of-data-(no-subsetting)\Distribution of empirical frequencies\Distribution-of-Elevation-Levels.R"

7.3.2 Dependence of Precipitation on the Latitudinal Degree

In the same subsection mentioned above, which is concerned with the results of the estimation of the GAMM, alongside the relationship of the expanding confidence interval and the decreasing observations with increasing altitude, it was also argued, that a result of exploratory data analysis seems to buttress the above mentioned assertion, whereafter the graph of the smooth term of latitudinal degree does not necessarily display a strong relationship, but that the dependence is still more discernible than for longitudinal degree and julian date.

Here, a Hovmöllerplot follows, which indeed displays that there is a relationship between precipitation and latitudinal degree, even though less pronounced if compared to the relationship between precipitation and the two predictor variables mean temperature and altitude:
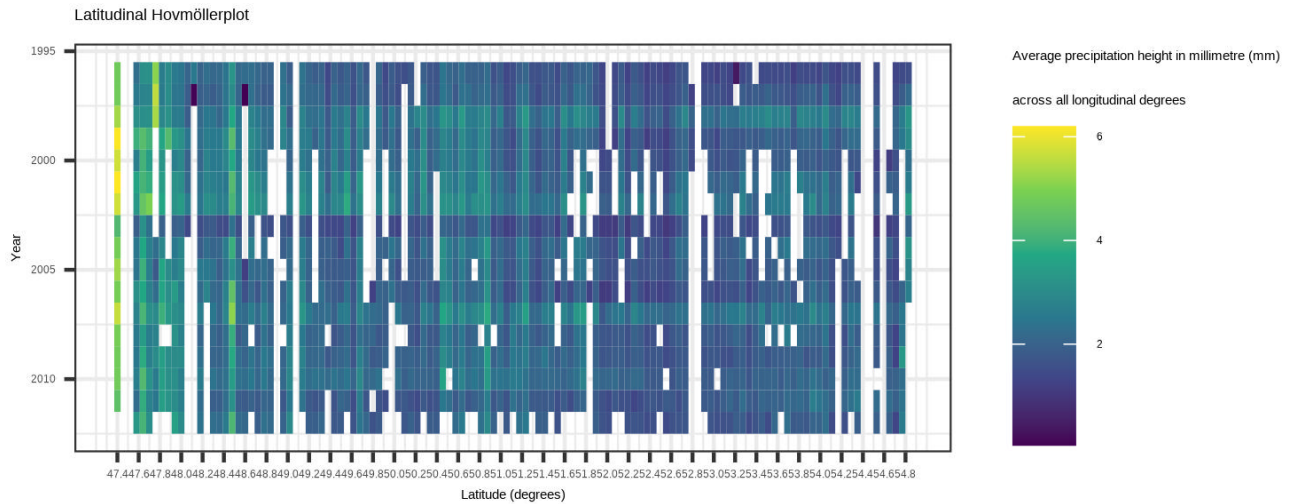

Figure 9: Latitudinal Hovmöllerplot: Mean Precipitation of Space-time bands. Own Representation

When exmaning the Hovmöllerplot, it appears, that the mean precipitation height is highest for time-space bands that involve the lowest latitudinal degrees among the weather stations situated in Germany. This supports the argument, that there is a dependency.

8 References

Benjamin Auer and Horst Rottmann. *Statistik und Ökonometrie für Wirtschaftswissenschaftler.* Springer Fachmedien Wiesbaden, Wiesbaden, 2020. ISBN 978-3-658-30136-1. doi: 10.1007/978-3-658-30137-8.

Roger Bivand, Edzer J. Pebesma, and Virgilio Gómez-Rubio. *Applied spatial data analysis with R.* Use R! Springer, New York and Heidelberg and Dordrecht and London, second edition edition, 2013. ISBN 978-1-4614-7618-4.

Hone-Jay Chu. Assessing the relationships between elevation and extreme precipitation with various durations in southern taiwan using spatial regression models. *Hydrological Processes*, 26 (21):3174–3181, 2012. ISSN 08856087. doi: 10.1002/hyp.8403.

Noel A. C. Cressie and Christopher K. Wikle. *Statistics for spatio-temporal data.* Wiley series in probability and statistics. Wiley, Hoboken, N.J, 2011. ISBN 9781119243069. URL https://ebookcentral.proquest.com/lib/kxp/detail.action?docID=4427975.

P. Dauni, M. D. Firdaus, R. Asfariani, M. I. N. Saputra, A. A. Hidayat, and W. B. Zulfikar. Implementation of haversine formula for school location tracking. *Journal of Physics: Conference Series*, 1402(7):077028, 2019. ISSN 1742-6588. doi: 10.1088/1742-6596/1402/7/077028.

