Freie Universität Berlin

Doctoral Thesis

Extreme Rainfall Events:
Incorporating Temporal and Spatial Dependence to Improve Statistical Models

Inaugural-Dissertation
to obtain the academic degree
Doctor rerum naturalium (Dr. rer. nat.)

submitted to the Fachbereich Geowissenschaften

of Freie Universität Berlin

by

Oscar Esli Jurado de Larios

2022

1st reviewer:	Prof. Dr. Henning Rust Institut für Meteorologie Freie Universität Berlin
2st reviewer:	Prof. Dr. Marco Oesting Institut für Stochastik und Anwendungen Universität Stuttgart
Date of defense:	28.11.2022

Selbstständigkeitserklärung

Hierdurch versichere ich, dass ich meine Dissertation selbstständig verfasst und keine anderen als die von mir angegeben Quellen und Hilfsmittel verwendet habe.

Berlin, den 21.09.2022

Oscar Esli Jurado de Larios

Abstract

The proper design of protective measurements against floods related to heavy precipitation has long been a question of interest in many fields of study. A crucial component for such design is the analysis of extreme historical rainfall using Extreme Value Theory (EVT) methods, which provide information about the frequency and magnitude of possible future events. Characterizing an entire basin or geographical catchment requires the extension of univariate EVT methods to capture the spatial variability of the data. This extension requires that the similarity of the data for nearby stations be included in the model, resulting in more efficient use of the data.

This dissertation focuses on using statistical models incorporating spatial dependence for modeling annual rainfall maxima. Additionally, we present ways of adapting the models to capture the dependence between rainfall of different time scales. These models are used in order to pursue two aims. The first aim is to improve our understanding of the mechanisms that lead to dependence on extreme rainfall. The second aim is to improve the resulting estimates when incorporating the dependence into the models.

Two published studies make up the main findings of this dissertation. The models used in both studies involve the use of Brown-Resnick max-stable processes, allowing the models to explicitly account for the dependence on either the temporal or the spatial domain. These conditional models are compared for both cases to a model that ignores the dependence, allowing us to determine the impact of the dependence in both situations. Contributions to three other studies using the concept of dependence are also summarized.

In the first study, we assess the impact of including the dependence between rainfall series of different aggregation durations when estimating Intensity-Duration-Frequency curves. This assessment was done in a case study for the Wupper catchment in Germany. This study found that including the dependence in the model had a positive effect on the prediction accuracy when focusing on rainfall with short durations ($d \leq 10$ h) and large probabilities of non-exceedance. Therefore, we recommend using max-stable processes when a study focuses on short-duration rainfall.

In the second study, we investigate how the spatial dependence of extreme rainfall in Berlin-Brandenburg changes seasonally and how this change could impact the estimates from a max-stable model that includes this dependence. The seasonality was determined by estimating the parameters of a summer and winter semi-annual block maxima model. The results from this study showed that, for the summer maxima, the dependence structure was adequately captured by an isotropic Brown-Resnick model. On the contrary, the same model performed poorly for the winter maxima, suggesting that a change in the assumptions is needed when dealing with typical winter events, typically frontal or stratiform for this region. These results show that accounting for the meteorological properties of the rainfall-generating processes can provide useful information for the design of the models.

Overall, our findings show that including meteorological knowledge in statistical models can improve their resulting estimations. In particular, we find that, under certain conditions, using statistical dependence to incorporate knowledge about the differences in temporal and spatial scales of rainfall-generating mechanisms can lead to a positive impact in the models.

Zusammenfassung

Die richtige Auslegung von Schutzmaßnahmen gegen Überschwemmungen im Zusammenhang mit Starkniederschlägen ist seit langem eine Frage, die in vielen Studienbereichen von Interesse ist. Eine entscheidende Komponente für eine solche Planung ist die Analyse extremer historischer Niederschläge mit Methoden der Extremwertstatistik, die Informationen über die Häufigkeit und das Ausmaß möglicher künftiger Ereignisse liefern. Die Charakterisierung eines ganzen Einzugsgebiets oder einer geografischen Einheit erfordert die Erweiterung der univariaten Extremwerstatistik-Methoden, um die räumliche Variabilität der Daten zu erfassen. Diese Erweiterung erfordert, dass die Ähnlichkeit der Daten für nahe gelegene Stationen in das Modell einbezogen wird, was zu einer effizienteren Nutzung der Daten führt.

Diese Dissertation konzentriert sich auf die Verwendung statistischer Modelle, die die räumliche Abhängigkeit bei der Modellierung von jährlichen Niederschlagsmaxima berücksichtigen. Darüber hinaus werden Möglichkeiten zur Anpassung der Modelle vorgestellt, um die Abhängigkeit zwischen Niederschlägen auf verschiedenen Zeitskalen zu erfassen. Diese Modelle werden zur Verfolgung zweier Ziele eingesetzt. Das erste Ziel besteht darin, unser Verständnis der Mechanismen zu verbessern, die zur Abhängigkeit von extremen Niederschlägen führen. Das zweite Ziel besteht darin, die resultierenden Schätzungen zu verbessern, wenn die Abhängigkeit in die Modelle einbezogen wird.

Zwei veröffentlichte Studien bilden die wichtigsten Ergebnisse dieser Dissertation. Die in beiden Studien verwendeten Modelle basieren auf max-stabilen Brown-Resnick-Prozessen, die es den Modellen ermöglichen, die Abhängigkeit entweder auf der zeitlichen oder auf der räumlichen Ebene ausdrücklich zu berücksichtigen. Diese bedingten Modelle werden für beide Fälle mit einem Modell verglichen, das die Abhängigkeit ignoriert, so dass wir die Auswirkungen der Abhängigkeit in beiden Situationen bestimmen können. Es werden auch Beiträge zu drei anderen Studien zusammengefasst, die das Konzept der Abhängigkeit verwenden.

In der ersten Studie bewerten wir die Auswirkungen der Einbeziehung der Abhängigkeit zwischen Niederschlagsreihen unterschiedlicher Aggregationsdauern bei der Schätzung von Intensitäts-Dauer-Frequenz-Kurven. Diese Bewertung wurde in einer Fallstudie für das Einzugsgebiet der Wupper in Deutschland durchgeführt. Diese Studie ergab, dass sich die Einbeziehung der Abhängigkeit in das Modell positiv auf die Vorhersagegenauigkeit auswirkt, wenn man sich auf Niederschläge mit kurzen Dauern ($d \leq 10$ h) und großer Nichtüberschreitungwahrscheinlichkeit konzentriert. Daher empfehlen wir die Verwendung von max-stabilen Prozessen, wenn sich eine Studie auf Regenfälle von kurzer Dauer konzentriert.

In der zweiten Studie untersuchen wir, wie sich die räumliche Abhängigkeit von Extremniederschlägen in Berlin-Brandenburg saisonal verändert und wie sich diese Veränderung auf die Schätzungen eines max-stabilen Modells auswirken könnte, das diese Abhängigkeit berücksichtigt. Die Saisonalität wurde durch die Schätzung der Parameter eines halbjährlichen Sommer- und Winter-Blockmaxima-Modells bestimmt. Die Ergebnisse dieser Studie zeigten, dass die Abhängigkeitsstruktur für die Sommermaxima durch ein isotropes Brown-Resnick-Modell angemessen erfasst wurde. Im Gegensatz dazu zeigte dasselbe Modell eine schlechte Leistung für die Wintermaxima, was darauf hindeutet, dass eine Änderung der Annahmen erforderlich ist, wenn es um typische Winterereignisse geht, die in dieser Region typischerweise frontal oder stratiförmig sind. Diese Ergebnisse zeigen, dass die Berücksichtigung der meteorologischen Eigenschaften der Niederschlagsprozesse nützliche Informationen für die Gestaltung der Modelle liefern kann.

Insgesamt zeigen unsere Ergebnisse, dass die Einbeziehung von meteorologischem Wissen in statistische Modelle die daraus resultierenden Schätzungen verbessern kann. Insbesondere stellen wir fest, dass unter bestimmten Bedingungen die Nutzung der statistischen Abhängigkeit zur Einbeziehung von Wissen über die Unterschiede in den zeitlichen und räumlichen Skalen der regenerzeugenden Mechanismen zu einer positiven Wirkung in den Modellen führen kann.

List of Publications

This dissertation includes material (either in full or partial) from the following publications, with the specific contribution of the author stated below:

1. **Jurado, O. E.**, Ulrich, J., Scheibel, M., & Rust, H. W. (2020). Evaluating the Performance of a Max-Stable Process for Estimating Intensity-Duration-Frequency Curves. Water, 12(12), 3314. http://doi.org/10.3390/w12123314

Contribution I developed the initial research questions with H.R. The methodology was developed and implemented by me, as well as the analysis and discussion of results. I wrote the initial manuscript, which was edited in collaboration with H.R. and J.U. I was in charge of data processing, using data provided by M.S.

2. **Jurado, O. E.**, Oesting, M., & Rust, H. W. (2022 - In submission). Implications of Modeling Seasonal Differences in the Extremal Dependence of Rainfall Maxima. *Currently submitted to Stochastic Environmental Research and Risk Assessment.* Preprint available in http://arxiv.org/abs/2207.03993

Contribution I developed the initial research questions with H. W. Rust. The methodology was developed in collaboration with H.R. and M.O. The implementation (including creation of software) was performed by me, as well as the analysis and discussion of results. I wrote the initial manuscript, which was edited in collaboration with H.R. and M.O.

3. Ulrich, J., **Jurado, O. E.**, & Rust, H. W. (2020). Estimating IDF curves consistently over durations with spatial covariates. Water, 12(11), 1-22. http://doi.org/10.3390/w12113119

Contribution I supported the discussions regarding the methodology, and in particular, the possible role of dependence in the final estimates. I helped with some of the development of the published software. Furthermore, I revised and helped to edit the manuscript.

4. Fauer, F. S., Ulrich, J., **Jurado, O. E.**, & Rust, H. W. (2021). Flexible and consistent quantile estimation for intensity-duration-frequency curves. Hydrology and Earth System Sciences, 25(12), 6479–6494. http://doi.org/10.5194/hess-25-6479-2021

Contribution I developed the concept and wrote software for the section regarding the coverage of confidence intervals for different levels of dependence between durations. Additionally, I helped write the corresponding section in the manuscript. I also revised and helped to edit the manuscript.

5. Otero, N., **Jurado, O. E.**, Butler, T., & Rust, H. W. (2022). The impact of atmospheric blocking on the compounding effect of ozone pollution and temperature: A copula-based approach. Atmospheric Chemistry and Physics, 22(3), 1905–1919. http://doi.org/10.5194/acp-22-1905-2022

Contribution

> I supported N.O. in the implementation of Copulas and their interpretation. I provided software for implementing the copulas in the analysis, and helped with the design of the statistical methods. I also revised the section of the paper involving copulas, and helped with the analysis of results.

6. Berghäuser L., Schoppa L., Ulrich J., Dillenardt L., **Jurado, O. E.**, Passow C., Mohor G. S., Seleem O., Petrow T., Thieken A. H. (2020). Starkregen in Berlin: Meteorologische Ereignisrekonstruktion und Betroffenenbefragung, Technical Report, Universität Potsdam, 44pp. http://doi.org/10.25932/publishup-50056

Contribution

> I collected and processed the precipitation data used for the statistical analyses of the event. I researched and wrote the section on the meteorological background for the 2017 event, as well as the section explaining the radar data. I also supported the efforts of J.U. for the statistical analyses.

Acknowledgements

First of all, I wish to thank my supervisor and first reviewer, Prof. Dr. Henning Rust, for all his support during the last four years. Thank you for all the discussions about my research and your guidance and advice that allowed me to become a better researcher and finish this project.

Furthermore, I wish to thank the second reviewer Prof. Dr. Marco Oesting for his support and help with all my questions during my stay with his working group.

The work done for this dissertation was possible thanks to the financial support of the Mexican National Council of Science and Technology (CONACyT) and the German Academic Exchange Service (DAAD) through the joint project "Country-related cooperation programme with Mexico: CONACyT PhD 2018". Additionally, I thank the DFG graduate research school NatRiskChange for their financial support.

I thank all my colleagues from the Institute of Meteorology for their academic support. In particular, I am grateful to all the members of the Statmet and the CliDia working groups for all their constructive and supportive discussions.

Moving to a new country can be rather challenging, so I give a big thank you to all the great friends I met during this time: Felix, Sebastian, Andreas, Edgar, Joscha, and Christoph. Thanks for all the great moments!

Special thanks go to Jana and Christian, who were there for me from the very beginning. Jana - thank you for all your input and help with all my papers. I truly would not have finished this dissertation without your help! Christian, thanks for all the coffee and banana breaks – they were always a highlight of my day. ¡Gracias amigos!

The biggest credit goes to my family back in Mexico, who have supported me for the longest time and patiently waited for me to finish this dissertation. Agradezco sobre todo a Alma, Rosita, Ricardo, Jagger, Sarah, Pupus, Castor, Mary, Beto, Marisol y Hector. ¡Gracias por su apoyo todos estos años!

Finally, the largest thank you goes to my wonderful husband H. Michan, whose support was invaluable for finishing this project. Thanks for sticking with me all these years!

Contents

List of Abbreviations

GEV	Generalized Extreme Value (distribution)
GPD	Generalized Pareto Distribution
EVT	Extreme Value Theory
FTT	Fisher-Tippett(-Gnedenko) Theorem
POT	Point-Over-Threshold
GP	Gaussian process
KL	Kullback-Leibler (divergence)
CV	Cross-Validation
AIC	Akaike Information Criterion
WAIC	Widely Applicable Information Criterion
LOO-CV	Leave-One-Out-cross-validation
PSIS	Pareto-S-Importance-Sampling
GLM	Generalized Linear Model
VGLM	Vector Generalized Linear Model
DM	Bayesian Distributional Model
BHM	Bayesian Hierarchical Model
QS	Quantile Score
QSS	Quantile Skill Score
QSI	Quantile Skill Index
DWD	German Weather Service
IPCC	Intergovernmental Panel on Climate Change
d-GEV	Duration-dependent GEV distribution
IDF	Intensity-Duration-Frequency (curve)
rd-GEV	Reduced d-GEV-based approach for modeling IDF curves
MS-GEV	Max-stable-based approach for modeling IDF curves
BR	Brown-Resnick model

“*Ojalá que llueva café en el campo*”

Juan Luis Guerra

1 Introduction

In the decade 2010-2019, the reported monetary losses resulting from disasters associated with extreme weather events were, on average, US$ 383 million per day (World Meteorological Organization, 2021). These staggering losses were compounded by the loss of human life, reported to be 185,000 deaths for the same decade. Of all the different weather extremes (e.g., hot extremes on land and in the ocean, drought, wind storms, fire weather, among others), the overwhelming bulk of damages and human loss was reported for precipitation events, which include droughts, tropical cyclones, and floods. Figure 1.1 shows that flooding due to heavy precipitation (excluding tropical cyclones) accounted for 44% of reported weather-related disasters, 16% of reported deaths, and 31% of worldwide economic losses due to extreme weather from 1970-2019 (World Meteorological Organization, 2021). Moreover, climate change has increased the likelihood and severity of extreme weather events causing impactful floods and droughts (IPCC, 2022; Caretta et al., 2022).

Given the high societal and environmental impact posed worldwide by heavy precipitation events, a considerable scientific and engineering endeavor has been undertaken in the past decades to understand better the risks and impacts of extreme precipitation and to design adaptation measures against such events. These measures have contributed to a significant decrease in reported deaths related to floods throughout the decades, even as the total number of events has increased steadily in the same period (World Meteorological Organization, 2021).

One of the greatest challenges in designing adaptation measures against heavy precipitation events is predicting the magnitude and frequency of expected events. This dissertation focuses on how a particular type of statistical models can provide valuable predictions for the design of the previously mentioned infrastructure.

In the following section, an explanation of the connection between adaptation measures against heavy precipitation and statistical modeling is explored. Furthermore, several aspects of what is considered to be extreme rainfall are explored before moving on to the statistical section. Afterward, the aims and goals of this dissertation, as well as the structure will be described.

1.1 Relevance of rainfall modeling for adaptation measures

Flood protection measures can be broadly cataloged as either hard infrastructure measures or soft measures. **Hard infrastructure measures** physically control water flow through streams to prevent water levels from overflowing. Examples include

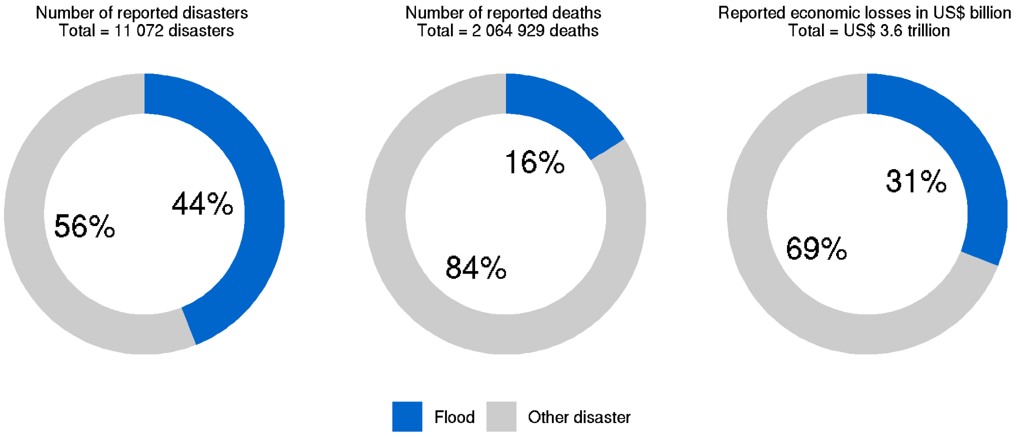

FIGURE 1.1: Proportion of flood-associated (left) number of disasters, (middle) number of deaths, and (right) economic losses globally for 1970-2019. Figure adapted from World Meteorological Organization (2021).

dams, levees, dikes, storm drains, and flood control gates. On the other hand, **soft measures** are human actions that generate and increase the awareness of floods, resulting in higher resilience. Examples of soft measures include proper city planning, management, and strong community support in preparation for a flooding event.

Of the two types of adaptation measures, the most common ones are hard infrastructure measures for flood protection (Andrew et al., 2017). Therefore, properly designing these infrastructure projects is crucial to avoid maladaptation. Typically, the design of hydrological infrastructure is done in order to protect against a flooding event of a certain critical level: if this level is exceeded within the structure's lifetime, then the structure's failure is expected to occur (Vogel et al., 2017). Therefore, hydrological structures are commonly designed to withstand events with very low occurrence probabilities during the structure's lifetime. Here we encounter a trade-off when choosing the level of protection: as the level of protection increases, the price of construction and maintenance grows significantly. However, choosing a level that is too low may result in heavy economic losses. Therefore, an equilibrium is required between choosing a level of protection that is neither too low nor too high. This design choice is the basis of hydrological design.

The design of hydrological infrastructure to protect against flooding due to heavy precipitation requires the knowledge of what magnitude of heavy precipitation events are to be expected and their frequency. Furthermore, this information needs to be known well in advance before the occurrence of the extreme event, as the construction of said infrastructure typically lasts in a timespan that ranges from years to decades. Therefore, the use of deterministic weather forecasting tools, like Numerical Weather Predictions, which have a relatively short lead time before the event, does not provide the necessary information with the required anticipation time. Thus, what is needed is a way of using previous knowledge about past events in order to predict how possible events will look in the future.

A way to predict possible future extreme events is to look back at previous events to get an idea of what is likely to occur. In principle, if we had a long continuous series of past rainfall extremes for a given location, we could estimate the stochastical behavior of these processes from which we could estimate the probability of events that are more extreme than any that have been observed. However, two problems arise in that case: the first one is that extreme events are, by definition, rare, so the amount of data available for characterizing the process is somewhat limited. Secondly, we could be interested in predicting events for locations where no previous records exist. In the following chapters, extreme value theory and spatial extremes will be introduced, which offer solutions to both problems. First, nevertheless, a more detailed definition of what is considered to be extreme rainfall is introduced.

1.2 Extreme rainfall

Precipitation is any form of water (liquid or solid) that falls from a cloud and reaches the ground. Of the different types of precipitation (e.g., hail, snow, sleet), the focus of this work falls exclusively on the liquid component, typically known as **rainfall**. Rainfall occurs when a series of complex processes result in cloud water droplets growing large enough[1] to overcome the upward forces of motion that otherwise would keep them aloft in the form of what we know as clouds.

This dissertation focuses on modeling of **extreme rainfall** regardless of its physical cause. However, an early problem arises when talking about extremes in rainfall events: In contrast to other meteorological extreme events like heat waves or tornados, there is no fixed definition of what exactly constitutes extreme rainfall. From a logical point of view, it is clear that rainfall events that lead to unusually severe and costly flooding should be considered extreme events. However, many different factors can influence the impact of any given rainfall event. For example:

- the duration of the rainfall event;
- the overall intensity;
- the spatial extent of the rainfall field;
- the physical properties of the geographical catchment where the event occurred; or
- the existing hydrological infrastructure and human settlements.

Defining whether a rainfall event can be classified as extreme can then depend on the context. Furthermore, events considered extreme in the past could be no longer extreme in the future, as climate change likely increases the frequency and magnitude of such events.

An important consideration of rainfall events is that their respective rainfall generation process can lead to different durations/time-scales, which in turn, leads to significant differences in the impact in the hydrological infrastructure. For context, an analysis following the methodology of Bohnenstengel et al. (2011) of 10 measuring stations in Berlin containing hourly precipitation height from the DWD revealed that events that last longer than 9 hours correspond to only 10% of the total count of events, but account for 47% of the total precipitation amount. In contrast, events that were shorter than 1 hour accounted for 30% percent of all events, but only contained 2% of the total precipitation. This imbalance further highlights the importance of distinguishing between types of rainfall-generating processes when modeling them.

For this dissertation, we focus on two types of rainfall-processes typically linked to extremes: convective and frontal/stratiform events. Fronts are defined in Glickman (2000) as the interface or transition zone between two air masses of different density.

[1] For context, a typical cloud droplet is 100 times smaller than a typical raindrop.

Furthermore, Lackmann (2011) indicates that not all airmass boundaries should be classified as fronts on surface analysis; the analyst should rather account for additional variations in variables like density, potential temperature or surface charts. A critical aspect of frontal zones is that they are long and narrow: along-front they present synoptic scales of 1000 km, while cross-front they have mesoscale scales of 100 Km. Rainfall events are usually generated on the along-front section, resulting in long and narrow rainfall events.

On the other hand, convective systems are a result of regional atmospheric instability that leads to upward movement of an air mass, resulting in cloud development and eventually, in localized storms that lead to heavy precipitation. The atmospheric instability is typically a result of strong temperature differences, which is common in summer, when the surface is heated by strong solar irradiation. Compared to frontal events, convective events have a much smaller spatial scale of around 10-20 km and a shorter time scale of minutes to hours. However, the resulting storms from these events are capable of causing the most extensive damage of all severe weather events.

Two types of methods exist to define a rainfall event as extreme. The first one is by using so-called climate indices, which are based on basic statistics from past observed events. These indices are helpful when studying and classifying past or recently forecasted events. However, the indices are exclusively derived from observations that have been observed; they provide no information about *possible* large events that have never been observed. For this, it is required to use the distribution of past rainfall extremes (for example, by looking at the distribution of the largest rainfall event from every year). Rainfall events that fall in the far-right of such a distribution would then be considered extreme events. This approach to defining (and modeling) events is given by Extreme Value Theory (EVT), which will be detailed in chapter 5.

This dissertation will examine extreme rainfall events from the extreme value theory point of view. That is, we will consider a rainfall event as extreme when it falls in the far-right tail of the empirical distribution of rainfall maxima obtained from the existing records in that location. This definition means that events considered extreme in one location could be considered non-extreme in locations with different rainfall distributions. However, a big focus of this thesis is the idea that nearby locations share many properties, so the threshold to denote an event as extreme is typically similar for neighboring locations.

1.2.1 Example of an extreme rainfall event

An example of what can be considered an extreme event from an EVT point of view is given by the rainfall event that occurred in the region of Berlin on the 29-30th of July 2017, as detailed in the report of Berghäuser et al. (2021). This event resulted from the collision of two low-pressure areas: the one known as Rasmund II, which formed over the Czech Republic and extended over Poland to Berlin, and another low-pressure system that approached from southern Germany. This collision resulted in a widespread rain area of long-lasting convective rainfall.

Figure 1.2 shows the 24-hour accumulated precipitation sum derived from the RADKLIM data provided by the German Weather Service (DWD). This figure shows that most rainfall fell in a circular region centered around the Tegel station. Around this area, precipitation height values of up to 170 mm can be observed (in fact, this is likely an underestimation of the approximated radar value, as ground-based stations reported 24-hour accumulated height values of up to 197 mm (Gebauer et al., 2017)). Most of the damages reported were confined to the area near the storm's center, as

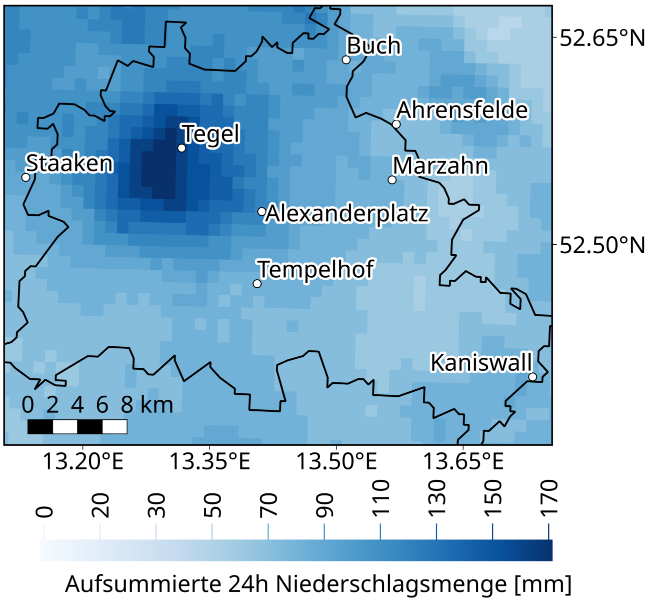

FIGURE 1.2: 24-hour precipitation sums derived from radar data for the event of 29-30th of July 2017. Image reproduced from Berghäuser et al. (2021) with permission of the authors.

seen by the geographical distribution of emergency calls made during/afterward the event (see Fig. 5 of Berghäuser et al. (2021)).

To investigate if the 2017 rainfall event in Berlin could be classified as an extreme event, Berghäuser et al. (2021) used the EVT approach mentioned in the last section to calculate the annual probability of non-exceedance for different amounts of precipitation. Their report found that in the vicinity of the point with the highest recorded rainfall, the aggregated 24-hour precipitation depth could be classified as an extreme event, with a probability $p < 0.005$ of happening in any given year. In contrast, when looking at lower aggregation durations, like one or 3 hours, the event could no longer be classified as extreme, as the measured value had a probability of occurrence in any given year in the range of $25\% \leq p \leq 50\%$. Furthermore, Fig. 1.2 shows that this event had considerable spatial heterogeneity: some points merely 10 km away from the center of the storm reported much lower rainfall. For example, the amount of rainfall recorded in station Tempelhof had a relatively large probability of occurrence for all aggregation durations. All of this points to the fact that defining a rainfall event as extreme depends on many considerations.

1.3 Aims and relevance of this dissertation

There are two primary aims of this dissertation:

- To improve our understanding of the mechanisms that lead to statistical dependence in extreme rainfall datasets.

- To improve the extreme rainfall estimates made from models by incorporating the dependence in a meaningful way.

To fulfill these aims, this dissertation contains two novel studies, as well as contributions made to two other collaborative studies:

- In the first study, presented in **chapter 4**, a novel method used to integrate statistical dependence between rainfall maxima of different aggregation durations is evaluated to determine its impact in the final estimates. This evaluation required the proposal of a logarithm distance to measure how "close-by" different durations were: this so-called log-distance was found to overperform over the existing euclidean distance. This study finds that the dependence structure for this dataset is rather complex and that, under certain conditions, it can be ignored for certain modeling purposes. Moreover, the results and methods developed for this study were essential for the other two studies described in ch. 6.

- The second study, presented in **chapter 5**, changed the focus from the dependence between durations to the spatial dependence of rainfall maxima. This novel study investigates how different regimes of rainfall-generating mechanisms can lead to different dependence structures. Moreover, the study develops a method to determine the impact of the differences in the dependence structure for estimates originating from the same model. This study finds that convective and stratiform/frontal events show different dependence structures. Therefore, information about the dominating rainfall-generating regime needs to be integrated into the design of models that include spatial dependence.

- Finally, in **chapter 6**, the contributions made to other three studies performed in the context of this dissertation are presented. The first of these studies took direct advantage of the results seen from ch. 4, which justified some of the modeling choices made for the study. Additionally, the methods developed in ch. 4 are shown to be an essential part of another study, where they formed the basis for coverage analysis. On the whole, this chapter shows how the topic of dependence can be generalized to improve the understanding of other topics related to statistical modeling of extreme rainfall.

The importance and originality of this study are that it explores a novel interpretation of already existing statistical methods for incorporating dependence in stochastical models from a meteorological perspective. These relatively recent methods were developed in mathematics, without focusing deeply on meteorology. Therefore, this dissertation offers some important insights into the applicability and caveats of adapting such statistical methods for modeling extreme rainfall. An example is the proposal of the log-distance in ch. 4, which comes from scaling considerations of rainfall processes for different durations. Another example comes from ch. 5, where the physical characteristics of different rainfall processes are the starting point for studying how the spatial dependence structure changes seasonally. To our knowledge, very little research has been done on this latter topic; therefore, our study makes a major contribution and a great starting point for further research.

1.3.1 Structure of this dissertation

This dissertation is organized as follows: First, the theoretical framework behind the studies mentioned above is detailed in chs. 2 and 3. Chapter 2 encompasses a primer on statistical modeling, introducing the basis for the models used in the subsequent studies. Chapter 3 builds upon this introduction by describing the state-of-the-art behind current multivariate methods for models that include dependence structures. Chapters 4 - 6 comprise the published studies developed during the course of this dissertation, which use the methods described in the first chapters. Finally, the ch. 7 offers an overview, summary, and outlook on the topic.

Part I

Theoretical Framework

2 Statistical Modeling

The last chapter introduced some of the societal impacts posed by extreme rainfall events and their adaptation measures. These measures against extreme rainfall require information about their characteristics like frequency, magnitude, and extent. Deriving this information from past observations requires using a special tool: the statistical model. This chapter will provide an overview of statistical models and their steps. The methods described in this chapter are very general; a particular extension for extreme rainfall will be provided in the next chapter.

2.1 Overview of statistical models

Statistical models are mathematical constructions used to describe data that shows some random component. As such, they contain an element of randomness, which differs from deterministic models. As their primary goal, these kinds of models attempt to represent what is called the **data-generating process**. Knowing more about this process can answer several questions of interest for an analyst. Some applications where these models give solutions include predictions of unseen data, extraction of causal information, and description of the stochastic structure in the data.

This work mainly focuses on using statistical models to **predict** unseen values from the existing observations. That is, it does not attempt to use statistical models to identify causal relationships. The task of identifying causal relationships from models differs greatly from prediction, and as such, it has its own specialized procedure, which is out of scope for this dissertation. In general, statistical models with the goal of prediction follow the following overarching steps:

1. Identify the data relevant to the research question.
2. Specify a model that adequately describes the data-generating process for the data coming from the last step.
3. Calibrate the model so that it can represent the data.
4. Verify the assumptions used for creating the model.
5. Obtain predictions from the model.
6. Validate the predictions made from the model.

The procedure detailed above is not entirely rigid; in fact, sometimes it is necessary to perform some steps in disorder or repeat others to reach a model that the analyst is satisfied with. In particular, the last four steps are commonly iterative, as sometimes it is necessary to recalibrate the model to improve the resulting predictions from the model.

In the first step, identifying the relevant data, it is also necessary to explore how the data is structured and identify what variables are to be predicted and which ones could work as predictors, helping to explain the variability in the data. This step, naturally, is preceded by the actual collection of data, which is a complete topic in and of itself. Furthermore, it is important in this step to define the data's scale and identify if we are dealing with a discrete, continuous, or categorical variable.

For the second step, a statistical model that describes a possible data-generating process is selected. This selection typically requires exploratory data analysis and domain expertise for the particular problem. Typically, the chosen model comes from an already existing parametric family; however, this model is not yet calibrated for the specific data identified in the first step. Calibrating the model to describe the specific data is done by finding values of quantities known as model parameters. The next step in the modeling process is using the data to find proper values for these parameters. It should be noted that the random nature of statistical models makes finding the exact true value of the parameters usually unattainable, so uncertainty about their value will always exist.

Every statistical model makes certain assumptions about the data and its generating process. Therefore, after finding and calibrating the model, the next step is always to check the assumptions. This check ensures that the model is consistent with its own logical rules; however, it does not say how well the model describes the actual data-generating process. Unfortunately, finding whether a model is a realistic description of the data-generating process is usually unfeasible without additional tools to study causality, a topic that, for this thesis, is out of scope. In any case, we want a model that is at least logically consistent with its own assumptions.

The final two steps are to obtain predictions from the model and to check how good the predictions are. The last step depends on the goal set for the model predictions. In light of this, there is no single definition of what constitutes a perfect prediction. For the studies described in this thesis, the goal is to get predictions that are good representations of unobserved values. As we will see below, getting good predictions for unobserved values requires that the model does not learn too much from what was observed already. This is done, for example, by penalizing models that learn too much from the data, using quantities known as Information Criteria. The ultimate goal of these kind of methods is to avoid the phenomenon known as overfitting, as overfitted models typically provide very poor estimates of unobserved future values.

In this chapter, every one of the steps mentioned earlier will be explained in more detail, except for the first step, which is skipped. I begin with the process of identifying a model to describe the data, for which I also define what a model is in more detail. The rest of the chapter will more or less cover the other steps point by point, finishing with a summary of them all.

2.2 Model Specification

The goal of statistical modeling is not only to summarize what has already been observed but additionally, to infer the characteristics of the process that generated the data. However, several difficulties need to be overcome to achieve this goal: Firstly, it may not be easy to infer what we want to know from the observed data. Additionally, observations from real-world phenomena virtually always contain some random variability; thus, if the data-gathering process were repeated several times, different data would be obtained for each replication. The randomness in the observations creates what, at first glance, looks like an insurmountable problem: We typically only have access to a single set of data, but due to the random variability, any conclusions drawn from this sample could potentially not be valid for all possible samples. Statistical models counteract this problem by incorporating a random component whose properties can be adjusted to end up with a model that describes how the data *might* have been generated.

A statistical model is a simplified description of the data-generating process that produced the observations. These models are simply mathematical formulas that characterize the trends and spread in the data and include a stochastic component that can reproduce the variability of the observations. These models combine known information (observations and predictors) with unknown information (stochastic component). The stochastic component of models makes them incredibly useful for describing many natural processes that can be seen as random. However, no single "universal" model that can explain every observed phenomenon exists. In reality, many different statistical models can usually describe the same observed data. Therefore, statistical modeling aims to find the model that gives the most credible description of the data.

Let Y represent the measured quantity, from which the actual individual measurements are denoted as y. The measurements y are assumed to be random, which in this context means that for every measurement of Y, we obtain a different y value, where some values are more likely than others. Without going now into much mathematical rigor, a statistical model is essentially a function that assigns a probability to events associated with Y. The specifics of the function depend on whether Y is discrete or continuous. For a discrete Y, the model f takes the form of a **probability mass function** $f(x) = \Pr[X = x]$. In contrast, when Y is continuous, the model is given as the **probability distribution function** $F(x) = \Pr[X \leq x]$. The distribution function can be differentiated to obtain the probability density function.

For both the discrete and continuous cases, the model's formula contains quantities, called **parameters**, that determine the exact shape of the resulting function. The parameters of a model are typically denoted by the greek letter θ. The models can then be written as $f(x|\theta)$ or $F(x|\theta)$. The formulas used for the models almost always come from already existing families of models, which have helped describe to describe different kinds of processes.

An essential assumption for modeling is that every measurement y from the variable Y comes from the same data-generating process. This assumption implies that the model is an accurate description for every observed y. In statistical jargon, one would say that the data are identically distributed. Additionally, a common assumption is that every y does not depend on previous or future realizations of Y (i.e., that they are independent). If both assumptions are made, the data is said to be **independent and identically distributed**, or i.i.d., in short. The latter assumption can be relaxed for many statistical models, as seen in the next chapter, where the addition of dependence is explored for different types of models.

We can then distinguish two different components for a statistical model:

- The mathematical formula of the model, which describes the general shape and behavior. For example, consider the Gaussian distribution (also known as the normal distribution), which is a model that has a well-known bell-shaped function and is very often used to describe data.

- The parameters θ that control different aspects of the final shape given by the formula. For example, the Gaussian distribution mentioned above contains two parameters: the mean and the standard deviation. The Gaussian distribution always has the same bell shape, but the two parameters determine precisely how wide the bell is and where in the real line it will be centered.

When designing a model, the model parameters θ are always initially unknown (albeit a prior idea of their value sometimes exists). Different values of θ will lead to different data-generation processes being simulated. Thus, the goal is to find values of θ that result in a data-generation process capable of reproducing the observed data. The resulting value of the parameters can then be used to answer questions of interest about the data-generating process of Y. In other words, a statistical model is a recipe by which Y might have been generated, given appropriate values for θ (Wood, 2015).

Statistical models can range from very simple, containing only one or two parameters, to highly complex, containing hundreds to thousands of parameters. All things considered, there are two main desiderata that statistical models should possess. The first is that the formula of the model should be understandable, with a meaningful and interpretable number of parameters [1]. The most commonly used models, like the Gaussian distribution, have well-defined formulas with parameters that are easy to identify. The second desideratum for a model is that the resulting function should be similar to the data; after all, a model that cannot reproduce the data used to create it is rather useless.

Example of a statistical model

Let Y represent the mean annual 2 m air temperature in a specific location measured for 30 years. In this case, y_i would represent the mean annual temperature for the year i. Our goal is to find a model that can both describe the random variability of Y and provide information about the data-generating process.

First, we assume that the observations y_i are i.i.d. Then, we propose that a Gaussian distribution can replicate the data-generating process for the observations. This statement can be formulated as:

$$Y \sim \mathrm{N}(\mu, \sigma^2), \tag{2.1}$$

where μ and σ represent the unknown parameters θ of the model. The Gaussian distribution has a well-known mathematical formula for the probability density function, which for a single observation y_i is given by:

$$f(y_i|\theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-(y_i - \mu)^2}{2\sigma^2}\right). \tag{2.2}$$

For the expression above we can see how the two parameters, μ and σ, control the resulting shape of the distribution. Thus, after choosing a distribution, the goal is to find values of the parameters that result in a model that resembles the data.

2.2.1 Generalized Linear Models

The model for the mean annual temperature shown in eq. (2.1) is relatively simple in that it does not incorporate any other information that could have been measured simultaneously as Y. For example, it is well known that mean annual temperature

[1] This desideratum could be no longer true in the age of Neural Networks.

TABLE 2.1: Some of the exponential family distributions.

Distribution	Type	Typical link function
Gaussian	continuous	identity
Gamma	continuous	negative inverse
Exponential	continuous	negative inverse
Poisson	discrete	log
Binomial	discrete	logit
Bernoulli	discrete	logit

has been increasing due to climate change, which is information that is not accounted for in the simple model above. This additional information is typically denoted as the **predictors** x_i. Incorporating additional information in the form of the predictor matrix X can be done straightforwardly using **Generalized Linear Models** (GLMs).

For a simple extension of the Gaussian model, we assume that the mean parameter μ is now a linear function of some measured predictor x_i. In that case, we can rewrite the model of eq. (2.1) as:

$$y_i \sim \mathrm{N}(\mu_i, \sigma) \tag{2.3}$$

$$\mu_i = \alpha + \beta x_i. \tag{2.4}$$

Note that this model has the additional parameters α and β. This kind of model, where the location parameter of the distribution is modeled using a linear combination of predictor variables, is known as a Generalized Linear Model.

The model in eq. (2.4) is more flexible than the model of eq. (2.1), as it can incorporate the information from the predictor x_i. The price to pay is the additional parameters that need to be estimated, but in return, new predictions can be made from previously unobserved predictors.

The GLM formulated in eq. (2.4) is one of the most widely used models. However, the outcome Y does not necessarily need to follow a Gaussian distribution. In fact, any distribution from the exponential family can be used to generate a GLM. For example, assume that Y is now an indicator variable that there was rain or no rain for a particular day. This time y represents the count outcome of rainy days within n days. In this case, the best model is no longer Gaussian, but instead the binomial distribution. For the binomial distribution, a possible GLM is then given by

$$\begin{aligned} y_i &\sim \mathrm{Binomial}(n, p_i) \\ f(p_i) &= \alpha + \beta x_i. \end{aligned}$$

For this model, the parameter to be estimated is the probability p of having a rainy day. Because p is a probability, it must be a number between 0 and 1. However, the linear function $\alpha + \beta x_i$ in the second line can return values outside this range. To ensure that the resulting p is inside the proper range, the **link function** $f(\cdot)$ is introduced in the model. For the GLM depicted above, a widely used link function is the logit function.

To summarize, GLMs have the following form:

$$y_i \sim \mathrm{EF}(\mu_i, \phi) \tag{2.5}$$

$$f(\mu_i) = \beta_0 + \sum_{i=1}^{k} \beta_i x_i, \tag{2.6}$$

where EF denotes some exponential family distribution with mean μ_i and other (scale, shape, etc.) parameters ϕ and $f(\cdot)$ is the link function. Some members of the exponential family distributions and their typical link functions can be found in table 2.1.

The formulation for GLMs given in eq. 2.6 is an incredibly powerful one. However, GLMs are not applicable when one desires to use a model not from the exponential family or when one wants to model a parameter other than the location parameter. To achieve this, Yee (2015) introduced an extension known as **Vector Generalized Linear Models** (VGLMs). VGLMs allow the use of non-exponential distributions like the GEV distribution introduced in the next chapter; furthermore, they allow the analyst to model not only the location parameter, but all of the parameters simultaneously. VGLMs will be one of the main tools for modeling used in the studies described in this thesis.

This section has attempted to provide a summary of statistical models, which are mathematical constructs that explain the variability of the data. These models always include parameters, which are quantities that control how the model assigns probabilities to different outcomes. Therefore, the task of statistical modeling can be seen as two-pronged: first, one needs to identify a proper model to describe the data; then, appropriate values for the parameters should be selected. Information about the data is then contained inside the parameters.

2.3 Statistical Inference

Once an adequate mathematical model to describe the overall random behavior of some experiment has been found (e.g., a certain distribution or a stochastic process), the next step is to arrange the properties of the model in a specific way that it is a good description of the data we have observed. As indicated above, the properties that control statistical models are known as the model parameters, typically denoted by the letter $\boldsymbol{\theta}$. Different models contain different amounts and types of parameters; for example, the Gaussian distribution $\mathrm{N}(\boldsymbol{\theta})$ is described by the two parameters $\theta = (\mu, \sigma)$, where μ is the location and σ is the scale parameter. In contrast, the exponential distribution Exponential(θ) contains only the parameter $\theta = \lambda$. Once a model has been selected, the task of statistical modeling is reduced to finding adequate values of $\boldsymbol{\theta}$ so that the model not only properly describes the sample we have observed but also gives us information about the population that the sample stems from.

Statistical inference is the process of finding values for the model parameters $\boldsymbol{\theta}$ using the observed sample in such a way that the resulting model is considered to be adequate[2] description for the population of the sample. The result of the inference process is some estimate of the parameters, either as a single number (i.e., a point estimate, denoted by $\hat{\theta}$) or as a probability distribution.

The methods for statistical inference can be divided into several paradigms. For this work, we describe the two most common ones: the frequentist and the Bayesian paradigm. The main difference between these two approaches is in how uncertainty

[2] The exact definition of what is adequate depends on the technique used to do the inference.

is handled, and in a more philosophical context, in the meaning of probability. In practical terms, both approaches commonly give similar results, but their interpretation is profoundly different. The following sections describe the background of both approaches and the primary methods for inference that exist for each approach.

2.3.1 Frequentist inference

Frequentist inference is based on the **frequentist** interpretation of probability, which defines the probability of an event as the limit (or long-run) of its relative frequency after performing many trials. As this value is utterly devoid of opinions, it is known as the objective interpretation of probability. Most of the ideas behind frequentist inference were developed in the early 20th century and constitute most of the methods used for statistical analyses today.

Inference for frequentist methods can be divided into parametric or nonparametric methods. Parametric methods assume that the model for the data contains a finite number of parameters $\boldsymbol{\theta}$ for a well-known parametric probability distribution. Nonparametric methods are used when the stochastic process cannot be described with either a finite number of parameters or with a well-known parametric distribution. For this thesis, only parametric methods will be considered.

A cornerstone idea within the frequentist paradigm is that the parameters $\boldsymbol{\theta}$ are seen as fixed states of nature. Therefore, the parameters are unknown quantities, with no room for uncertainty: we either know their true value or we do not; parameters are inherent properties of a population, from which we commonly can only take samples. The samples are used to estimate the true value of the parameters. This estimate is commonly denoted as $\hat{\theta}$.

This idea gives way to the sampling distribution. When dealing with different samples of the same population, it is almost always the case that each sample will show some variability from the other. Thus, estimates or statistics computed from these different samples will also show some variability. The so-called **sampling distribution** describes this variability between estimates of different samples. To construct this distribution, one must first define the "cloud" of all possible sample outcomes, which is a function of the assumed model, the sampling procedure employed, and the given sample size.

The sampling distribution provides a probability model that describes the relative frequencies for each possible value within the cloud of possible outcomes given fixed values of the parameters $\boldsymbol{\theta}$. The sampling distribution can be denoted as $p(D_{\theta,I}|\theta, I)$, where $D_{\theta,I}$ is the data within the cloud of outcomes that *should* be observed (assuming that the chosen underlying model is true), θ is the different parameter values and I is the stopping and testing intention. For the same experiment, I could differ between analysts, so that the resulting cloud of possible outcomes would be different.

In summary, the sampling distribution describes the probabilities of possible data if we run an experiment many times given a particular model with fixed parameter values. It does not, however, give information about the probability of the parameter values given the data (that is, it does not provide information about $p(\theta|D)$). Within the frequentist framework, the sampling distribution can be used to estimate parameter values, to obtain inference about the uncertainty of estimated parameter values or statistics (e.g., confidence intervals) and to perform tests to assess if certain parameter values can be "rejected" because they are deemed too improbable within the cloud of all possible outcomes (e.g., null hypothesis significance testing).