Bradley Efron. Controversies in the foundations of statistics. *The American Mathematical Monthly*, 85(4):231–246, 1978. ISSN 0002-9890. doi: 10.1080/00029890.1978.11994566.

Ludwig Fahrmeir, Thomas Kneib, Stefan Lang, and Brian D. Marx. *Regression: Models, Methods and Applications.* Springer Berlin Heidelberg, Berlin, Heidelberg, 2. 2nd ed. 2021 edition, 2021. ISBN 9783662638828. URL http://nbn-resolving.org/urn:nbn:de:bsz:31-epflicht-1987674.

Luis Gimeno, Rogert Sorí, Marta Vázquez, Milica Stojanovic, Iago Algarra, Jorge Eiras-Barca, Luis Gimeno-Sotelo, and Raquel Nieto. Extreme precipitation events. *WIREs Water*, 9(6), 2022. ISSN 2049-1948. doi: 10.1002/wat2.1611.

David W. Hosmer, Stanley Lemeshow, and Rodney X. Sturdivant. *Applied logistic regression*, volume 398 of *Wiley series in probability and statistics*. Wiley, Hoboken, third edition edition,

2013. ISBN 978-0-470-58247-3. doi: 10.1002/9781118548387. URL https://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=566576.

Sina Keller and Andreas Atzl. Mapping natural hazard impacts on road infrastructure—the extreme precipitation in baden-württemberg, germany, june 2013. *International Journal of Disaster Risk Science*, 5(3):227–241, 2014. ISSN 2095-0055. doi: 10.1007/s13753-014-0026-1.

Heidi Kreibich, Meike Müller, Kai Schröter, and Annegret H. Thieken. New insights into flood warning reception and emergency response by affected parties. *Natural Hazards and Earth System Sciences*, 17(12):2075–2092, 2017. doi: 10.5194/nhess-17-2075-2017.

Christian Kuhlicke, Ines Callsen, and Chloe Begg. Reputational risks and participation in flood risk management and the public debate about the 2013 flood in germany. *Environmental Science & Policy*, 55:318–325, 2016. ISSN 14629011. doi: 10.1016/j.envsci.2015.06.011.

Annegret H. Thieken, Sarah Kienzler, Heidi Kreibich, Christian Kuhlicke, Michael Kunz, Bernhard Mühr, Meike Müller, Antje Otto, Theresia Petrow, Sebastian Pisi, and Kai Schröter. Review of the flood risk management system in germany after the major flood in 2013. *Ecology and Society*, 21(2), 2016. ISSN 1708-3087. doi: 10.5751/ES-08547-210251.

Annegret H. Thieken, Philip Bubeck, Anna Heidenreich, Jennifer von Keyserlingk, Lisa Dillenardt, and Antje Otto. *Performance of the flood warning system in Germany in July 2021 – insights from affected residents*. 2022. doi: 10.5194/egusphere-2022-244.

W. R. Tobler. A computer movie simulating urban growth in the detroit region. *Economic Geography*, 46:234, 1970. ISSN 00130095. doi: 10.2307/143141. URL http://www.jstor.org/stable/143141.

S. Trömel and C.-D. Schönwiese. Probability change of extreme precipitation observed from 1901 to 2000 in germany. *Theoretical and Applied Climatology*, 87(1-4):29–39, 2007. ISSN 0177-798X. doi: 10.1007/s00704-005-0230-4.

Dennis Wagenaar, Stefan Lüdtke, Kai Schröter, Laurens M. Bouwer, and Heidi Kreibich. Regional and temporal transferability of multivariable flood damage models. *Water Resources Research*, 54(5):3688–3703, 2018. ISSN 0043-1397. doi: 10.1029/2017WR022233.

Christopher K. Wikle, Andrew Zammit Mangion, and Noel A. C. Cressie. *Spatio-temporal statistics with R*. Chapman & Hall/CRC. CRC Press Taylor & Francis Group, Boca Raton and London and New York, 2019. ISBN 978-1-138-71113-6.

Simon N. Wood. *Generalized additive models: An introduction with R*. Texts in statistical science. CRC Press Taylor & Francis Group, Boca Raton and London and New York, second edition edition, 2017. ISBN 9781498728348. URL https://ebookcentral.proquest.com/lib/kxp/detail.action?docID=4862399.

Nina Zumel and John Mount. *Practical data science with R*. Manning, Shelter Island, 2014. ISBN 9781617291562.