Maximum Likelihood Estimation

We now come back to the original question of frequentist parametric inference:

- Given a model with parameters $\boldsymbol{\theta}$ and some data D, what are reasonable guesses for the values $\hat{\theta}$ that are consistent with the assumed data generating process of D?

Within the frequentist paradigm, several approaches exist to answer this question. These include the method of moments, the least squares method, and maximum likelihood estimation (MLE). The latter one stands out in terms of its practical utility and theoretical properties, making it particularly useful for extremes.

In a nutshell, **maximum likelihood estimation** methods find the values of the parameters $\boldsymbol{\theta}$ that maximize the probability of having observed D. We know that different values of $\boldsymbol{\theta}$ lead to differences in the probability assigned from the model to each possible outcome. Therefore, a reasonable idea is to find parameter values for which the observed data D has a relatively high probability. We assume that parameter values that make the observed D highly probable are likely more correct than parameter values that make the observed D improbable. This assumption is the main idea behind MLE methods, which holds exceptionally well for many situations.

The probability of having observed D given a specific model with parameters $\boldsymbol{\theta}$ is given by the **likelihood function** $p(D|\boldsymbol{\theta})$. For n i.i.d. data points, the likelihood function can be written as:

$$L(\boldsymbol{\theta}|D) = p(D|\boldsymbol{\theta}) = \prod_{i=1}^{n} p(D_i|\boldsymbol{\theta}). \tag{2.7}$$

In this case, $L(\boldsymbol{\theta}|D)$ is a function of the parameters while the data is fixed, and as such, it does not constitute a probability density function over the parameter values. Therefore, the likelihood function $L(\boldsymbol{\theta}|D)$ should not be confused with $p(\boldsymbol{\theta}|D)$. To obtain this latter term, the probability of different parameter values given the observations, we require an application of Bayes' rule, as explained in the next section.

To get the maximum likelihood estimator $\hat{\boldsymbol{\theta}}$, one needs to find the parameter values that maximize $L(\boldsymbol{\theta} \mid D)$. However, from eq. (2.7), we can see that this would imply working with the product of many potentially minuscule quantities, which can lead to numerical instability. Therefore, it is more common to work with the **log-likelihood** given by

$$l(\boldsymbol{\theta} \mid D) = \log(L(\boldsymbol{\theta} \mid D)) = \sum_{i=1}^{n} \log(p(D_i \mid \boldsymbol{\theta})). \tag{2.8}$$

Because the log-likelihood uses sums instead of products, it becomes much easier to handle. The logarithm is a monotonically increasing function, so that the maximum value of this function occurs at the same point as the original likelihood function.

Given the log-likelihood, the maximum likelihood estimate of $\boldsymbol{\theta}$ is then given by

$$\hat{\boldsymbol{\theta}} = \arg\max_{\theta} l(\boldsymbol{\theta} \mid D). \tag{2.9}$$

Figure 2.1 shows an example of finding the value of the θ parameter for a Bernoulli distribution. This is a classical optimization problem, and as such, many different optimization algorithms exist to solve it. Some examples include the BFGS algorithm (Broyden, Fletcher, Goldfarb, and Shanno, 1970), the CG algorithm (Fletcher and Reeves, 1964), and the L-BGFS-B algorithm of Byrd et al. (1995). These algorithms can be easily implemented using the `optim` function in `R`.

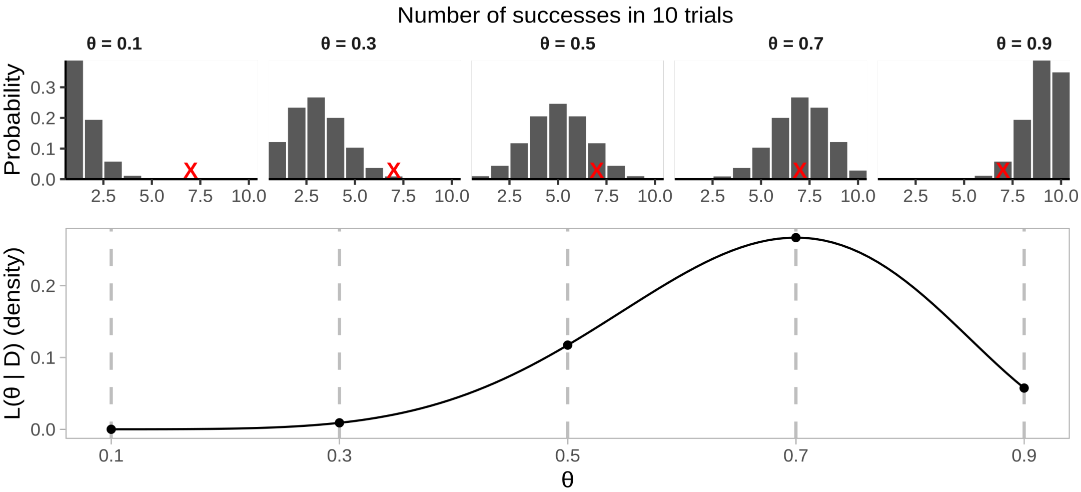

FIGURE 2.1: Graphical depiction of how MLE methods work. The plots in the top row show the pmf of the Bernoulli distribution for different values of θ, with the red cross showing the actual observed number of successes for 10 trials. The bottom plot shows the likelihood function for the Bernoulli distribution. MLE finds the value of θ for which the likelihood is maximum, which in this case is $\theta = 0.7$.

The MLE estimate $\hat{\boldsymbol{\theta}}$ is not only intuitive, but it also possesses some nice theoretical properties. When dealing with the sampling distribution of $\hat{\boldsymbol{\theta}}$ (known as the estimator), we would expect to have different values of the estimated values for each sample. Under some mild regularity conditions, it can be shown that the MLE estimator is asymptotically unbiased (i.e., $E[\hat{\boldsymbol{\theta}}] = \boldsymbol{\theta}$), and additionally, that its variance can be modeled with

$$\hat{\boldsymbol{\theta}} \sim N(\boldsymbol{\theta}, \boldsymbol{\mathcal{I}}^{-1}), \tag{2.10}$$

where $\boldsymbol{\mathcal{I}}$ denotes the **Fisher information matrix**. Furthermore, it can be shown that MLE estimators are usually consistent, that is, that as the sample size tends to infinity, $\hat{\boldsymbol{\theta}} \to \boldsymbol{\theta}$. These theoretical properties have made MLE estimation a reliable and popular method for the estimation of parameter values in recent years.

2.3.2 Bayesian inference

Bayesian inference (or statistics) is a branch of statistics that assigns probabilities or distributions to events and parameters based on previous knowledge before experimentation and data collection and that applies Bayes' Theorem to revise the probabilities and distributions after obtaining experimental data. Under this paradigm, probability is considered to represent the degree of belief that a certain outcome from a pool of possible outcomes will occur. This interpretation is also known as the subjective interpretation of probability, as different people can have different degrees of belief for the same outcome. These degrees of belief are mathematically equal to frequentist probabilities, as they share the same axioms.

From the subjective probability point of view, Bayesian inference amounts to “the reallocation of credibility across a space of candidate possibilities” (Kruschke, 2014). This idea can be illustrated briefly by the following example: Imagine that a person comes outside to find that the ground is wet. They want to know the most probable explanation for the ground being wet. Many possible explanations could reasonably explain this: it rained earlier, someone was washing their car nearby, a water pipe burst, someone dropped a water cup, etc. Initially, the person only knows that some part of the ground is wet, and without extra information, they can only assign different degrees of belief to each possibility based on their previous knowledge and experience. Crucially, however, the person can make additional observations to reallocate some credibility: for example, if they observe that the sky is very cloudy and everything around them is wet, their belief that it rained will increase. This increased belief comes at the expense of lowering belief in other possibilities, like that someone was washing their car. On the other hand, if the person observes that only the ground around a shiny-looking car is wet, their belief that someone washed their car will increase, while their belief that it rained will decrease.

The process described in the example above of [previous belief - observation - updated belief] can be repeated ad infinitum until only one of the possibilities has a significant degree of belief assigned to it. This idea is the basis behind Bayesian inference[3]. The operation described above can be summarized in the expression known as **Bayes' Rule**:

$$Pr(A \mid B) = \frac{Pr(B \mid A)Pr(A)}{Pr(B)}, \tag{2.11}$$

where A and B are two events from the same sample space.

Bayes' Rule as formulated in 2.11 is an incredibly powerful tool for many applications of discrete problems, like estimating the chance someone has a disease given a positive test result. However, it is not immediately clear how to adapt it to use for estimating parameter values that are essentially continuous, as they can take virtually any value within some interval. When dealing with statistical models, the idea of Bayesian inference is to reallocate belief toward the parameter values $\boldsymbol{\theta}$ that are consistent with the data and away from parameter values that are inconsistent with the data. This requires the expression given in 2.11 to be transformed to

$$p(\theta|D) = \frac{p(D \mid \theta)p(\theta)}{p(D)}, \tag{2.12}$$

which replaces probabilities for probability density, and where D represents the observed data and $\boldsymbol{\theta}$ the model parameters.

Each term of eq. (2.12) has its own name and explanation. On the left-hand side of eq. (2.12), $p(\boldsymbol{\theta}|D)$ represents the probability assigned to each parameter value given the data (i.e., how much belief we assign to each parameter value after seeing the data). As this quantity is derived *after* seeing the data, it is commonly known as the **posterior distribution**. The posterior distribution represents what we typically want to know from the application of a statistical model. On the right-hand side, the term $p(D \mid \boldsymbol{\theta})$ represents the probability of particular data values given the model's structure and parameter values. This term is known as the **likelihood**, mathematically identical to the frequentist one described by eq. (2.7). The additional term $p(\boldsymbol{\theta})$ in the right-hand numerator represents the probability distribution for the model parameters *before* seeing the data, and its typically known as the **prior distribution**. The prior distribution encodes the previous knowledge about the problem, and it is central to Bayesian inference. Finally, the right-hand denominator $p(D)$, sometimes known as the **marginal likelihood**, is the overall probability of the data according to the model, determined by averaging across all possible parameter values weighted by the probability of each parameter value. The marginal likelihood is merely a numerical constant that ensures that the integral of the resulting posterior is equal to one.

[3] Indeed, Sherlock Holmes was applying Bayesian inference when he famously claimed "When you have eliminated the impossible, all that remains, no matter how improbable, must be the truth" (Doyle 1890, chap. 6).

For the continuous case the marginal likelihood is calculated by $p(D) = \int p(D \mid \boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}$. Therefore, eq. (2.12) can be rewritten in terms of the likelihood and prior as:

$$p(\boldsymbol{\theta}|D) = \frac{p(D \mid \boldsymbol{\theta})p(\boldsymbol{\theta})}{\int p(D \mid \boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}}. \tag{2.13}$$

This latter equation is the most commonly used form of Bayes' rule for performing parameter estimation.

Steps of Bayesian inference

Bayesian inference involves using eq. (2.13) to find the posterior distribution of parameter values given the observations. In a nutshell, Bayesian inference requires the following steps:

1. Define an appropriate model that describes the variability of a phenomenon that can be seen as a stochastic process.

2. Identify the parameters $\boldsymbol{\theta}$ that characterize the model, and propose the prior distribution $p(\boldsymbol{\theta})$ that assigns a probability to each parameter value before seeing the data (i.e., based on previous knowledge).

3. Recollect the observed data D and calculate the probability of having seen the particular values of D with parameter values θ, that is, the likelihood $p(D \mid \boldsymbol{\theta})$.

4. Use Bayes' theorem (Eq. (2.12)) to combine the likelihood and the prior to obtain the posterior probability distribution of $\boldsymbol{\theta}$, given the data ($p(\boldsymbol{\theta}|D)$).

5. Use the posterior distribution to make inferences about the most probable values of $\boldsymbol{\theta}$, as desired.

Note that for step 4, it is necessary to deal with the integral in the denominator of eq. (2.13). This can be very challenging, as most of these integrals do not have analytical solutions. An initial alternative could be to use numerical approximations; however, from eq. (2.13), we can see that this integral is performed in the joint parameter space, which involves the combination of *all* parameter values. Within most real-world applications, the number of parameters $\boldsymbol{\theta}$ is in the order of dozens to hundreds, meaning that the resulting combination has an enormous number of dimensions. Therefore, most, if not all, numerical approximations are infeasible to solving this integral for many real-world applications.

Historically, the computational challenge posed by the integral in eq. (2.13) limited the use of Bayesian inference to a small group of well-known problems where the integral could be simplified. In recent years, however, several methods have been developed to avoid computing this integral by instead using a large number of samples that contain representative combinations of the posterior distribution. These developments have brought over an explosion in the applications of Bayesian inference, cementing it as a powerful and modern tool for statistical inference today. The work done in this dissertation uses one of the most commonly used sampling methods, known as Markov Chain Monte Carlo (MCMC). MCMC methods are incredibly powerful, but have a high computational cost. They will now be described in detail.

Bayesian inference using MCMC

As mentioned above, the main challenge in applying Bayesian inference to real-world applications is dealing with the integral in the denominator of eq. (2.13), which is

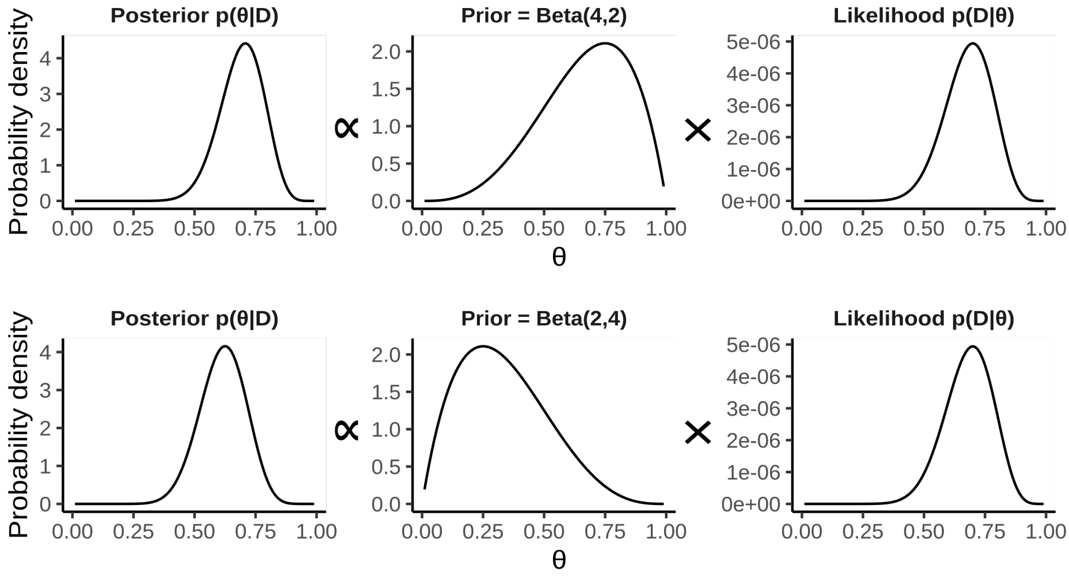

FIGURE 2.2: Change of the posterior according to the used prior. The top row uses a prior whose density is concentrated towards larger values of θ, while the bottom row uses a prior whose density is concentrated towards lower values. The resulting posterior is then shifted towards different locations. Notice that the likelihood used for both examples is the same: a Bernoulli distribution for 7 out of 10 observed successes.

usually intractable. However, it turns out that this integral is merely a multiplicative constant of proportionality so that we can rewrite Bayes' rule as

$$\begin{aligned} p(\theta|D) &\propto p(\boldsymbol{\theta}, D) \\ &\propto p(D \mid \boldsymbol{\theta})p(\boldsymbol{\theta}). \end{aligned}$$

In other words: the posterior is proportional to the likelihood times the prior. Figure 2.2 shows how two different prior distributions can change the resulting posterior. In order to approximate the posterior, we would only need to find a way to sample from $p(\boldsymbol{\theta}, D)$ with the observed data values plugged in for D. Once enough samples are taken, the resulting histogram can be used to approximate the distribution of the posterior. The problem then reduces to finding a way to sample effectively from $p(D \mid \boldsymbol{\theta})p(\boldsymbol{\theta})$.

Markov Chain Monte Carlo (MCMC) methods are a general family of sampling schemes for arbitrary distributions known up to a multiplicative constant. In principle, MCMC methods only require that the expression for $p(D \mid \boldsymbol{\theta})$ and $p(\boldsymbol{\theta})$ can be computed for any specified values of D and $\boldsymbol{\theta}$. MCMC methods then return an approximation of $p(\theta|D)$ in the form of a large number of values from $\boldsymbol{\theta}$ sampled from that distribution. These samples can then be used to get several descriptive statistics of $\boldsymbol{\theta}$: for example, the values with the largest probability density (mode) or intervals of specific probability values (credibility intervals).

Under the hood, MCMC algorithms search for a Markov Chain with a stationary distribution equal to the posterior distribution in eq. (2.13). Let $\{\boldsymbol{X_1, X_2, ...}\}$ be a sequence of random vectors. This sequence constitutes a first-order **Markov chain** if, for any j,

$$p(\boldsymbol{x}_j \mid \boldsymbol{x}_{j-1}, \boldsymbol{x}_{j-2}, ..., \boldsymbol{x}_1) = p(\boldsymbol{x}_j \mid \boldsymbol{x}_{j-1}). \tag{2.14}$$

In short, this means that the probability of $\boldsymbol{X}$ having a certain value in the current

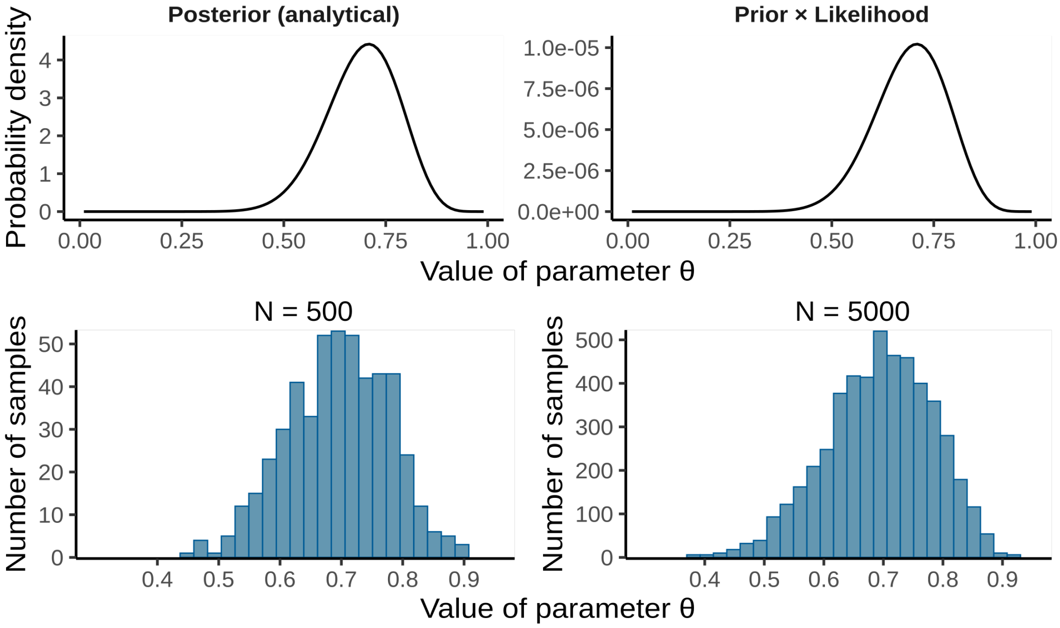

FIGURE 2.3: Top row: Comparison of the posterior distribution for the example shown in Fig. 2.2 obtained analytically (left) with the distribution obtained by multiplying the likelihood with the prior (right). Notice that the shape of both distributions is identical, but that the y-axis changes. The bottom row shows an approximation to the posterior using samples from a MCMC sampling scheme. The left plot uses 500 samples, and the right plot 5000 samples.

state j depends only on the value it had at the previous step $\boldsymbol{x}_{j-1}$. Markov chains can be expressed in terms of the transition probabilities from one state to the next, usually denoted as the transition matrix $\boldsymbol{P}$, whose individual components contain transitions such as $p(\boldsymbol{x}_{j-1} \rightarrow \boldsymbol{x}_j)$. Under certain conditions of irreducibility, the transition probabilities after advancing many steps (i.e., the long-run probabilities) will converge to a stationary distribution $\pi(\boldsymbol{x}_j)$, such that

$$\pi(\boldsymbol{x}_j) = \pi \boldsymbol{P}. \tag{2.15}$$

The idea of MCMC methods is then to get enough samples from a Markov Chain whose stationary distribution π is equal to $p(\boldsymbol{\theta} \mid D)$, where we know this posterior up to a proportionality constant ($p(\boldsymbol{\theta} \mid D) \propto p(D \mid \boldsymbol{\theta})p(\boldsymbol{\theta})$). This requires that the constructed Markov Chain is irreducible, meaning that regardless of the starting position there is a positive probability of visiting all possible values of $\boldsymbol{X}$. Additionally, MCMC methods require that the Markov chain is recurrent, meaning that in the long-run the chain will visit any non-negligible set of values an infinite number of times. For an irreducible and recurrent chain, a good approximation of the posterior can be obtained if enough samples are recorded. The bottom row of Figure 2.3 shows how the resulting histogram of the samples converges to the posterior as the number of samples grows. However, it is not trivial to determine how many samples are enough to ensure that the resulting distribution has converged to the desired one.

Theoretically, one needs only to construct an irreducible and recurrent Markov Chain with the correct stationary distribution to get samples from the posterior distribution. The question is now how exactly to construct such a chain. To design such a Markov chain, different algorithms exist. One of the most simple but effective ones is the **Metropolis-Hastings** algorithm (Hastings, 1970), a special case of the

Metropolis algorithm developed by Metropolis et al. (1953). The Metropolis-Hastings algorithm has the following steps:

1. Pick a starting position of $\boldsymbol{\theta}_{j-1}$ such that $p(\boldsymbol{\theta}_{j-1}|D)$ is non-zero.
2. Create a **proposal distribution** $p(\boldsymbol{\theta}_j \mid \boldsymbol{\theta}_{j-1})$ from which a candidate value $\boldsymbol{\theta}_{proposed}$ can be formulated.
3. Generate a value for $\boldsymbol{\theta}_{proposed}$.
4. Compute the probability p_{move} to accept or reject $\boldsymbol{\theta}_{proposed}$ according to the expression
$$p_{move} = \min\left(\frac{p(D \mid \boldsymbol{\theta}_{proposed})p(\boldsymbol{\theta}_{proposed})}{p(D \mid \boldsymbol{\theta}_{current})p(\boldsymbol{\theta}_{current})}, 1\right).$$
5. Accept or reject the proposed move by sampling from a uniform distribution over $[0, 1]$. If the sampled value is between 0 and p_{move}, accept the proposal by setting $\boldsymbol{\theta}_j = \boldsymbol{\theta}_{proposed}$. Otherwise, reject the proposal and set $\boldsymbol{\theta}_j = \boldsymbol{\theta}_{j-1}$
6. Increment j and iterate from step 2. Repeat until a sufficient number of samples have been reached.

The key to the Metropolis-Hastings algorithm is in how to choose the proposal distribution of step 2. Metropolis-Hastings uses a multivariate Gaussian distribution centered on the current value of θ with a standard deviation set by the user. While this algorithm ensures that the resulting Markov chain will converge to the posterior distribution, it can take many samples before doing so. Therefore, different algorithms (e.g., Gibbs sampling) that use more clever proposal distributions have been developed in recent decades. These algorithms are more efficient when sampling, requiring fewer samples to reach convergence of the posterior.

In order to use Metropolis-Hastings (or any other MCMC sampling scheme, for that matter), the following conditions must be met:

- We must be able to generate random values from the proposal distribution to obtain as many $\boldsymbol{\theta}_{proposed}$ as needed.
- Both $p(D \mid \boldsymbol{\theta})$ and $p(\boldsymbol{\theta})$ must be able to be computed for any valid value of $\boldsymbol{\theta}_{proposed}$.
- We must be able to sample a random value from a uniform distribution to decide whether to accept or reject p_{move}.

At this point, the reader may ask why all these steps are needed to approximate the posterior if we already know its shape and other properties from multiplying the likelihood times the prior (e.g., Figure 2.2). The answer is that real-world applications typically contain dozens, if not hundreds, of parameters, in which case the posterior is a high-dimensional distribution. In this case, obtaining information about a single parameter would require marginalizing (i.e., integrating) over the entire distribution, which requires that the integral in eq. (2.13) be solved anyway. The great advantage of MCMC methods is that they operate under any number of dimensions, so obtaining information about the posterior of any given parameter requires only the histogram of the resulting MCMC samples for a such parameter.

Hamiltonian Monte Carlo

Of the different MCMC algorithms that exist today, one of the most efficient ones is the so-called **Hamiltonian Monte Carlo** (HMC) algorithm, which has been coded for general use as part of the software Stan. HMC is based on the Metropolis algorithm, but the way it chooses the proposal distribution is very different from the other algorithms, such as Gibbs sampling or Metropolis-Hastings. For HMC, instead of using a fixed proposal distribution with a symmetrical shape for each step, the distribution adapts according to the current position of the parameter within the distribution. This adaptation is made in such a way as to maximize the chance of accepting the following proposal so that HMC typically requires much fewer steps than other algorithms to converge to the posterior. This adaptation, however, comes at a greater computational cost.

The basic idea for choosing a proposal distribution in HMC is first to use the negative logarithm of the posterior to create a "surface." Then, random momentum is given to a marble in a certain initial position, which then "rolls" on the surface for some predetermined number of steps. The position where the marble ends up is recorded, and the whole procedure is repeated many times until a distribution of end positions for the marble is obtained. This is the new proposal distribution from which the proposal move is taken. The trajectory of the marble is computed using the gradient of the likelihood with the Hamiltonian equations of motion (thus the "Hamiltonian" part of the name). Since the marble typically moves towards downhill positions in the negative logarithm of the posterior (i.e., the marble moves towards larger density values of the posterior), the proposals are almost always accepted, making this method more efficient than other Metropolis algorithms.

When designing a particular HMC algorithm, several choices have to be made. First of all, the initial momentum given to the marble has to be carefully chosen. Typically, this momentum comes from a zero-centered Gaussian distribution with a specified standard deviation. As the standard deviation grows, the proposal distribution grows with it. Thus, finding a standard deviation that is neither too small nor too large is necessary to explore all the space in the distribution properly. The second and third choices are the total number of steps per proposal and the length of these steps in discrete space. These two parameters can tune the HMC sampling to have a proper acceptance rate, typically desired to be 65%. Stan incorporates several adaptative steps to choose the values of these parameters correctly. To do this, a so-called *warmup* phase is necessary, for which the resulting samples are not representative of the posterior but instead used to choose the parameters of the algorithm. The warmup samples are always discarded and typically make up 20-40% of the total number of samples used for any particular MCMC sampling using Stan.

In order to use HMC sampling with Stan, in addition to the conditions needed to use MCMC, the following conditions must be met:

- The product of the prior times the likelihood must have an analytically-derived gradient at any value of $\boldsymbol{\theta}$, computed by Stan using autodiff.
- The posterior must be continuous, as the gradient for discrete distributions cannot be computed.

MCMC sampling was applied exclusively using HMC with Stan for all Bayesian studies in this dissertation. This implementation was accomplished by using the R-plugin of Stan, `rstan` (Stan Development Team, n.d.), which offers a convenient and easy-to-use interface between both languages. In addition, the package `brms` was used for the distributional models described in sec. 5.2.3. The brms package is a convenient extension of rstan to automatize the use of Bayesian inference for linear models.

MCMC diagnostics

Once an MCMC algorithm has been designed to sample from the posterior of a particular study, an important step is to ensure that the number of samples was enough for the resulting distribution to have converged to the posterior. Remember that MCMC methods ensure that the distribution eventually converges to the posterior, but they say nothing about how fast this will happen. Fortunately, several diagnostics exist to check when the samples have *not* yet converged properly (there are, however, no diagnostics to know if the distribution has already converged). A part of any modern Bayesian study that uses MCMC sampling always includes checking and reporting the value of these diagnostics to discard any severe issues with the samples.

Most MCMC diagnostics are based on a method similar to ensemble forecasting: Several **chains** are initialized for sampling the same posterior but using different initial values. The different chains are then run until the required number of samples (including warmup) are obtained. Theoretically, if a sufficient number of samples is reached for each chain, every chain should arrive at the same posterior distribution. However, when the different chains show different posterior distributions, it is a sign that the samples have not yet converged to the posterior and that more samples are needed. Therefore, two different diagnostics are used to determine how similar the chains are: The trace plot and the $\hat{R}$ diagnostic.

The **trace plot** is a graphical representation of every step in sequential order from every chain in the MCMC sampling. An example of a trace plot is shown in Fig. 2.4. Good chains should show three properties in the trace plot: (i) stationarity, (ii) good mixing, and (iii) convergence. Stationarity means that the path of each chain stays more or less in the same high probability region of the posterior (that is, the chain should not wander too much around the parameter space). Good mixing occurs when the chain rapidly explores the entire region of the posterior: this can be seen as each chain having a zig-zag path centered around a particular median value. Finally, convergence means that the different chains are all concentrated in the same high-probability region of the posterior; this is seen in the trace plot when the chains are superimposed on top of each other, looking like a “fuzzy caterpillar.” The trace plots in the left column of Fig. 2.4 are an example of trace plots for good chains, while those in the right are examples of bad chains.

A numerical complement to trace plots is a numerical diagnostic known as the $\hat{R}$ convergence diagnostic (pronounced “er-hat”), also known as the **Gelman-Rubin statistic** (Gelman et al., 1992; Vehtari et al., 2021). The $\hat{R}$ measures how much variance there is between chains relative to how much variance is within chains. The idea is that after convergence, the chains should show the same between-chain variance as within-chain variance, as the samples supposedly come from the same distribution. When a chain is stuck in a different region of the posterior, the between-chain variance increases its value relative to the within-chain variance. For chains that have converged, $\hat{R} \in [1, 1.1]$. Values of $\hat{R}$ greater than 1.1 typically signal a grave issue with the MCMC samples and should always be suspected before using for the analysis.

The final standard diagnostic for MCMC chains is the **Effective Sample Size** (ESS). Typically, samples from MCMC show a high degree of autocorrelation, due to the underlying Markov Chain used to construct the samples. The problem is that autocorrelation decreases the new information given by each new sample: samples with really high autocorrelation will contain very little new information about the posterior.

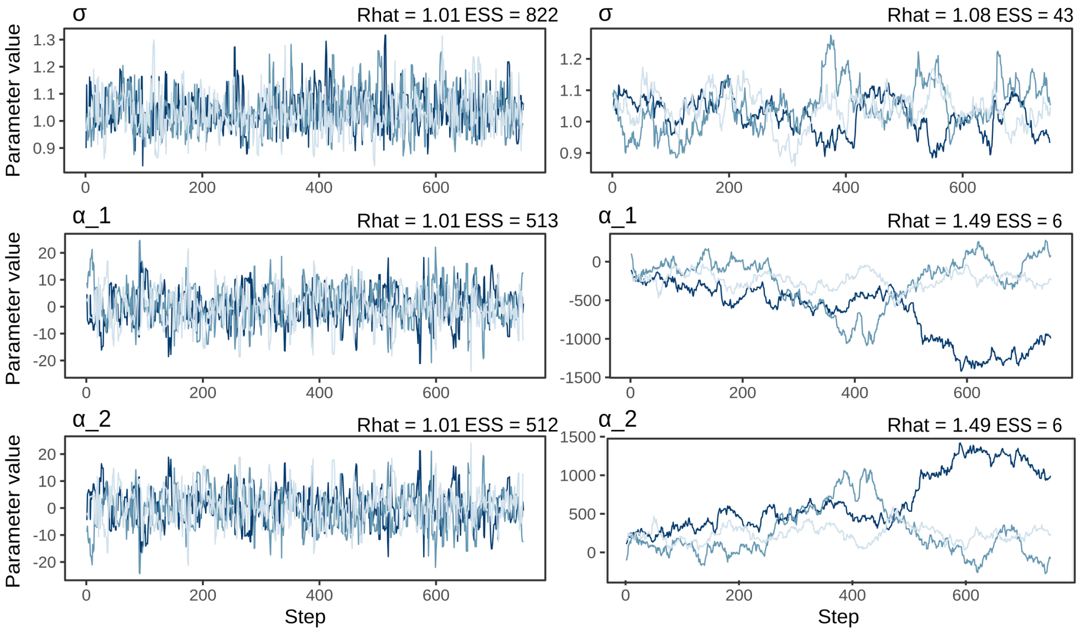

FIGURE 2.4: Standard diagnostics for evaluating MCMC samples. The left column shows samples from a 3-parameter model that shows good signs of stationarity, proper mixing, and convergence. The right column shows samples from the same model but with chains that show poor diagnostics. The warmup period samples were not included in the plots.

The ESS was created to get an idea of how much information is in the samples; it is computed by dividing the actual sample size by the amount of autocorrelation. This number is always lower than the total number of samples. The "proper" number of ESS depends on the goal of the analysis: If the goal of the study is only to know the posterior's median, relatively low values of ESS are acceptable. If, however, the goal is to estimate the posterior's tails to create 95% credible intervals, a large ESS is needed. A commonly used rule of thumb is that when estimating 95% credible intervals, an ESS of 10,000 is required. Note that low sample sizes are typically not an issue when using Stan, as HMC methods are typically much less autocorrelated than other MCMC schemes.

To summarize, the common heuristic to check that the resulting MCMC samples of the posterior distribution are representative is the following:

1. Run the MCMC sampling for 3 or 4 chains.

2. Check the trace plot for stationarity, good mixing, and convergence.

3. Check that the value of $\hat{R} < 1.1$.

4. Check that the ESS is sufficiently large, based on the goal of the analysis.

Finally, Stan contains a final diagnostic not present in the other MCMC schemes based on the conservation of energy when solving the Hamiltonian equations of motion. When the sampling is done in a very challenging/degenerate region of the posterior, the energy will sometimes not be conserved. This is known in Stan as a **Divergent transition** and will be reported in the final results of the sampling. Posterior distributions that showed divergent transitions should be handled with care, as the samples could have skipped parts of the posterior that should have been sampled. Strategies to deal with divergent transitions can be found in Stan Development Team (2022).

2.4 Uncertainty estimation

The statistical models described in the last section combine observations with randomness to gain knowledge about a particular process. This newfound knowledge, however, comes at a cost: The random component used in the model is inherently uncertain, which, in turn, makes every component of the model, from the parameters to the predictions, uncertain. This uncertainty does not mean that statistical models are useless for real-world applications. On the contrary, "all models are wrong, but some are useful" (George Box) Therefore, a crucial step in the modeling process is to get an idea of the degree of uncertainty from the model. Then, we can judge how valuable a particular model can be based on the particular application.

In specific terms, uncertainty in statistical models typically consists of three parts:

- Uncertainty in the measuring of the observations,
- uncertainty in the value of the model's parameters, and
- uncertainty in the model's predictions.

The first source of uncertainty, associated with the measurements, is a relevant but often ignored[4] source of uncertainty in the models. Ignoring the uncertainty of the measurements is common due to the complexity of incorporating such uncertainty in the models, although Bayesian inference has relatively straightforward ways of doing so. The next source of uncertainty, that of the estimated values for the parameters $\boldsymbol{\theta}$, is a result of using (inevitably) incomplete information when performing inference. Lots of observations for a model typically result in less uncertainty for the estimated parameters, but the information will never be perfect, always leading to some uncertainty remaining. Moreover, in the case of Bayesian inference, previous knowledge can also add (or subtract) uncertainty to the estimated parameters, depending on the used prior. Finally, the last source, the uncertainty of the prediction made from the models, stems from the "propagation" of uncertainty from the model parameters to the model predictions. In this section, the last two sources of uncertainty are explored, as well as some typical ways of presenting them.

Uncertainty is an inevitable part of statistical models, but what exactly *is* uncertainty? This question has (surprisingly) many different answers, highly related to the respective analyst's definition of probability. The two most relevant interpretations of uncertainty are the frequentist and the Bayesian interpretations of uncertainty. For the frequentist approach, the true value of $\boldsymbol{\theta}$ is *fixed*, and therefore, it has no uncertainty. Instead, what is uncertain is the variability of the estimator that comes from using different samples from the same population (as depicted by the sampling distribution). Under this paradigm, uncertainty cannot be expressed for one single event; instead, it exists for a series of (imaginary) repeated experiments. In contrast, for the Bayesian approach there is no true value of $\boldsymbol{\theta}$, so the uncertainty represents the different degrees of belief that the analyst has on every specific value of the parameters. Each approach has a different way of conveying uncertainty as a mathematical expression, which we explore below.

2.4.1 Representing uncertainty: CIs

The uncertainty about the value of the model parameters or its predictions is typically denoted using numerical intervals around some pointwise estimate. These intervals

[4] Alas, as it happens, we also ignore the uncertainty of the measurements in this thesis.

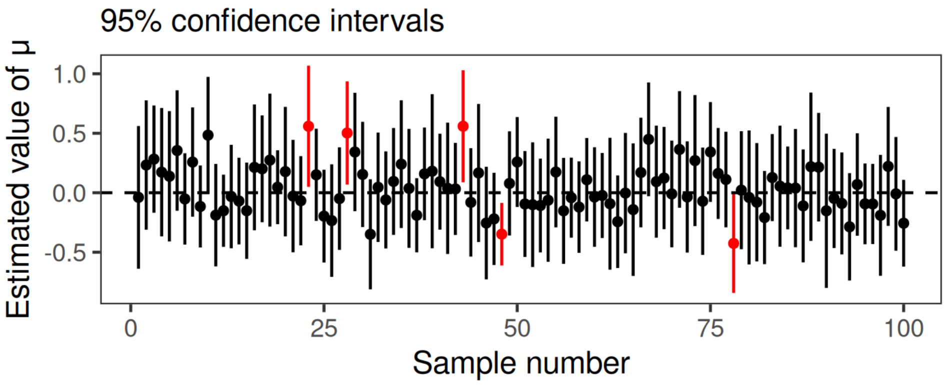

FIGURE 2.5: 95% Confidence intervals constructed from 100 simulated samples for the estimate of the location parameter μ of a Gumbel distribution. The dashed line shows the true value of the Gumbel distribution used to get the samples. The red intervals highlight those that do not contain the true value inside.

are known as **confidence intervals** for frequentist statistics and **credibility intervals** for Bayesian statistics[5]. At a basic conceptual level, the width of CIs can be used to judge how uncertain an estimate is: the larger the width of a CI, the more uncertain the value is. However, the frequentist and Bayesian approaches differ greatly in interpreting such values. Thus, explaining what CIs in both approaches do and do not represent is essential.

Confidence intervals

Following the logic used to build the sampling distribution, we expect that there will be some variability from sample to sample when estimating the point estimate $\hat{\theta}$. Confidence intervals construct a certain interval $[\hat{\theta} \pm Z]$ such that, in the long run, the true value of $\boldsymbol{\theta}$ will be contained in $(1-\alpha) \cdot 100\%$ of cases, where α denotes the level of confidence. For example, a CI with $\alpha = 95$ will be an interval that contains, in the long run, the true value of the parameter for 95 samples out of 100 total samples.

Confidence intervals cannot say anything about the probability of θ being inside a single particular interval: For any given interval, the true value is either inside or outside. Thus, the question "what is the probability that the true value of θ is inside the CI?" is ill-posed, as it makes no sense in this context. Instead, an $(1-\alpha) \cdot 100\%$ confidence interval means that if sampling from the same population is repeated many times, we expect that $(1-\alpha) \cdot 100\%$ of CIs contain the true parameter value for the population. Figure 2.5 shows an example of 100 estimated sample location parameters from the same population, where a 95% confidence interval was constructed for each estimate. Here, it can be seen that 95 intervals out of the total 100 contain the true value, as expected. Also important to note is that confidence intervals are not probability distributions, meaning that the values around the center of the interval are not more likely to be the true parameter than the values near the limits of the intervals.

The width of confidence intervals is mainly influenced by sample variability and sample size. When the samples from the population are very different, the resulting estimates will vary greatly, increasing the width of the interval. This variability is

[5] This naming standard is rather unfortunate, as both confidence and credibility intervals use the same acronym *CI*. However, context usually dictates what kind of CI the analyst refers to. Most CIs seen "in the wild" are frequentist confidence intervals.

usually out of the analyst's control, as it comes from the inherent variability in the observed process. As for sample size, bigger samples almost always result in shorter confidence intervals. This occurs because adding more information reduces the uncertainty of the estimated parameter values. Therefore, a straightforward way of reducing uncertainty is to increase the sample size or, at least, use the existing observations to extract more information from them. This latter approach is the one used for the spatial methods explained in the next chapter.

The construction procedure of confidence intervals is usually based on asymptotical methods. These methods work in the large-sample setting, where it is assumed that the $\hat{\theta}$ estimates will be distributed according to a Gaussian distribution, as per the central limit theorem. In fact, for unbiased estimators like the MLE under the large-sample setting, the estimates are Gaussian distributed (eq. (2.10)), with a variance given by the Fisher information matrix. Confidence intervals are then constructed by finding the corresponding range of parameters where the α probability level of this Gaussian distribution is contained. Asymptotical methods can also propagate the uncertainty from the parameters to the predictions. For an example see the delta method, which also assumes asymptotical normality of scalar functions $g(\hat{\theta})$ to construct confidence intervals.

Constructing confidence intervals using asymptotical methods requires making several assumptions that do not always hold when using data. Therefore, a way of assessing confidence intervals helps check the assumptions. This assessment is given by the **coverage probability** of confidence intervals. The coverage probability is the probability of the true value being inside the confidence intervals. In this case, the *nominal* coverage is the desired theoretical coverage of the intervals, usually taken to be 95%. In contrast, the *empirical* coverage is the one actually shown by the constructed intervals. The difference between the nominal and empirical coverage can be large when either the asymptotical assumptions are violated, or the model is misspecified. The calculation of coverage probabilities is done using simulations: A large number of n-samples from a population with a known parameter value θ is obtained, from which n-confidence intervals are constructed. The empirical coverage is then the proportion of intervals containing the parameter's true value (Schall, 2012).

Credibility intervals

In contrast to the frequentist approach, the Bayesian approach assumes that the parameters $\boldsymbol{\theta}$ and the observations are uncertain. This contrast is reflected by the fact that in Bayesian inference, every parameter has its own probability distribution given by the prior and the posterior. The uncertainty of the parameter value is then given directly by the resulting distribution of each parameter; no extra computation is needed.

After performing inference with Bayes' rule, the uncertainty of the parameter's value is given by the posterior distribution $p(\boldsymbol{\theta} \mid D)$. From the posterior, credibility intervals with a probability level of α can be constructed as the equal-tailed region of the distribution where $\alpha \cdot 100\%$ of the density is concentrated. For example, a 50% credible interval is the equal-tailed region that contains 50% of the density of the posterior. Similar to confidence intervals, it is common to use levels of 90% and 95% for constructing credible intervals.

An alternative to credible intervals is given by the **Highest Density Interval** (HDI), which is constructed similarly but is not always equal-tailed. The difference between CIs and HDIs is seen chiefly for non-unimodal posterior distributions. Figure 2.6 shows the credible intervals and the HDI for the example in Fig. 3.9.

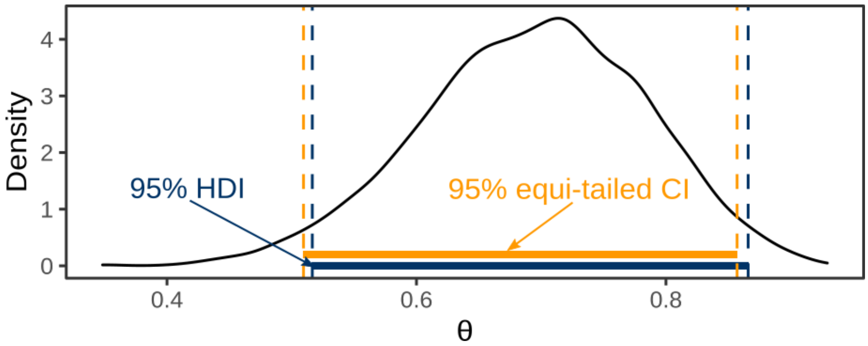

FIGURE 2.6: 95% HDI and equi-tailed credibility interval for the posterior distribution from the example seen in the bottom right plot of Fig. 3.9. Notice that the HDI and the CI differ by a little amount.

In contrast to confidence intervals, the interpretation of credible intervals is more intuitive. Credibility intervals (or HDIs, for that matter) answer the question, "given the observations and our previous knowledge, what is the probability that the true value of the parameter will be inside the interval?". Thus, if one has a 95% credible interval, it is correct to say that any value inside of the interval has a probability of 95% being the true value, given the previous assumptions. Note that the interpretation of credible intervals always includes previous knowledge, as the choice of prior has an important influence on the resulting credible intervals.

Credibility intervals and HDIs are merely summaries of the posterior distribution and, therefore, do not represent distributions. Thus, when given only a credible interval, it is generally not possible to know if the values in the middle of the interval are more probable than values near the boundaries of the interval. This information is given only by the full posterior distribution, and as such, it is always recommended to report the full posterior when performing Bayesian inference.

2.5 Model predictions

Once a statistical model for the random variable Y has been chosen, and its parameters $\boldsymbol{\theta}$ have been estimated with the methods of the last section, the next step is to obtain predictions from the model. Predicting from a statistical model involves finding what values of Y are likely to occur in different contexts. For example, we could be interested in:

- finding what value of Y is the most likely for a particular value of a predictor x,
- finding the most likely value of Y that has a 0.95 probability of not being exceeded,
- finding how likely it is that a certain value of Y is exceeded,
- finding 1000 values of Y, distributed according to the data-generating process,
- among other aspects.

All of the above are examples of **prediction** using statistical models. These procedures require knowing the model's probability mass function (for discrete variables) or the probability density function (for continuous variables). However, a problem arises when considering that the parameters entirely determine the shape of the resulting mass/density function: as we have seen in the last section, the value of the estimated parameters is uncertain. As every parameter value leads to a different function, the resulting probability mass/density functions are also uncertain.

In this section, we will tackle the problem of propagating uncertainty from the parameters to the predictions made from a model. The specifics of how to get predictions from an extreme-valued function will be seen in the next chapter, after introducing the GEV distribution.

2.5.1 Uncertainty propagation

As discussed in the last section, all statistical models always possess an element of uncertainty. Because the value of the parameters can never be known with complete certainty, this always means that the predictions from a model will also contain some uncertainty. After all, every possible parameter value will eventually lead to different predictions, so if the parameter values are uncertain, so are the predictions.

Calculating the uncertainty of the predictions made by a model is done by "propagating" the uncertainty from the parameter values to the predictions. As uncertainty is handled very differently by the frequentist and Bayesian paradigms, this propagation is also very different for each approach. We now explore the main ways of propagation for the frequentist and Bayesian approaches.

Frequentist propagation of uncertainty

Propagation of uncertainty for frequentist methods is done via two main approaches: asymptotical methods and resampling methods. The former approach, using asymptotical methods, is, in essence, the same as the asymptotical methods used to construct the confidence intervals in the first place. The difference is that instead of using the approximate normality of $\hat{\theta}$ to construct the CIs, the approximate normality of a scalar function of the estimator $g(\hat{\theta})$ is used to construct the CIs. The usual way of performing this approximation is the **delta method**, explained in more detail by Coles (2001). The delta method is, however, not the only asymptotical method that exists to propagate uncertainty. Other methods include the deviance function or the profile likelihood functions.

The other approach to uncertainty propagation, resampling methods, is performed by creating many artificial samples from the original dataset. For each of these samples, an estimate of the parameter is obtained, which is then used to get a prediction. In the end, all the predictions from the many samples are collected, from which the desired CI can be obtained by simply taking the corresponding empirical quantiles. The most widely used resampling method is known as the **bootstrap**. A comprehensive introduction to the bootstrap method is given by Davison et al. (1997).

Bayesian propagation of uncertainty

The result of Bayesian inference is a distribution of the parameters $\boldsymbol{\theta}$ conditional on the observations (that is, the posterior distribution $p(\boldsymbol{\theta} \mid D)$). From the posterior, information about the uncertainty of parameter values is given as credibility intervals. However, for prediction studies, we also desire to know the uncertainty of predicted values, denoted by y_{pred}. Thus, a way to propagate the uncertainty from the parameters to the predictions is needed.

Take as an example a model used for precipitation intensity. This model uses a Gamma distribution to describe the intensity (a choice justified in Wilks (2011)). We fix the value of the β parameter to be constant to simplify the example. Using very broad priors for the α parameter of the Gamma distribution, we arrive at the posterior distribution $p(\alpha \mid D)$. From this posterior, we extract the 95% credibility intervals $[2.99, 4.38]$, which represent the interval where the value of α has a 95% probability of being located, according to our previous assumptions. The question is now: what is the distribution of possible rainfall intensity values given this distribution of parameters? Remember that for every possible value of α and β, there will be a corresponding distribution of intensity values.

A first approach is to take the value of $p(\alpha \mid D)$ with the highest density (e.g., the mode) and use this single value as the pointwise estimate of α. However, by doing this, the resulting distribution of predictions will be overly confident, as it has ignored all the uncertainty in α.

To overcome this issue, instead of taking a single value, we can marginalize the predictions y_{pred} out of the posterior by integrating over all the values of $\boldsymbol{\theta}$. In mathematical terms,

$$p(y_{\text{pred}} \mid D) = \int_{\theta} p(y_{\text{pred}} \mid \boldsymbol{\theta}, D) p(\boldsymbol{\theta} \mid D) d\boldsymbol{\theta}. \tag{2.16}$$

In practical terms, this integral can be seen as a weighted average of all the possible distributions given by $\boldsymbol{\theta}$, where the weight is given by the probability of each parameter value given by the posterior. This idea is illustrated in Fig. 2.7. The resulting distribution from this integration is known as the **posterior predictive distribution**. From the posterior predictive distribution, credible intervals can describe the uncertainty of the predicted y_{pred}.

The integral of eq. (2.16) could be challenging to solve. Fortunately, for MCMC samples it is sufficient to sample a value from the distribution given by each sample; the resulting distribution of the sampled values will be equivalent to $p(y_{\text{pred}} \mid D)$. This avoids the need to actually perform the integral.

2.6 Model validation

Statistical inference gives us values of the model parameters $\boldsymbol{\theta}$ that align with our assumptions about the model and the data. The question is then to determine how good the resulting models are. However, there is no single universal definition of what a “good” model is, as this is determined by the ultimate goals for the model. When the goal is to make predictions about unobserved events, good models should have the following characteristics:

- The model should be a self-consistent logical representation of the observed data.
- The model should make predictions close to the unobserved data’s true values.

The first point, model consistency with the assumptions, can be performed using what is sometimes known as **model diagnostics**. It is important to remember that a model with perfect logical consistency does not assure that the model will be at all a good representation of the real world. Nevertheless, a model should always (at least!) be consistent with the logic used to create it.

The second point, related to the accuracy of the predictions, is a much more complex topic. The main issue is that the true value of the unobserved data is unknown so it is impossible to check how good the model was at predicting them directly. Nevertheless, several techniques use the observed data in clever ways to get an idea of how good a model is in predicting unobserved values (i.e., the out-of-sample accuracy). These methods include cross-validation and information criteria.

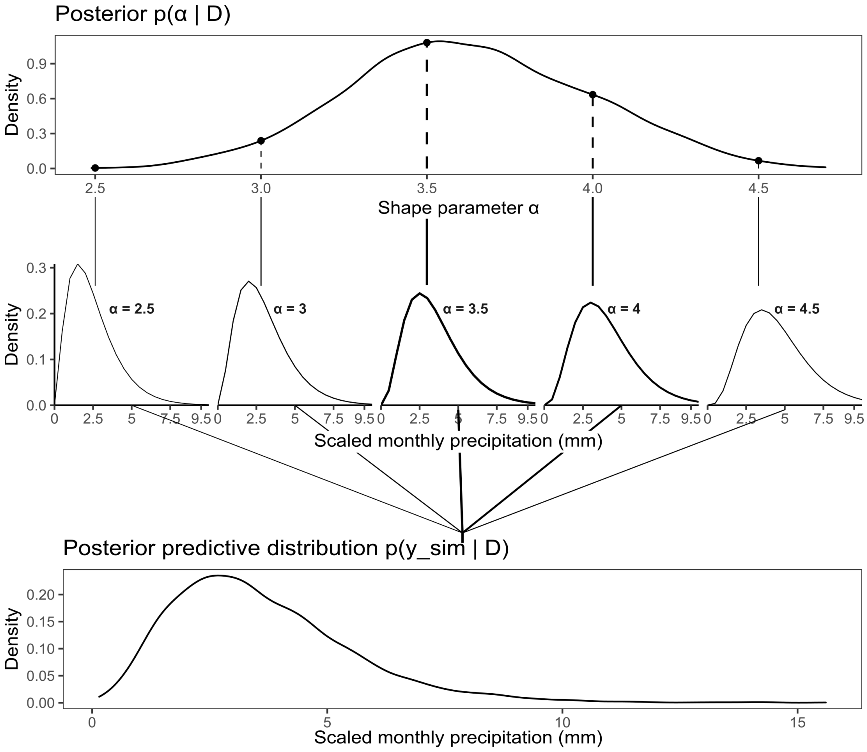

FIGURE 2.7: Example of a posterior predictive distribution for the Gamma distribution model. The top plot shows the distribution of possible values for the shape parameter α of a standard gamma distribution. The middle plot shows five standard gamma distributions resulting from using the α values marked in the posterior with dots. The line width is proportional to the probability density of each α in the posterior. The bottom plot shows the resulting distribution of samples after applying the weighted average described by eq. (2.16).

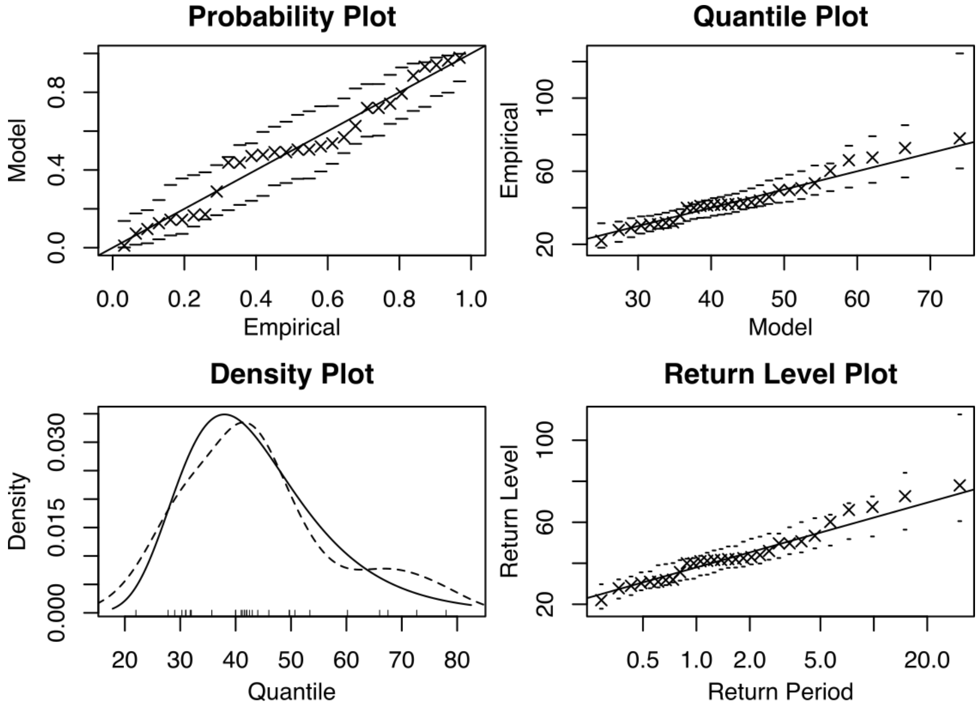

FIGURE 2.8: Model diagnostics for the GEV model fitted in Fig. 3.9. The plots shown here were generated with the *evd* package (Stephenson, 2002).

The next section will first explore the primary tools used to check the assumptions and self-consistency of statistical models, which are mainly graphical. Then, the two main methods to check model predictions will be explored.

2.6.1 Checking model assumptions

Every statistical model contains background assumptions about the underlying data-generating process and the data itself. These assumptions include, for example, assuming that the data follows a certain distribution or that the sample is representative of the population. When the assumptions made for a model are incompatible with the observed data, the model will not be a good representation. Thus, it is crucial to check that the model is a good representation of the data used to estimate it. This check is accomplished via a series of methods known as **model diagnostics**. I present here four graphical types of diagnostics, which are used in the studies presented later.

The first graphical method is to superpose the probability density given by the fitted model with either the histogram or the estimated density of the data (bottom left plot of Fig. 2.8). The resulting plot is sometimes known as the **density plot**. The density given by the model should be similar to that of the observations; significant differences between them signal that the model is not a good representation. This method is possibly the easiest one to interpret, but it is also the most unreliable: It is hard to judge differences in the tails from a simple visual analysis. Therefore, the density plot is a useful first approximation but must always be followed by another diagnostic plot.

The next two diagnostic plots are based on the **empirical distribution function**. Let $\{x_1 \leq x_2 \leq \cdots \leq x_n\}$ represent an ordered sample of independent observations. Following Coles (2001), the empirical distribution function F is defined by

$$\tilde{F}(x) = \frac{i}{n+1} \, \text{for} \, x_i \leq x < x_{i+1}. \tag{2.17}$$

From this formulation, it is apparent that for any x_i, exactly i of the n observations have a value less than or equal to x_i. Therefore, an estimate of the probability of an observation being less or equal to x_i is given by $\tilde{F}(x_i)$. The great advantage of $\tilde{F}(x)$ is that it is a model-free estimate of the actual distribution function F. As such, it can be used to judge how appropriate a certain model $\hat{F}(x)$ is: If $\hat{F}(x)$ represents an appropriate model to describe the data, then $\hat{F}(x)$ should be similar to $\tilde{F}(x_i)$: deviations from the model to the empirical distribution indicate problems with the background assumptions. Two graphical models exist to compare $\hat{F}(x)$ with $\tilde{F}(x_i)$: probability-probability plots and quantile-quantile plots.

Probability plots, also known as PP plots, consist of a simple scatter plot comparing $\hat{F}(x)$ and $\tilde{F}(x_i)$. For $\hat{F}(x)$ to be considered a good model for the data, the dots should lie close to the unit diagonal. An example of a PP plot is shown in the upper left panel of Fig. 2.8.

The probability-probability plot compares the full support of both the empirical and theoretical distribution. Thus, this plot focuses on the central region of the distributions, making a comparison of the tails very difficult. When the analysis aims to focus on the tails of a distribution we need to transform the visualization to focus on this region. This is done via the **quantile-quantile plot**, also known as the QQ plot. The QQ plot is also a scatter plot, but in contrast to the PP plot, the QQ plot plots $\hat{F}^{-1}(x)$ vs. x_i. The inverse of the distribution function F^{-1} is commonly referred to as the **quantile function**, from which the name QQ plot stems. As before, the model $\hat{F}(x)$ is considered a good representation of the data when the points lie

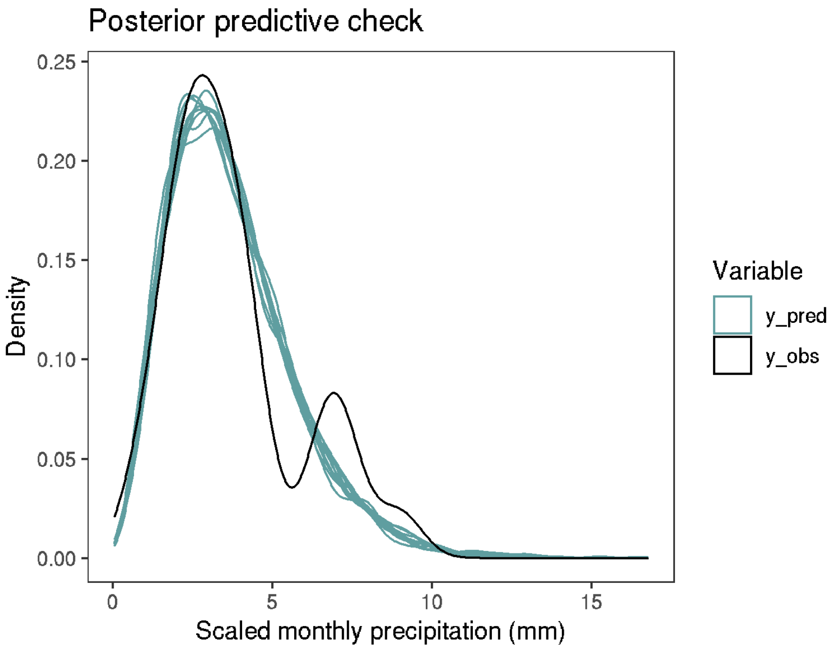

Figure 2.9: Example of a posterior predictive check using the posterior predictive distribution from Fig. 2.7. The estimated density for the observed data (black line) is superposed with 10 samples from the posterior predictive distribution. In this case, the simulated data is more or less capable of replicating the observed data, a positive sign for this model.

close to the unit diagonal. The QQ plot is handy for diagnosing the extreme-valued models explored in the next chapter. An example is given in the upper right panel of Fig. 2.8.

Another diagnostic plot used in this thesis is the **return level** plot. A more precise definition of return levels is given in the next chapter, but in a nutshell, they represent the magnitude of the random value for a certain fixed probability of non-exceedance. An example is seen in the bottom right of Fig 2.8. As before, an model is appropriate when the values lie close to the diagonal line.

The previously described diagnostic plots are valid for both the frequentist and Bayesian approaches. However, in the case of the Bayesian approach, those plots are mainly based on a pointwise estimate of the parameter values, commonly taken as the posterior distribution's median. An alternative that uses the full posterior distribution is based on the posterior predictive distribution. The PPD is useful to getting the uncertainty of y_{pred} and as a model diagnostic. For the diagnostic, the distribution of the observations is plotted simultaneously with the posterior predictive distribution. The reasoning is that we should expect the model to be able to predict the observations used to fit it; a model whose predictions were very different from the observations should be cast as highly suspicious. An example of this procedure, known as the **posterior predictive check**, is seen in the rainfall intensity example in Fig. 2.9.

2.6.2 Checking model predictions

A crucial step in prediction using statistical models is to determine just how accurate the predictions are for unobserved values. This evaluation is known as the **out-of-sample prediction accuracy**. This problem is not trivial; by definition, we do not have access to unobserved values. Furthermore, we first need to find a measure of accuracy for our model, that is, a measure of the distance between a model and some target. This last problem requires a detour in the field of information theory to arrive at the concept of deviance.

Information Theory: entropy and divergence

In order to know how accurate our model is, we need to find a measure of distance between the model and our target. In the context of statistical modeling, both the model and the target are statistical distributions. Thus, we need to find a measure of distance between two distributions. This measure is given by the field of information theory, developed in the 1940s. The basic concept is that of information, which is defined as “the reduction in uncertainty when an outcome is learned.” **Information entropy** gives the measure of this uncertainty for any probability distribution. For a random variable X with n different possible mutually exclusive and collectively exhaustive events, where each event i has probability p_i the information entropy is defined as:

$$H(X) = -E[\log(p_i)] = \sum_{i=1}^{n} p_i \log(p_i). \tag{2.18}$$

By themselves, the values of $H(X)$ for any given distribution do not have much meaning other than larger values being somewhat associated with higher uncertainty. Instead, entropy becomes interesting when it is used to compare several distributions, as it can be used to build the measure of accuracy we need.

Assume that we have two different distributions, p and q, where p represents the true target distribution of the data, and q represents the distribution assigned as the model. We want to find a measure of how much uncertainty is added when using the model q to describe the true probability of the events given by p. This added uncertainty is given by the **cross-entropy** of (p, q), defined as:

$$H(p, q) = -\sum_{i=1}^{n} p_i \log(q_i). \tag{2.19}$$

Therefore, a measure of the accuracy of the model q can be derived by calculating how much additional uncertainty was added by using q to describe the events of p (given by $H(p, q)$) compared to the initial uncertainty of p (given by $H(p)$). In mathematical terms:

$$D_{\mathrm{KL}}(p, q) = H(p, q) - H(p) = \sum_{i=1}^{n} p_i(\log(p_i) - \log(q_i)). \tag{2.20}$$

This quantity is known as the **Kullback-Leibler divergence** or simply the KL divergence. When $p = q$, the KL divergence is $D_{\mathrm{KL}} = 0$. A perfect model would have zero KL divergence; the bigger the divergence, the more distant our model is from the target. Note that the KL divergence is not symmetrical, as $H(p, q) \neq H(q, p)$. From eq. (2.20), it can be seen that the KL divergence is merely the average difference in log probability between the target (p) and the model (q).

The KL divergence has a deep connection with all methods for statistical inference. For example, it can be proven that finding the MLE estimator $\hat{\theta}$ described in section 2.3.1 is asymptotically equivalent (for large n) to finding a parameter $\hat{\theta}$ for a probability distribution that minimizes the KL divergence with the true distribution from which the data was generated. For Bayesian inference, KL Divergence can be seen as how much uncertainty was added when going from the prior to the posterior.

This helps in finding priors that maximize the entropy of the problem, which lead to posteriors that maximize the information.

The fact that KL divergence increases as the model is more different than the target can be used to compare between different models: If we had to choose between two candidate models, q and r, the one with the least KL divergence would be the best choice. The difference between both divergence values is given by:

$$\begin{aligned} D_{\mathrm{KL}}(p,q) - D_{\mathrm{KL}}(p,r) &= \sum_{i=1}^{n} p_i(\log(p_i) - \log(q_i)) - \sum_{i=1}^{n} p_i(\log(p_i) - \log(r_i)) \\ &= \sum_{i=1}^{n} p_i(\log(p_i) - \log(q_i) - \log(p_i) + \log(r_i)) \\ &= \sum_{i=1}^{n} p_i(\log(q_i) - \log(r_i)) \\ &= \mathrm{E}[\log(q)] - \mathrm{E}[\log(r)]. \end{aligned}$$

As can be seen, most terms with the true probability of the target (p) cancel out. This is desired, as usually we do not know what p is. As long as we can find the model's average log-probability $\mathrm{E}[\log(q)]$, we can get a clear picture of which one is closer to the target. Furthermore, while the p term inside the expectation is unknown, it turns out that simply taking the sum of the log-probabilities is a good approximation of the average log-probability:

$$\mathrm{E}[\log(q)] \approx S(q) = \sum_{i=1}^{n} \log(q_i). \tag{2.21}$$

Therefore, an approximation of the difference in the KL divergence between the two models can be given as $\sum \log(q_i) - \sum \log(r_i)$. The absolute value of either $S(q)$ or $S(r)$ has no meaning; only the difference between them gives us information about the accuracy of the model. The quantity $S(q)$ is sometimes known as the **log-probability score**, as it can be used to compare between models. The quantity $-2S(q)$ is sometimes known as the **Deviance**, another commonly used score used to compare between models.

When working with Bayesian inference, direct use of both the log-probability score and the Deviance can be challenging, as every event i has a distribution of possible values of p_i. The first approach would be to simply get the median probability for every p_i in the posterior. However, by doing this, we are wasting most of the information contained in the posterior. Thus, finding the logarithm of the average probability for each i is better, where the average is taken over the entire posterior distribution. The result of this operation is known as the **log-pointwise-predictive-density** (lppd). In mathematical terms:

$$\mathrm{lppd}(x, G(\boldsymbol{\theta})) = \sum_{i} \log \frac{1}{N} \sum_{n} p(x_i \mid \boldsymbol{\theta}_n), \tag{2.22}$$

where x represents the observed data, $G(\boldsymbol{\theta})$ is the posterior distribution, $\boldsymbol{\theta}_n$ is the n-th sample from the posterior distribution, and N is the number of MCMC samples. The lppd is the Bayesian analog of the Deviance.

To summarize, this section addressed two problems that arise when estimating the predictive accuracy of a statistical model, namely:

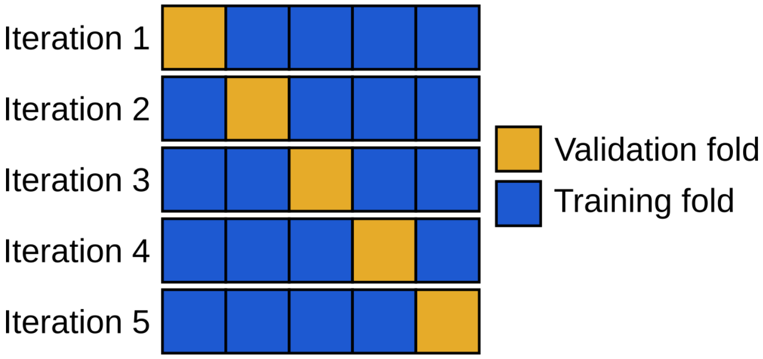

FIGURE 2.10: Graphical depiction of how $K-$fold CV works. A dataset is divided into K-folds, from which the model is fitted using $K-1$ folds. The remaining one is used as the validation set. This process is iterated until every fold is used as the validation one.

- How to construct a measure of the distance between a model q and the target distribution q (the KL Divergence), and
- how to estimate this distance to compare two models q and p with the information at hand (the Deviance and lppd).

From predictions to out-of-sample model accuracy

So far, we have seen that the KL Divergence can be used to judge between two models under the precept that a model that is closer to the target distribution is better. However, a not-so-surprising result arises when using only the Deviance/lppd/log-score as a measure of accuracy: More complex models always lead to better score values [6]. This is a problem because models with a lot of parameters (relative to the number of data points) always lead to **overfitting**. Overfitting simply means that the model has learned too much from the data and, as a result, cannot make good predictions for data it has not observed before. The contrary phenomenon, **underfitting**, occurs when the model learns too little and thus cannot predict observed or unobserved data. Thus, we want to construct a way of using the accuracy scores such that it avoids both under- and overfitting. This is accomplished in two different ways: cross-validation and information criteria.

Cross-validation is one of the simplest and most commonly used methods to estimate the out-of-sample model accuracy. The idea is to split the data into a training and a validation set, where the training set is used to do inference on the model parameters, and the validation set is used to compute the accuracy of the predictions. However, because throwing away data for the validation is undesirable, the data is first split into K-chunks (typically known as folds). Each of the $K-$folds is then used as the validation set, using the other folds for the training. This means that all the data is used to train and validate the model, avoiding waste. This procedure, known as **$K-$fold cross-validation**, can be graphically seen in figure 2.10.

An important design choice when performing Cross-Validation is the number of folds (K). A typical choice is $K = 10$, which has been justified as a good default choice by several authors (Hastie et al., 2009). 10-Fold cross-validation is the default for the verification section of the work done in this thesis. An alternative is to use as many folds as there are data points: this is known as **Leave-One-Out-Cross-Validation** (LOOCV).

[6] Complexity in this context means adding parameters to the model

By itself, cross-validation is a very general method to estimate out-of-sample model accuracy. It is not exclusively used to determine the out-of-sample divergence. For example, many different metrics can be used with cross-validation, such as RMSE or the Quantile Score. Let $\mathcal{K} : \{1, ..., N\} \rightarrow \{1, ..., K\}$ be an indexing function that denotes the fold to which the randomization assigns the observation i. The fitted model is then q^{-k}, where the kth part of the data was removed when training the model. The cross-validated estimate of the out-of-sample accuracy is then

$$\text{CV}(p) = \frac{1}{N} \sum_{i=1}^{N} S(y_i, q^{-k(i)}), \tag{2.23}$$

where $S(\cdot)$ represents the scoring function used, and $k = 1, ..., K$.

Depending on the number of folds, $K-$fold cross-validation is a simple way to estimate the out-of-sample accuracy for our model p. Common examples of scoring functions include the root mean square error (RMSE) or the mean average error (MAE). For the studies described in this thesis, two different score functions $S(\cdot)$ are used: the Quantile Score and the Deviance (or lppd in the Bayesian case). The Quantile Score is used to measure the accuracy of the predictions for the final model, while the Deviance/lppd was used to guide the choice of model. The specific formulation of the Quantile Score is given in sections 5.2.4 and 4.2.2.

The method described in eq. (2.23) requires that the model be fitted K-times, which in the case of large K can be quite demanding. The most extreme case, LOO-CV, is challenging for Bayesian inference, as it would require K posterior distributions to be sampled. For example, if we had 100 observations, we would have to compute 100 posterior distributions (which depending on the model, could take a couple of days per posterior). To circumvent this, Vehtari et al. (2017a) developed a method to approximate the LOO deviance without having to fit the model many times. This handy method, called the **Pareto-Smoothed importance sampling cross-validation** (PSIS), is based on the so-called importance sampling, where each observation is weighted according to how common it is. The weights are calculated in the background using the Generalized Pareto Distribution, which compares how extreme each observation was compared to what was expected. See Vehtari et al. (2017b) for more information on PSIS and how to use it to estimate the LOO-CV deviance. For the Bayesian sections of this dissertation, the LOO-CV deviance was estimated using PSIS.

An alternative to cross-validation is **Information Criteria**, where instead of "brute-forcing" the estimate by reusing the data, a theoretical estimate of the relative out-of-sample KL divergence is constructed. The most famous (and commonly used) example of information criteria is the **Akaike Information Criterion** (AIC) [pronounced a-ka-e-ke]. The AIC has the following formula:

$$\text{AIC} = D_{\text{train}} + 2p = -2\text{lppd} + 2p. \tag{2.24}$$

Here D represents the deviance of the training sample, which is approximated by the lppd. The p represents the number of parameters present in the model. Because of the -2 present in the first term, the preferred model is the one with the lowest AIC. From eq. (2.24), it is apparent that the more parameters in the model, the more the AIC will grow, due to the term $2p$ acting as a penalty term to avoid overfitting.

Just as the other derivated measures of relative KL divergence, AIC (and, in fact, all information criteria) do not provide any information about the absolute accuracy of a model. They can only be interpreted in the context of comparing several models. After choosing a model with information criteria, the predictions should still be checked for out-of-sample accuracy using a method like cross-validation.

The AIC is an effective tool for model selection, used in many publications. However, it makes rather strict assumptions about the model - for example, it assumes that the sample size N is much larger than the number of parameters; furthermore, for the Bayesian approach the AIC assumes a very flat prior over the parameters and that the posterior distribution is multivariate Gaussian. To counteract this, newer approximations with less restrictive assumptions have been developed, which have surpassed the AIC in almost every aspect. For the Bayesian paradigm Watanabe et al. (2010) developed the **Widely Applicable Information Criterion** (WAIC), which is an ideal alternative to AIC as it makes no assumptions about the shape of the posterior. The WAIC is calculated as

$$\mathrm{WAIC(x, G(\boldsymbol{\theta}))} = -2\left(\mathrm{lppd} - \sum_{\mathrm{i}} \mathrm{var}_\theta \log \mathrm{p(x_i} \mid \theta)\right). \tag{2.25}$$

This time the penalty term is $\sum \mathrm{var}_\theta \log p(x_i \mid \theta)$, which is proportional to the variance of the posterior predictions. The WAIC is computed pointwise (i.e., individually for each observation), so every observation has its own penalty score. Just as with the AIC, the model with the lowest WAIC value should be preferred.

Before explaining how these out-of-sample accuracy measures work in practice, it is important to point out some of their limitations. First, by design, information criteria like AIC or WAIC will not always point to the "true" model; in statistical jargon, it means that information criteria are inconsistent for model identification. The inconsistency of information criteria arises because, in many situations, the model with the best predictions will have either biased parameter values or a different number of parameters than the true model. As a sidenote, this identification problem is part of why causal analysis is a very different problem from prediction, and thus, the chosen parameters of a model from Information Criteria should not be used to infer causal relationships.

Finally, it is worth mentioning that an interesting relationship arises between WAIC/AIC and the PSIS-LOO-CV. First, both PSIS-LOO and WAIC are pointwise, so each observation gets its own accuracy value. More importantly, it turns out that when the number of observations is really large, the value of LOO-CV and WAIC/AIC converge (in fact, this convergence can already be seen for moderately large data samples). Based on this, Watanabe postulated the following rule-of-thumb: If the PSIS-LOO and the WAIC for a certain model are very dissimilar, one (or both) of the two scores is probably unreliable.

2.6.3 Model validation in practice

This section began by describing the issue of measuring the accuracy of models and arguing that KL divergence is the measure that tells us which model is closer to the target. It went on to explain how to construct an estimate of this quantity, especially for the out-of-sample accuracy, using either resampling methods (cross-validation) or a theoretical approximation (information criteria). But how is this applied in practice in the context of a study where prediction with a statistical model is the goal?. This is usually done with the following steps:

1. If deciding between two or more models to use for the data, use information criteria like AIC or WAIC to choose the one with the best out-of-sample accuracy.

2. Check that the fitted model properly represents the data using model diagnostics like the PP or QQ-plots. In the case of Bayesian inference, check additionally that the model can generate the data using a posterior predictive check.

3. Get an estimate of the out-of-sample prediction accuracy using a cross-validated metric like the RMSE or the Quantile Score.

If performance is not what was expected in any of these steps, the model should be revised. The results for this validation were reported for all the studies included in this thesis.

It is important to remember that even a model that passes all steps can still be a poor representation of the actual data-generating process. A statistical model that has good out-of-sample prediction accuracy can still catastrophically fail when used in a different setting or when trying to explain causality.

2.7 Summary

In this section, we have seen an overview of the steps required to use a statistical model. The first step is to choose an appropriate model to represent a process that can be seen as stochastic. This model is a mathematical formula whose behavior is controlled by quantities known as parameters. In order to propose values of the parameters, inference is performed by using the data to find parameter values that result in a model that is able to properly capture the variability seen in the observations. However, the stochastical nature of these estimates always results in uncertainty about the value of the parameters, and in turn, also in uncertainty about the predictions made by the models. In this chapter, several methods of estimation were explored, which can be broadly classified as frequentist or Bayesian. Additionally, this chapter explored how to get predictions from a model, which required a way of propagating the uncertainty from the parameter values to the predictions. Finally, several aspects of model validation were explored, which are used to check that the predictions are in line with what is expected from the data.

3 Stochastic models for Rainfall Extremes

In this chapter, the tools presented for statistical modeling in the last chapter are extended to handle spatial data from extreme rainfall events. This extension first requires the introduction of univariate extreme value theory, which is combined with spatial statistics to arrive at the methods englobed by so-called spatial extremes.

3.1 Statistical dependence of extreme rainfall

One of the goals of this thesis is to take advantage of the existing statistical dependence on an extreme rainfall series to improve and reduce the uncertainties of statistical models' estimates. The base assumption here, of course, is that such a statistical dependence exists in the datasets we used.

This section explores why extreme rainfall datasets, comprised of records for several rain gauges in the same geographical catchment, virtually always show some level of spatial dependence. We will discuss how this is ultimately the result of the prevailing meteorological processes in the region. Additionally, we will show another type of dependence present in rainfall maxima of different durations; this "temporal" dependence can be modeled with the same methods used to deal with spatial dependence, also allowing the analyst to improve the estimates.

This section discusses the statistical properties of datasets made up of rainfall maxima. For clarification, we consider these datasets to be comprised of the annual or semi-annual maxima of precipitation height for each year in the observation record. The reasoning for this choice is explained in section 3.3.1.

3.1.1 Spatial dependence

Many of the methods described in this work for the statistical modeling of extreme rainfall take advantage of the spatial dependence in annual rainfall maxima for rain gauges located in nearby locations. Later in this chapter, we will see more about the concept of spatial dependence and how to include it in models. However, an important question should be answered before delving into the mathematical details: why do extreme rainfall observations typically show so-called spatial dependence?

The simplest answer to the above question is that extreme rainfall events are sometimes large enough to impact several rain gauges simultaneously. For example, the upper panel of Fig 3.1 shows an example of an extreme event where the spatial extent was such that several stations were affected simultaneously. When computing the station-wise annual maxima for this kind of dataset, the magnitude values in any given year for the stations that were hit simultaneously by the same event will likely

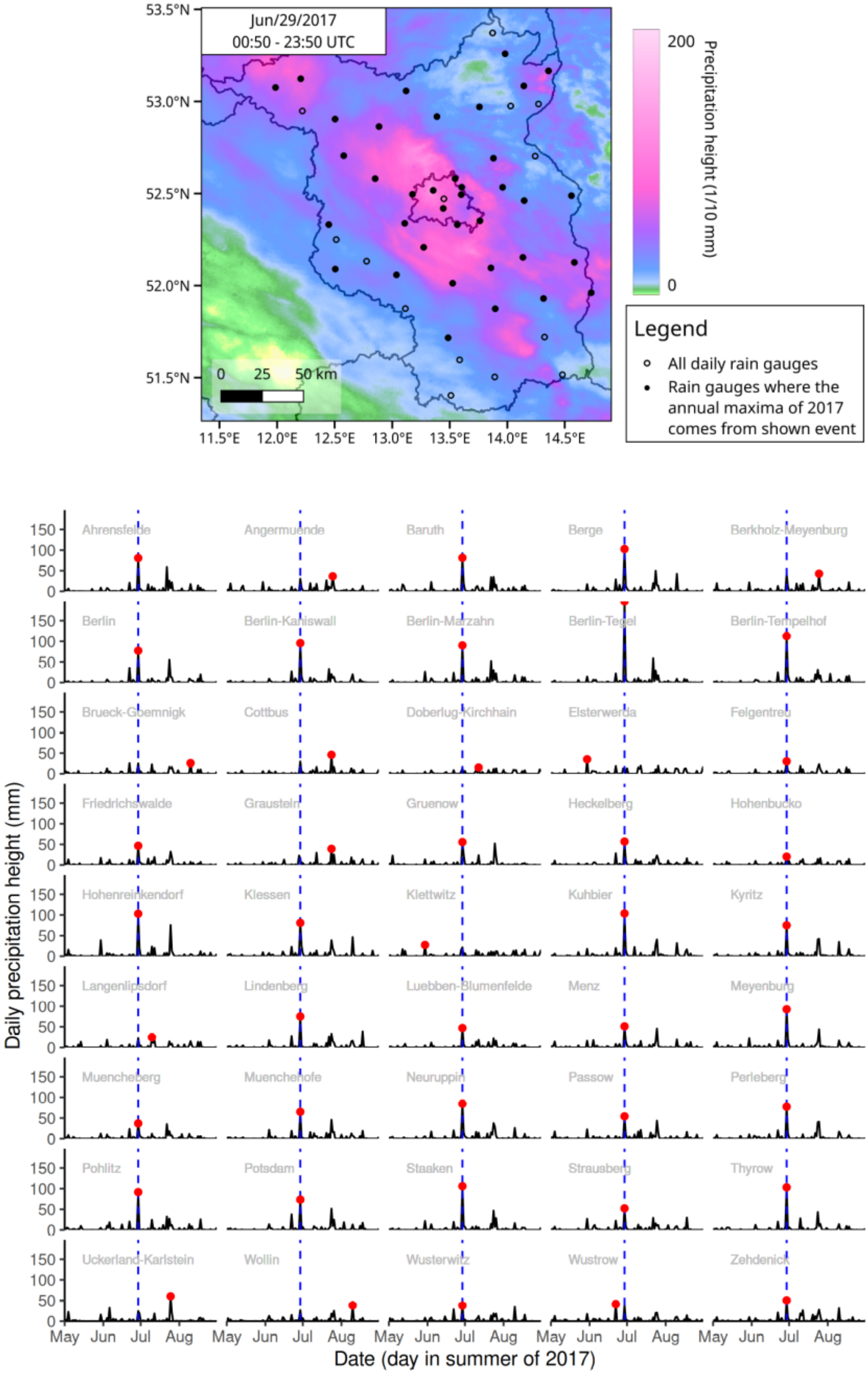

FIGURE 3.1: Upper panel: Map of Berlin-Brandenburg showing the accumulated 24-hour daily precipitation for the 2017 extreme rainfall event (data source: RADOLAN, DWD). Dots represent rain gauges from the DWD. Lower panel: Time series of the recorded daily precipitation of the 45 gauges shown in the map above for 2017. The blue dashed line denotes the day of the event denoted in the above map; the red dots indicate when the corresponding annual maxima were recorded.

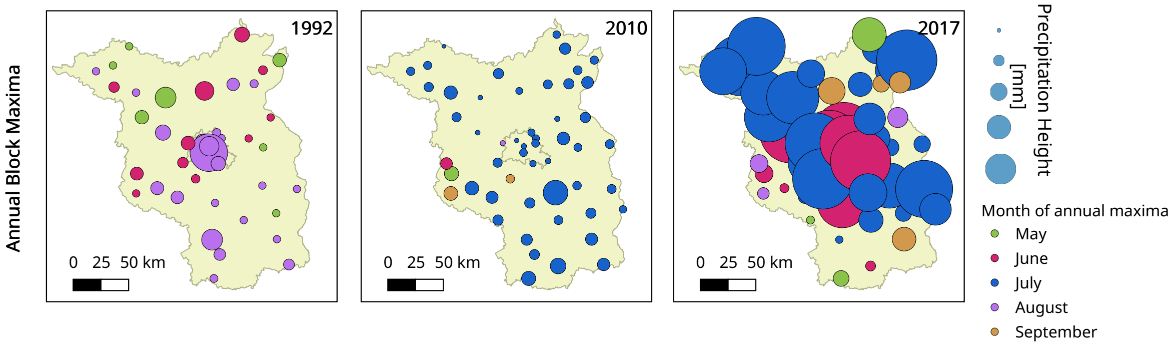

FIGURE 3.2: Maps of Berlin-Brandenburg showing the magnitude (area) and month (color) of the annual maxima registered in 1992, 2010, and 2017 for the 53 rain gauges described in section 5.2.1.

be similar. For example, the lower panel of Fig 3.1 shows that for most of the stations where the annual maxima are taken from this particular day, the measured rainfall is between 50 and 100 mm. This similarity between stations is likely repeated for several years, inducing a statistical dependence.

The explanation above helps to explain why it is common to find spatial dependence in annual maxima for gauges of the same geographical catchment. However, it is important not to be misled by thinking only about individual events: annual maxima are taken over *all* the events that occur during one year, meaning that one cannot ensure that the maxima came from the same event for any given year. For example, a quick look at the lower panel of Fig 3.1 reveals that, while most of the maxima come from the 29.06 event, the rest is distributed from June to August. In fact, we will see later from Fig. 5.9 that for the same region (with an extent of around 250 × 250km), the number of unique convective-based extreme rainfall events that result in annual maxima of 53 stations is, on average, around 12 per year.

Another example can be seen in Fig. 3.2, where the different colors indicate the date from which the annual maxima were registered. In 1992 and 2017, many different months can be seen, where clusters of nearby stations tend to have the same color. In contrast, 2010 shows that most of the annual maxima came from an event in July. Additionally, the figure shows that nearby stations also have similar magnitudes, represented by the size of the circle. It is the repeated occurrence of these patterns that eventually lead to the existence of a spatial dependence structure in block maxima data.

The examples above can be summarized by saying that the spatial heterogeneity of individual events makes it so that nearby stations could have yearly maxima stemming from different events of the same year. However, given a long-enough record, we expect, *on average,* to see similar magnitude values for nearby stations. Ultimately, each individual event's geometry is what determines how much dependence will exist in the dataset. Luckily, this geometry is not random: as explained above, each rainfall-generating process can be associated with different geometrical patterns with well-defined scales.

To better understand this idea, Fig. 3.3 shows the magnitude of block maxima for three years; however, in contrast with the last figure, here, the maxima were taken over two different seasons: winter and summer, resulting in semi-annual instead of annual block maxima. The reasoning behind this choice is that different types of rainfall processes dominate in winter and summer (this idea is explored in depth in ch. 5). For the summer of 2005, the magnitudes of the semi-annual maxima can be

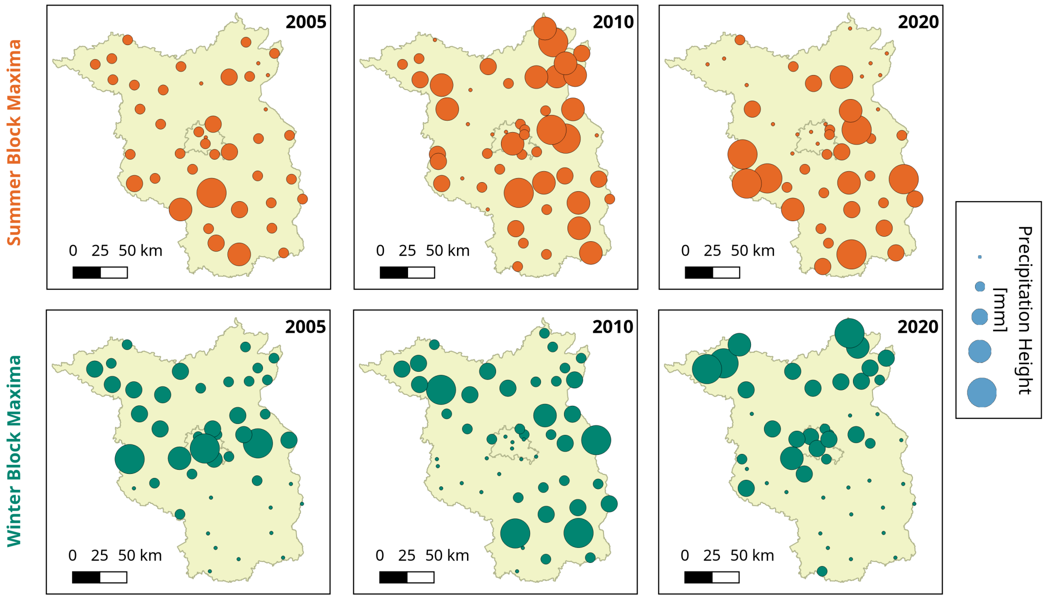

FIGURE 3.3: Maps of Berlin-Brandenburg showing the magnitude (area) of the semi-annual maxima registered in 2005, 2010, and 2020 for the 53 rain gauges described in section 5.2.1. The upper panel shows the semi-annual summer maxima, while the lower panel shows the semi-annual winter maxima.

seen to be making small clusters of several stations; this changes in the winter of the same year, where the geometry of similar magnitudes changes substantially to have a kind of barrier in the middle. This example helps to see that the dependence structure is ultimately a function of the prevailing rainfall-generating process regime.

3.1.2 Temporal dependence

Another type of dependence commonly encountered within the context of statistical extreme rainfall modeling is the one that exists between rainfall records of different time scales. To understand this, consider a rain gauge that measures precipitation height every hour, which is then hit by a storm that lasts for 6 hours. The amount of rainfall that is measured every hour (i.e., the **rainfall intensity**, commonly measured in [mm/h] or [mm/day]) changes during this interval so that at some hours, there is more rainfall than at others. An example of such an event can be seen in Fig. 3.4. We can summarize the event by getting an average intensity over the 6-hour interval; in fact, we can get the average intensity for any arbitrary duration, as seen in the left panel of Fig. 3.4.

A common strategy for studying rainfall in different time scales (i.e., different durations) is to take a rolling average of intensity for different durations. This operation is seen in the right panel of Fig. 3.4, where the initial window in the left panel is moved to the next available timestep. This operation can be repeated until all the recorded timesteps are covered, generating a time series of **aggregated** rainfall intensity values for different durations (i.e., time scales).

Once aggregated series of different durations have been generated, they can be used to study extreme rainfall events with different time scales: a possibility to do this is to take the annual maxima of intensity values for each duration. Figure 3.5 shows the resulting annual maxima for different aggregation durations for the complete data series used to generate Fig. 3.4. Notice that, in general, an increase in

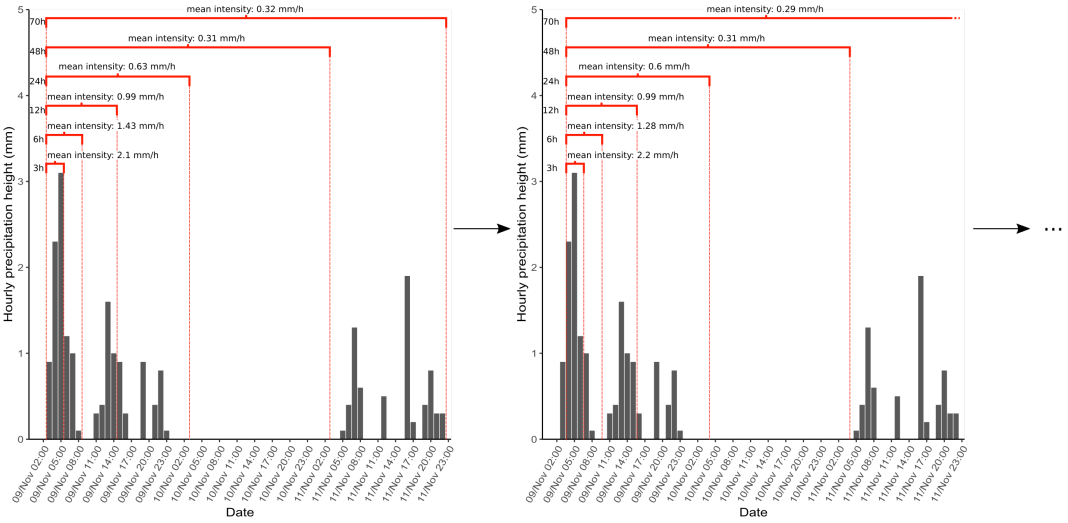

Figure 3.4: Time series showing hourly rainfall height (mm) for a timespan of approx. 3 days. The red intervals show increasing aggregation intervals where average intensity is calculated. Notice how typically intensity decreases with increasing duration. Data source: DWD

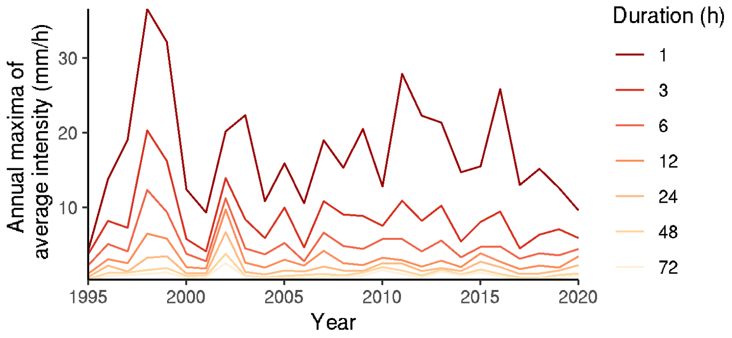

Figure 3.5: Time series of annual maxima for 7 aggregation durations. Data source: DWD

aggregation duration results in a decrease in intensity, with the lower durations showing the largest average intensity values. An extreme value distribution could then be fitted individually to each series to obtain the statistical properties of events with different durations. However, as we will explore later in this thesis, a certain model, known as the d-GEV, is capable of englobing the information of all durations into a single model, optimizing the use of information.

Figure 3.5 also shows an additional feature of this kind of maxima that is particularly relevant for this work. Notice that the line for the one and 3-hour aggregation duration have similar behaviors: when the intensity increases for the 1-hour intensity around 2007, it also increases for the 3-hour intensity of the same year. This similar behavior can be seen for durations that are "close by."[1] Most importantly, this similarity decreases when two durations are "far" from each other: the line for 1-hour intensity behaves much more differently than the line for 48-hour intensity. We can conclude from this example that annual maxima of rainfall aggregated from different durations presents what could be called a **temporal dependence**, which is strongest for similar durations.

The existence of a dependence structure for this kind of dataset follows from a very

[1] in a temporal sense

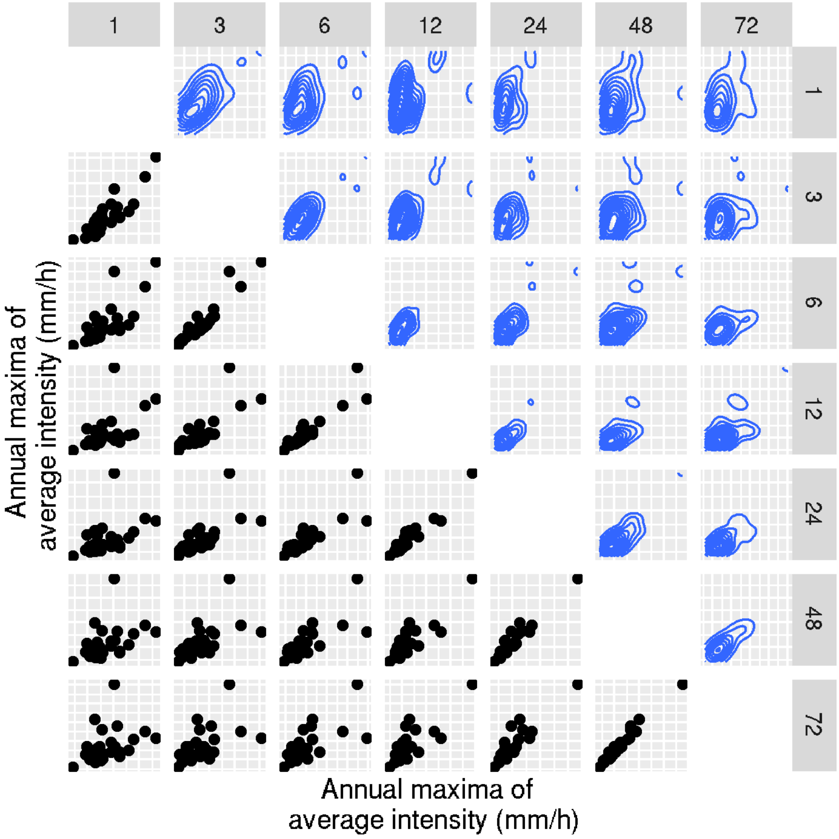

FIGURE 3.6: Correlogram for the aggregated intensity series for the durations shown in Fig. 3.5. The upper part shows the estimated bivariate density, and the lower part the scatterplot. Data source: DWD

logical idea: given a rainfall event with duration k, all the average intensity values with durations up to k (with the average taken over the time of the event) will be very similar to each other. Thus, if we have a period with a lot of short-lasting events, the short durations will have strong dependence; in contrast, if we have a period with only long events, the longer durations will tend to have a stronger dependence. Ultimately, series from short durations will tend to be similar to series from short durations; the same can be said for long durations. This idea can be seen in the correlogram of Fig. 3.6, where it can be seen that for close-by durations, a dependence structure exists.

The dependence between aggregated rainfall for different durations has been described in the literature before, commonly as a complement to the concept of "ordered random variables" (Nadarajah et al., 2019). A typical design choice for EVT models including different time scales is to ignore this dependence (Ulrich et al., 2020; Fauer et al., 2021), but recently, some studies have successfully incorporated it into the models (Tyralis et al., 2019). In a similar vein to that of spatial dependence, it could be that ignoring the dependence between durations results in estimates with underestimated uncertainties. Before this thesis, this effect was largely unknown, as the work of Tyralis et al. (2019) did not include an analysis of the impact on the model estimates or their uncertainty. In this thesis, the study presented in ch 4 explores in depth the impact that including this dependence has on the final model, thus extending the results of Tyralis et al. (2019).

Now that the two types of dependence present in the extreme rainfall datasets have been detailed, we move to the models used to represent it.

3.2 Basic probability concepts

The statistical modeling of phenomena such as extreme rainfall, temperature, or air quality entails using randomness to learn and predict properties from their behavior. However, using randomness in this form requires the assumption that these variables possess certain well-defined mathematical properties. Therefore, before delving into the specifics of how to model such environmental variables stochastically, it is important to first look at the concept of random variables and the extensions to random vectors and stochastic processes.

3.2.1 Random variables, vectors, and processes

The concepts of random variables, vectors, and fields require defining a **probability space**, which is a mathematical construct denoted by the three elements $(\Omega, \mathcal{F}, \mathbb{P})$. In general terms, these three elements are:

- The sample space Ω, that denotes the set of all possible outcomes from the experiment, which can be finite or infinite,
- the σ-algebra $\mathcal{F}$ that contains a subset of possible outcomes in the sample space, and
- the probability measure $\mathbb{P}$ that assigns a probability to each event in the event space.

In the case of environmental modeling, most sample spaces Ω are continuous (i.e., infinite). Uncountable sample spaces can create issues when choosing the σ-algebra $\mathcal{F}$, as one can show that including all possible subsets of Ω in $\mathcal{F}$ leads to the probability measure being undefined for uncountable Ω. The most common workaround for this problem is to restrict the subsets included in $\mathcal{F}$ to be the smallest possible σ-algebra that contains all open sets. Such a construction is known as the Borel σ-algebra. For the remainder of this thesis, we assume that the event space $\mathcal{F}$ is a Borel σ-algebra.

We can now think of some environmental phenomenon (e.g., rainfall depth or air temperature records) as a random experiment modeled by a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. A **random variable** (r.v.) is then defined as a function $\zeta : \Omega \to \mathbb{R}$ which is *measurable* (i.e., for every Borel set $B \subset \mathbb{R}$, it holds that $\zeta^{-1}(B) \in \mathcal{F}$). Denoting as ω the outcomes of the random experiment, $\zeta(\omega)$ represents the value of the r.v. for the outcome ω (i.e., $\zeta(\omega)$ is the data that was observed). Thus, every value $\zeta(\omega)$ of a r.v. can be assigned some probability measure $\mathbb{P}$. The function that assigns a probability measure to the set of all possible values of a r.v. is known as the **distribution** of the random variable. Distributions were already mentioned in the last chapter.

Stochastic processes extend the concept of random variables from a single function to a collection of many random variables $X = \{X_1, X_2, ...\}$. Consider the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and an arbitrary set $\mathcal{T}$ called the index set. Every r.v. in X must be uniquely associated with an element in $\mathcal{T}$ for it to be a valid index set. The collection of random variables $X = \{X(t) : t \in \mathcal{T}\}$ defined on $(\Omega, \mathcal{F}, \mathbb{P})$ is then called a stochastic process with index set $\mathcal{T}$. Therefore, for every element t in the index set $\mathcal{T}$, there corresponds a random variable $X_t : \Omega \to \mathbb{R}$ from which every outcome leads to $\omega \to X_t(\omega)$. In this case, the function $t \to X_t(\omega)$ is called the **sample path**

of the stochastic process X corresponding to the outcome ω. Stochastic processes are commonly denoted as $\{X(t)\}$, but depending on the context, they can be written simply as $X(t)$.

The choice of the index set $\mathcal{T}$ can lead to specific cases of well-known stochastic processes. For example, choosing an index set $\mathcal{T}$ with only one element leads to the stochastic process being reduced to a single random variable (meaning that random variables are a specific case of stochastic processes). Moreover, if one chooses $\mathcal{T}$ to be a finite set with n-elements, the stochastic process reduces to the collection $X = \{X_1, X_2, ..., X_n\}$ defined on a common probability space, which is typically denoted as a **random vector**. Of particular interest for this thesis is the choice of $\mathcal{T} = \mathbb{R}^d$, where $d \geq 2$. This last stochastic process is known as a **random field**. From both of these examples, it is clear that stochastic processes are a generalization of random variables, vectors, and fields.

Stochastic processes can be seen as descriptions of systems that change randomly in time, space, or a combination of both. Therefore, a noteworthy application of stochastic processes is the modeling of environmental phenomena that seems to behave randomly. For example, consider the spatial distribution of rainfall in a particular region. The amount of rainfall in every *single* location can be seen as a realization of a random variable. In turn, repeated measurements in time from each location result in a random vector. Additionally, the collection of rainfall values for *all* locations could be considered realizations of a random field. On the whole, these are all different types of stochastic processes.

The studies described in this thesis focus on the use of two types of stochastic processes for modeling extreme rainfall in space. The first one is the Gaussian process, commonly used in classical geostatistics. GPs work very well with spatial data, but as will be discussed later, they do not extend well to directly using extreme-valued data. The second process of interest is the max-stable process, which extends the concept of unidimensional modeling of extreme values to infinite dimensions. In the next section, we will explore Gaussian processes and their connection to spatial modeling. This is followed by an introduction to modeling extreme events in the univariate setting.

3.2.2 Gaussian Processes

Of all the existing stochastic processes, the **Gaussian process** stands out as one of the most useful and widely applicable for stochastic modeling. A Gaussian process is a stochastic process $\{G(t) : t \in \mathcal{T}; \mathcal{T} \subset \mathbb{R}^d\}$, where the joint distribution of any finite-number combination of random variables G_t is multivariate Gaussian-distributed. That is, if $\boldsymbol{f}$ represents any possible linear combination of the $G = \{G_1, G_2, ..., G_d\}$ random variables, a Gaussian process must fulfill the following:

$$\boldsymbol{f} \sim \mathrm{MVN}(\boldsymbol{\mu}, \boldsymbol{K}). \tag{3.1}$$

Here $\boldsymbol{\mu}$ denotes a vector of means for each G, and $\boldsymbol{K}$ denotes the **covariance matrix** (also known as the kernel). The covariance matrix gives the covariance for every two points (t, t'):

$$K_{ij} = \mathrm{cov}(G(t), G(t')).$$

The entries K_{ij} of the covariance matrix can sometimes be computed by the covariance function $C(t, t')$. The covariance function gives information about the similarity of two random variables $(G(t), G(t'))$, and can have many parametric forms. Covariance functions are explored in more detail in section 3.5.

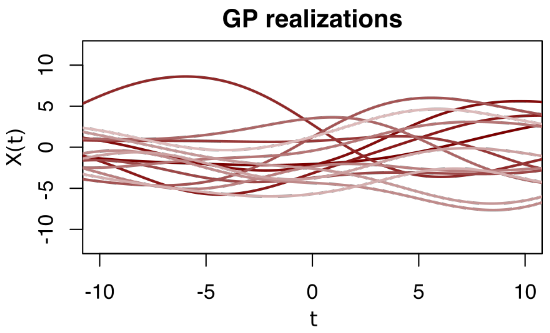

FIGURE 3.7: 15 realizations of a Gaussian Process $\{X(t) : t \in \mathcal{T} ; \mathcal{T} \subset \mathbb{R}\}$. The code used to create this example was taken from Betancourt (2020)

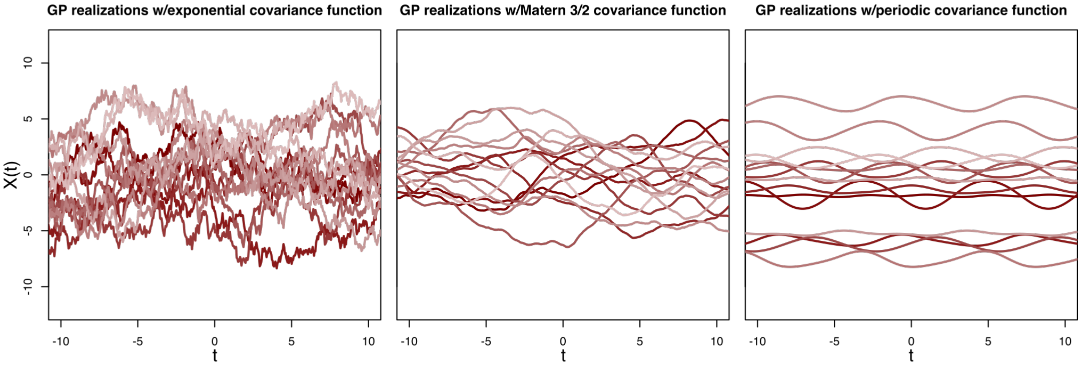

FIGURE 3.8: 15 realizations of a Gaussian Process $\{X(t) : t \in \mathcal{T} ; \mathcal{T} \subset \mathbb{R}\}$ using three different covariance functions. The code used to create this example was modified from Betancourt (2020)

An essential property of Gaussian processes is that every GP can be decomposed as the sum of a mean function with a zero-mean Gaussian process. Thus, if $\alpha(t) = \boldsymbol{\mu}$ denotes the mean function, we can write a GP as:

$$G(t) + \alpha(t) \sim \text{GP}(\boldsymbol{\mu}, \boldsymbol{K}) \tag{3.2}$$

$$G(t) \sim \text{GP}(0, \boldsymbol{K}). \tag{3.3}$$

Most applications that use GPs use this decomposition to facilitate the analysis. This is because working with zero-mean GPs shifts the focus of the characterization of the stochastic process to finding an appropriate covariance function.

Figure 3.7 shows an example of 15 realizations of a zero-mean Gaussian Process over a one-dimensional domain $\mathcal{T}$. Theoretically, the Gaussian process operates on all the infinite values of t, but in reality, we can only work with finite-dimensional spaces. Thus, in the plot of Fig. 3.7, the values of t were constrained to a grid, where the "lines" are actually an interpolation between neighboring points of t. Note that, on average, the resulting lines hover around the expected value of zero, but that does not mean that each line necessarily hovers around zero.

Gaussian processes can sometimes be seen as a distribution over functions. That is, for every point $t \in \mathcal{T}$, the Gaussian process can be seen as returning a certain function $f(t)$. This is already apparent from looking at Fig. 3.7. In this figure, we can think of each line as a different function that hovers around a particular mean μ given for each t. Furthermore, the covariance matrix $\boldsymbol{K}$ gives the overall shape of the functions. The covariance function used for creating Fig. 3.7 was from the exponentiated quadratic family. Figure 3.8 shows examples of realizations from GPs using three different covariance functions, where each family had the same magnitude and length scale parameters. Note how the covariance function completely determines how the lines behave in each case.

The properties of Gaussian Processes described above make them excellent candidates among stochastic processes for modeling environmental variables. The former is particularly true in a spatial setting, where the problem expands to infinite dimensions.

3.3 Modeling extremes

The studies described in this thesis involve modeling exceedingly rare (i.e., extreme) rainfall events. By definition, these events occur very infrequently, and thus, the available observations are very few. The goal in modeling these events is typically to infer more about the possible frequency and magnitude of future events – this information can then be used to design infrastructure or make decisions. Extreme events are not limited to meteorological ones; in fact, the theory was first developed by the insurance sector and is typically used by the financial sector.

The question then reduces to finding appropriate statistical models (of the kind described in section 2.2) that allow us to infer properties about extreme events that have not yet occurred. Statistical models that handle extremes are given by **Extreme Value Theory** (EVT), a branch of statistics dealing with this very question. A cornerstone of EVT is a limit theorem that states a limiting distribution for sampled maxima of i.i.d. random variables as the sample size goes to infinity. This theorem is analog to the central limit theorem, except that it operates for maxima instead of averages. The limiting distribution is always one of two families of distributions. The family depends on the approach used to define an event as *extreme*: for the approach known as "block maxima," the data converges to the **Generalized Extreme Value** distribution; for the one known as "Point-Over-Threshold," it converges to the **Generalized Pareto Distribution**. In any case, the result for both approaches is a distribution whose parameters $\boldsymbol{\theta}$ are inferred using the techniques described in section (2.3).

What follows is an overview of the univariate EVT approach used throughout this thesis: block maxima. This is followed by a description of the GEV distribution and one of its application-specific derivatives, known as the d-GEV. An explanation of the multivariate EVT approach needed for the spatial extremes is left for a later section of this chapter. Point-over-Threshold and the GPD distribution will not be covered, as they were not used for this thesis; for more information about PoT methods, the reader is referred to (Coles, 2001) and (Ribatet et al., 2016).

3.3.1 Block Maxima and their limiting distribution

The first approach to define extreme events is to take the observation with the largest magnitude (i.e., the maxima) over a certain fixed time length. This is then repeated without sampling from the same timestep more than once for every time interval with the same length. The different time intervals are known as blocks – this is why the approach is known as **block maxima**. Let $X_1, ..., X_n$ be a sequence of i.i.d. random variables that follow a common distribution F. Each X_i represents the values of the observed quantity measured at a regular interval for n timesteps. For example, if X_i represents the sea-level depth measured daily, $n = 30$ would correspond to a block

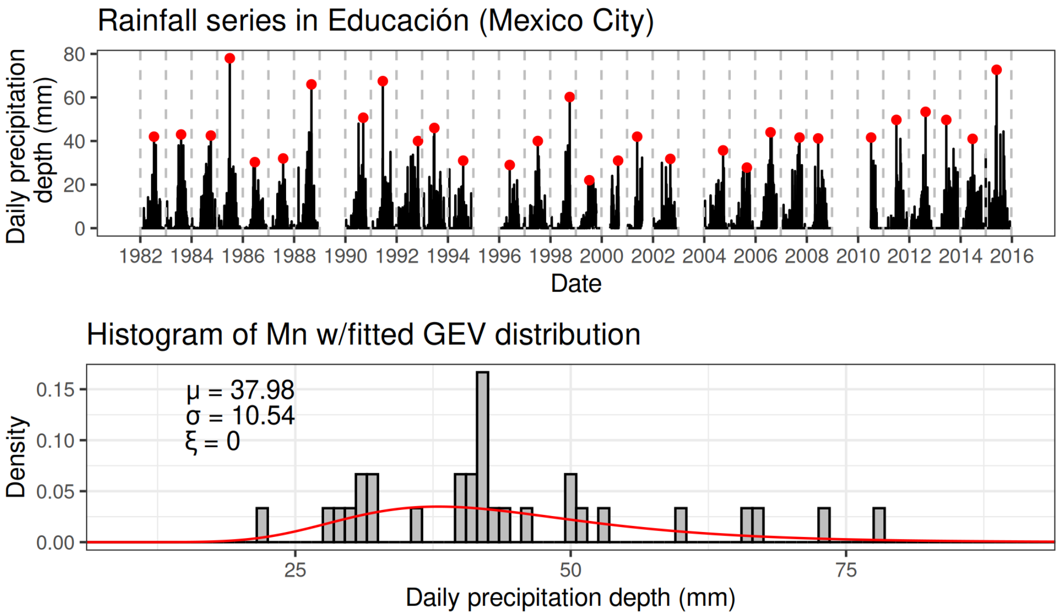

FIGURE 3.9: Top: Example of yearly block maxima for a daily-measured rainfall series. The dashed lines represent each block (n = 1 year), and the red dots indicate the value of the maxima for the given block. Bottom: Histogram of the resulting M_n block maxima from the example above, with the fitted GEV distribution superposed as a red curve. The MLE-estimated values of the GEV parameters are also indicated. Data source: (CLICOM-SMN, 2018)

size of a month. The block maxima M_n is then obtained via

$$M_n = \max\{X_1, ..., X_n\}. \tag{3.4}$$

From this equation, we can see that the choice of n will result in different sizes of the blocks used for block maxima. Furthermore, it is easy to show that $\Pr[M_n \leq z] = F^n$.

For convenience, it is common to name the block maxima according to the block size used. For example, yearly block maxima correspond to blocks spanning a year[2]. The top section of Fig. 3.9 shows the resulting yearly block maxima (red dots) for accumulated daily precipitation in Mexico City. Note that the actual number of elements used in each block depends on the time resolution of the observations, while the resulting maxima in M_n are a function only of the length of the complete series and the block size n. The applications with rainfall extremes described in later chapters work exclusively with yearly and monthly block maxima.

Once the M_n block maxima have been constructed, the goal of EVT is to find an appropriate distribution to describe M_n. Finding such a distribution is non-trivial, as estimating the exact distribution F (and by consequence, F^n) from data is typically unachievable for extreme data. However, it is possible to *approximate* F^n by finding a parametrical distribution G that approximates it when using the extreme data. Finding such a distribution is done by considering the limiting distribution of the adequately rescaled maxima M_n as the block size n tends to infinity, analog to the central limit theorem, but for maxima.

[2] Yearly and monthly block maxima do not always have the same length, as some years/months have differing amounts of days. However, this difference is usually too small to make a difference in the resulting distribution, so it is ignored.

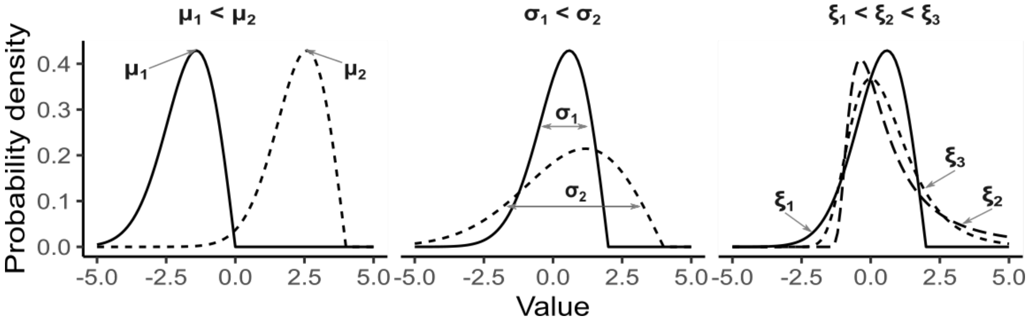

FIGURE 3.10: Influence of the GEV parameters in the resulting probability density function. Each plot shows how the change in one of the parameters changes the distribution shape.

The limiting distribution G of M_n is given by the theorem known as the **Fisher-Tippett-Gnedenko theorem**. This theorem states that if there exists a sequence of constants $a_n > 0$ and $b_n \in \mathbb{R}$ such that

$$\lim_{n\to\infty} \Pr\left\{\frac{M_n - b_n}{a_n} \leq z\right\} = G(z) \tag{3.5}$$

where G is non-degenerate, then the only possible limiting distribution G, up to location-scale transformation, is the **Generalized Extreme Value** (GEV) distribution.

Following Coles (2001), the GEV distribution with location μ, scale σ, and shape ξ parameters has the following distribution function:

$$\Pr[M_n \leq z] = G(x \mid \boldsymbol{\theta}) = \begin{cases} \exp\left\{-\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]_+^{-1/\xi}\right\} & \xi \neq 0, \\ \exp\left(-\frac{x-\mu}{\sigma}\right) & \xi = 0 \end{cases} \tag{3.6}$$

where $x_+ = \max(x, 0)$, and $\boldsymbol{\theta} = (\mu, \sigma, \xi)$. These parameters are restricted to $\mu \in \mathbb{R}, \sigma > 0$ and $\xi \in \mathbb{R}$. The GEV distribution encompasses three families of distributions, controlled by the shape parameter ξ. The Gumbel distribution occurs when $\lim_{\xi\to 0}$, the Fréchet distribution when $\xi > 0$, and the Weibull distribution when $\xi < 0$. Each of the three distribution families has historical developments and applications. An example of the GEV fitted to block maxima is given in the bottom part of Fig. 3.9

The GEV has the following density function (Dipak et al., 2016) (obtained by derivating eq. (3.6)):

$$f(x) = \frac{1}{\sigma} t^{\xi+1}(x) \exp[-t(x)], \tag{3.7}$$

where

$$t(x) = \begin{cases} \left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]_+^{-1/\xi} & \xi \neq 0, \\ \exp\left(\frac{x-\mu}{\sigma}\right) & \xi = 0. \end{cases}$$

The three parameters of the GEV distribution control different aspects of the resulting distribution, as seen in Fig. 3.10. The location parameter μ controls the position in the real line where the bulk of probability is located; the scale parameter σ controls how wide it is. For modeling purposes, the most important one is the shape parameter ξ (also known as the Extreme Value Index), which controls how the tails behave. For our stated goal of finding out the magnitude and frequency of future extreme events, the region of interest of the GEV distribution is the far-right side of the distribution's tail. This region is highly influenced by ξ, so finding an appropriate value for this parameter is crucial for any EVT application. Unfortunately, this parameter

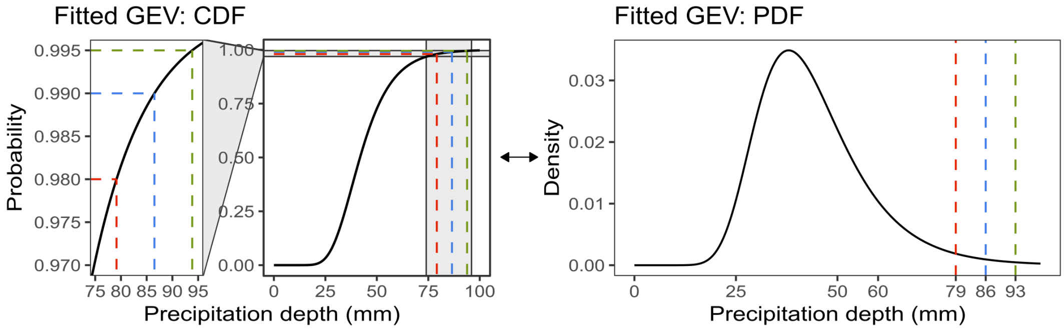

FIGURE 3.11: Resulting CDF and PDF from the fitted GEV distribution shown in Fig. 3.9. The color lines show the (0.98,0.99,0.995) probabilities of non-exceedance and their corresponding return levels. Data source: (CLICOM-SMN, 2018)

is also the hardest to model, as it is highly sensitive to the design of the inference method (Dipak et al., 2016).

Max-stability property

An essential consequence of the Fisher-Tippet-Gnedenko is the property known as **max-stability**. Following Coles (2001), a distribution G is said to be max-stable if, for every $n = 2, 3, ...,$ there are constants $\alpha_n \geq 0$ and β_n such that

$$G^n(\alpha_n z + \beta_n) = G(z). \tag{3.8}$$

However, we know that G^n is the distribution of $M_n = \max\{X_1, ..., X_n\}$, where M_n is composed of several random variables, each with distribution G. Therefore, for M_n maxima that follow the distribution G, the max-stability property implies that the sample maxima of M_n will also be G distributed, albeit with a change in scale and location. For the univariate case, it turns out that the GEV distribution is always max-stable.

The max-stable property is particularly important for multivariate extremes, where it will be the basis for a special stochastic process used to model extremes: the max-stable process. In a nutshell, the max-stability property is a desirable property for any model that deals with extreme-valued data.

3.3.2 Predictions from the GEV distribution

The goal of the studies described in this thesis is to find an appropriate statistical model for extreme rainfall events. We have seen that the GEV distribution is a theoretically-sound model for modeling the distribution of rainfall block maxima, which represent what we consider as extreme events. This means that once a precipitation series with enough records has been observed, the modeling problem is reduced to estimating values for the parameters (μ, σ, ξ) of the GEV distribution that best describe the observations. This estimation can accomplished by using the inference methods like MLE of Bayesian inference described in section 2.3. Other methods, like L-moments or the method of moments are also possible, but they will not be explored in this thesis. See Hosking et al. (1997) for an example of the L-moment method used for fitting the GEV distribution.

After fitting the parameters of the GEV distribution, the interest is to get information about what magnitude of the events to expect with certain probabilities in the future. Such information about possible future events based on the previous observations is given by the inverse of the distribution function in eq. (3.6), which is known as the quantile function:

$$Q(u) = \begin{cases} \mu + \frac{\sigma}{\xi}[(-\log u)^{-\xi} - 1] & \xi \neq 0, \\ \mu - \sigma \log(-\log u) & \xi = 0. \end{cases} \tag{3.9}$$

For this equation, it is enough to plug in the estimated values of (μ, σ, ξ) with the desired probability of non-exceedance p to get back the expected magnitude (typically in mm). This amount is commonly known as the **return level**, which has a probability of being exceeded of $(1 - p)$. An example of the procedure used to get return levels is seen in the color lines of Fig. 3.11: one starts on the Y-axis of the CDF (by choosing a probability of non-exceedance) and then notes the point in the X-axis where the CDF intersects the probability p. The resulting values of the x-axis are the return levels. This procedure is summarized by using the quantile function.

EVT studies are typically interested in very high non-exceedance probabilities (e.g., 0.9, 0.95, 0.99) located in a region of the distribution where (mostly) no observations exist. In other words, we want to extend the information from the magnitude and frequency of observed events to make assertions about the previously unobserved magnitude and frequency of future events. This is different from other branches of statistics, where the interest is in the central tendencies of the distribution. Because of this, EVT studies entail a "leap of faith" of sorts, where it is assumed that information from the bulk of the distribution is sufficient to get accurate information about the (unobserved) behavior of the far-right tail of the distribution. There is no theoretical justification for this assumption, but no more credible alternative has been proposed.

If the block size for M_n is of one year per block, p is then the probability that the magnitude z_p will (in the long run) not be exceeded every year [3]. This idea has given way to the creation of the quantity known as **return periods**. A return period is a conceptual device used to simplify the communication of probabilities of non-exceedance and return levels to a more general audience, but have been criticized for being easy to misinterpret and manipulate in recent years (Serinaldi, 2015). Return periods are obtained using the following expression:

$$\mathrm{RP}(z_p) = \frac{1}{\omega(1-p)}, \tag{3.10}$$

where z_p is the return level at probability of non-exceedance p, and ω is the sampling frequency of the data (for yearly block maxima, $\omega = 1$). Return periods are given in years, so that one can talk about the "1-in-RP year event". For example, the corresponding return period for $p = 0.99$ is $1/(1 - 0.99) = 100$ years. Therefore, the corresponding return level $z_{p=0.99}$ can be said to have a 100-year return period, meaning that, *on average*, it is expected that the $z_{p=0.99}$ level is exceeded *at least once* in 100 years. This does *not* mean, however, that the return level $z_{p=0.99}$ will be exceeded only once every 100 years nor that after exceeding the level, the next event will occur exactly in 100 years. Furthermore, return periods are valid only when assuming stationarity, an assumption that no longer holds in the age of climate change. Therefore, we avoid using return periods as much as possible in this thesis, opting instead for return levels with probabilities of non-exceedance.

[3] This statement assumes that the distribution is stationary, a claim that is in several studies increasingly disputed (Ganguli et al., 2017).

3.3.3 Special topics for EVT

Several special topics from EVT needed for this thesis are now described. First, a description is given of the duration-dependent GEV used in specific applications when the duration of rainfall is of interest. Then, an overview of using Bayesian methods for EVT will be discussed.

The duration-dependent GEV distribution and IDF curves

The societal impact of extreme rainfall is a function not only of magnitude, but also of duration. This is because hydrological infrastructure like storm drains or levees respond differently to events of different durations. For example, a typical storm drain should be able to handle 10 mm of rainfall that falls over 1 hour with no difficulty, but it would likely fail if 10 mm falls in 10 minutes. Therefore, different probabilities of non-exceedance correspond to different rainfall durations for the same amount of rainfall.

Depending on the specific application, rainfall can be expressed in either depth or intensity. Depth is simply the absolute amount of rainfall, measured in mm. On the other hand, intensity is the relative amount of rainfall over a certain period: commonly used periods are hour ($\mathrm{mm\,h^{-1}}$) or day ($\mathrm{mm\,day^{-1}}$). For example, 10 mm of rainfall measured in 10 minutes would have a depth of 10 mm and an intensity of $60\,\mathrm{mm\,h^{-1}}$. The transformation between the two values is straightforward so that no information loss occurs when using one or the other. For this thesis, I will exclusively use intensity in $\mathrm{mm\,h^{-1}}$.

From the definition of the GEV distribution in eq. (3.6), it can be seen that no information about the duration of extreme rainfall is included. However, extending the expression to account for the differences in the probability of non-exceedance for different durations would be a significant improvement for hydrological applications. This extension to the GEV model requires the specification of a relationship between intensity (or depth) and duration, as this information would allow us to include the duration d as an extra "random variable" of sorts in the GEV formulation. In his famous work, Koutsoyiannis et al. (1998) proposed the following relationship between intensity i and duration d:

$$i = \frac{\omega}{(d+\nu)^{\eta}}, \tag{3.11}$$

where ω,ν,η are non-negative coefficients. For any two return levels z_{p1} and z_{p2}, where $p_1 < p_2$, there exists the additional restrictions: $\nu_1 = \nu_2 = \nu = 0$, $0 < \eta_1 = \eta_2 = \eta < 1$, and $\omega_1 > \omega_2 > 0$. These restrictions ensure that the return levels will not intersect, meaning that the resulting intensity values are *consistent*.

Using the relationship between duration and intensity in eq. (3.11), Ritschel et al. (2017) and Ulrich et al. (2020) developed the **duration-dependent GEV distribution** (d-GEV). This distribution incorporates the restrictions mentioned above. The d-GEV is used to model the intensity $i(d)$ for a certain duration d and has the following distribution function:

$$G(x \mid \boldsymbol{\theta}) = \exp\left[-\left(1+\xi\left(\frac{x}{\sigma(d)}-\tilde{\mu}\right)\right)^{-1/\xi}\right], \tag{3.12}$$

where $\sigma(d) = \sigma_0/(d+\nu)^{\eta}$ is the duration-dependent scale parameter, and $\tilde{\mu} = \mu(d)/\sigma(d)$ is the modified location parameter. In this case $\boldsymbol{\theta} = (\mu(d), \sigma(d), \xi, \eta, \nu)$, where now μ and σ are functions of the duration $(\mu(d), \sigma(d))$. The shape parameter is held constant across all durations.

The distribution function for the d-GEV in (3.12) requires block maxima x that come from different durations. Typically, rainfall data provided by weather services come only as rainfall depth measured in 1-minute, 1-hour, or 24-hour intervals. Therefore, a transformation of those rainfall series to different durations is required to use the d-GEV. This transformation is done by aggregating the initial series from their measurement frequency $d = T$ to longer durations $d = nT$. Given the measured average hourly intensity $\zeta_{d=T,t}$ for the given time resolution of $d = T$ at time t, the n-hour aggregated series $\boldsymbol{\zeta}_{d=n}$ is obtained using

$$\zeta_{d=n,t} = \frac{1}{n} \sum_{i=0}^{n-1} \zeta_{d=T,t-i}, \tag{3.13}$$

which can be seen as a moving average with a time window of n time-units. Using eq. (3.13) allows us to get the aggregated intensity of any desired duration d, as long as this duration is a multiple of the measuring frequency. In practice, hydrological studies use an upper limit for d of 120 h. Most real-world applications are limited to shorter durations, typically from the range of 1 h to 24 h. After the aggregation process, the d-GEV requires block maxima to be taken from the aggregated intensity series $\boldsymbol{\zeta}_{d=n}$. The aggregation process was automatized in the R-package IDF (Ulrich et al., 2020).

After aggregating the rainfall data, taking the block maxima and using it to estimate the $(\mu(d), \sigma(d), \xi)$ d-GEV parameters, it is straightforward to calculate the return level $z_{d,T}$ for any arbitrary duration using

$$z_{d,T} = \mu(d) + \frac{\sigma(d)}{\xi} \left[\left(-\log\left(1 - \frac{1}{T}\right) \right)^{-\xi} - 1 \right]. \tag{3.14}$$

A common strategy is to use eq. (3.14) to estimate the return level for a range of durations with three or four fixed non-exceedance probabilities. The resulting return levels are then plotted simultaneously using a log-log scale, drawing a curve that goes over return levels with the same p. Each curve is known as a **Intensity-Duration-Frequency curve**, or IDF for short. IDF curves are commonly-used hydrological tools, especially for hydrological engineering. An example of an IDF curve is seen in Fig. 4.4.

Bayesian inference for the GEV distribution

As stated before, the problem of univariate EVT modeling is typically reduced to inferring appropriate values for the GEV distribution from the observations. MLE estimation provides a robust and theoretically-sound method for estimating the parameters. Unfortunately, the very nature of EVT models can lead to well-known computational issues with MLE estimates, especially when $\xi < -0.5$. Furthermore, the propagation of uncertainty from the parameters to the estimated return levels relies mainly in asymptotic assumptions that do not always hold (for more information on this subject, see Coles (2001)). It is thus desirable to explore other methods that are also based on the likelihood but possibly avoid the common pitfalls from MLE.

A valid alternative for estimating the GEV parameters is Bayesian inference, which avoids some computational pitfalls of MLE and gives a straightforward way to propagate the uncertainty to the return levels. For the Bayesian approach, the likelihood term $p(D \mid \boldsymbol{\theta})$ from Bayes' Rule is given by the GEV density (eq. (3.3.1)). Additionally, a prior distribution has to be given for the $\boldsymbol{\theta}$ parameters. A benefit of using prior distributions is that it decreases the uncertainty inherent when using small datasets by incorporating previous knowledge. However, the prior acts as a double-edged sword, as models using small datasets may be very sensitive to the choice of prior.

A typical design choice for Bayesian models is to use prior distributions $p(\boldsymbol{\theta})$ that have either an enormous variance (for example, Normal$(0, 10000)$) or are completely flat (e.g., Uniform$(-10000, 10000)$). These priors are commonly known as **noninformative priors** and are seen as an "objective"[4] way of doing Bayesian inference. This choice is especially relevant when dealing with large datasets; in this case the choice of $p(\boldsymbol{\theta})$ tends to be inconsequential, as the likelihood overwhelms the prior. For EVT models, however, the datasets are commonly too small for the likelihood to overwhelm the prior, so a careful choice of $p(\boldsymbol{\theta})$ is required.

Prior distributions can either be constructed by using previous knowledge *before* seeing the observations, or by using the observations to make a more informed prior. From a strictly Bayesian point of view, the latter approach is strictly invalid, as the prior must be set before seeing the observations. However, using data-driven priors has yielded positive results in many real-world applications. Some examples of the data-driven approach include the use of Jeffreys' prior (based on Fisher's information matrix) and the maximal data information (MDI) prior. Jeffreys' prior and the MDI are valid priors for the GEV parameters, as long as an adjustment is made to avoid problems with the shape parameter. On the other hand, using previous knowledge to build a prior is done via the process known as **prior elicitation**. Elicitation entails that the previous knowledge is expressed in the form of a proper distribution function; Many resources exist in the literature for this procedure (Mikkola et al., 2021).

The simplest prior elicitation for the GEV model is to assume that the parameters (μ, σ, ξ) are independent and assign a distribution with a large variance to each of them, using previous knowledge for the location hyperparameter of each distribution. This approach is used for the priors in the second study described in this thesis. Other approaches include using quantiles to approximate priors for the GEV parameters. For more information on this topic, see Stephenson (2016).

After choosing a prior, Bayesian inference for the GEV parameters requires dealing with the integral of Bayes' rule. This is typically solved by approximating the posterior distribution using the MCMC methods described in section 2.3.2. Fortunately, it has been seen that MCMC-based inference works just as well for the parameters of the GEV distribution as for other distributions. The second study of this thesis makes use of MCMC methods for inference of the GEV parameters.

3.4 Setting of the spatial modeling problem

Spatial data typically consists of a collection of measurements in several locations $X(s_i)$, where s_i denotes the location of one of N-observations inside the domain $\mathcal{S}$. For the classical setting, there is only one observation x in each location, but each location could also contain many observations. Figure 3.12 shows an example of an imaginary spatial dataset. When modeling rainfall, the positions s_i represent the location of the rain gauges, which can be seen as fixed in time and space. A distance measure h can be defined for the different s_i locations, typically taken to be the euclidean distance.

Two main questions typically arise from this dataset:

[4] I am of the opinion that no true objective experiment-design exists. Not only is the choice of a non-informative prior a subjective decision, but most importantly, the choice of a likelihood model is a subjective choice, with much larger repercussions than the choice of prior.

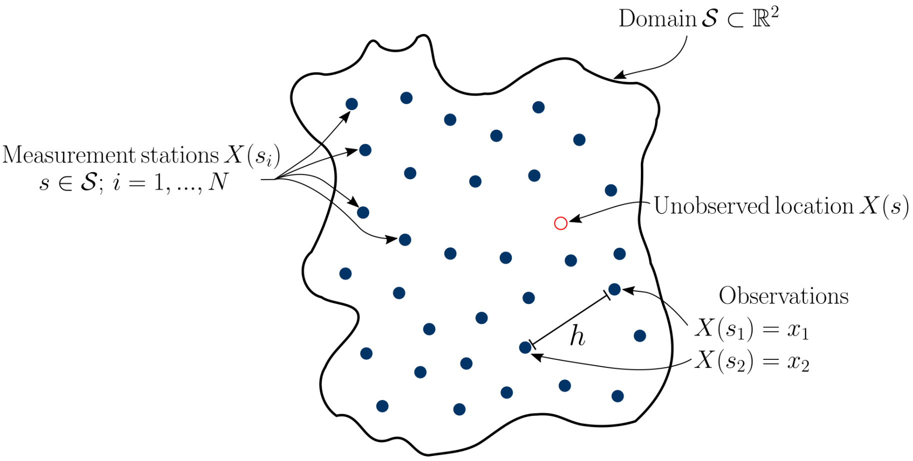
Domain $\mathcal{S} \subset \mathbb{R}^2$
Measurement stations $X(s_i)$
$s \in \mathcal{S}$; $i = 1, ..., N$
Unobserved location $X(s)$
Observations
$X(s_1) = x_1$
$X(s_2) = x_2$
h

FIGURE 3.12: Typical data setting for a spatial problem.

- What are the possible X values for some unobserved location s inside the domain? and
- What are the possible future values of X in observed locations s_i?.

The first question, predicting unobserved locations, involves using information from existing stations to interpolate. This interpolation results in estimates for unobserved locations that ideally should share some properties from the observed locations. The second question, predicting future values, is the same as the one discussed in section 2.5. In practice, we could just use univariate statistical methods to get individual predictions for each observed location. However, in this case, we would like to somehow *pool* the information from all observed locations in a way that improves the resulting estimates. Both problems can be solved by applying Tobler's famous first law of geography:

> "Everything is related to everything else, but near things are more related than distant things." -Tobler, 1970

What this means to our particular problem is that we expect to see a high similarity for the observations between two locations (s, s') when the distance is (relatively) short. The opposite is also true: when the distance between the locations is large, we expect to see less similarity. The idea is then to use this similarity to improve the model estimates: we want to find a family of models that gives similar estimates to nearby locations.

The first step towards this goal is to define what *similarity* means and how to measure it. Defining similarity in space is achieved using the **covariance**, a measure of how much two random variables change together. For two random variables, X and Y, the covariance is the difference between the expected product and the product of the expectations: $\text{cov}(X,Y) = \text{E}(XY) - \text{E}(X)\text{E}(Y)$. The resulting sign of $\text{cov}(X,Y)$ indicates the direction of the linear relationship between the variables: a positive sign means that if X increases, Y also increases; a negative sign means that if X increases, Y decreases. To interpret the magnitude of the covariance it is necessary to normalize it; the normalized covariance is known as the linear correlation coefficient. The variance can be seen as a particular case of covariance, as $\text{var}(X) = \text{cov}(X,X) = \text{E}(X^2) - \text{E}(X)^2$.

To apply the first law of geography to our model, we need to design a function that returns the value of the covariance for any given two locations inside the domain. This function is known as the **covariance function** $C(X, Y)$. This function essentially encodes the notion of nearness and similarity for our model. The following section explores the branch of statistics known as geostatistics, which proposes models to characterize $X(s)$ in terms of its mean and covariance function. Once a covariance function has been found, geostatistical methods can solve the questions posed earlier in this section.

3.5 Classical Geostatistics

Initially developed 70 years ago for the mining industry, **geostatistics** encompasses a collection of statistical models and methods that focus on spatial and spatiotemporal datasets like the one described in Fig. 3.12. The main goal of geostatistics is to perform interpolation in a spatial domain using Tobler's first law of geography; effectively, the idea is to pool information in space to "borrow the strength" from the spatial dependence shown by the data.

A basic geostatistical model claims that for every location s in the domain $\mathcal{S}$, the observed $X(s)$ can be formulated as

$$X(s) = \alpha(s) + e(s), \tag{3.15}$$

where $\alpha(s)$ denotes the mean value of X at location s, and $\{e(s)\}$ is a zero-mean stochastic process $\{e(s) : s \in \mathcal{S}\}$. Note that for this simple model, $X(s)$ contains only one observation per location. The first term represents the large-scale variability of $X(s)$. The second term represents the local/small-scale variability of $X(s)$. The idea of spatial similarity/nearness enter into the model in the second term, and as such, $\{e(s)\}$ is sometimes denoted as the **spatial dependence**. A useful analogy can be made thinking about the distinction between climate and weather: The mean $\alpha(s)$ acts as the "climate" component, which varies in space but at very large scales, and the stochastic component $\{e(s)\}$ is the "weather," composed by many local events which induce dependence for points that are near each other.

A common strategy to model the mean component $\alpha(s)$ is to use the regression model

$$\alpha(s) = \boldsymbol{A}(s)\boldsymbol{\beta}. \tag{3.16}$$

Here $\boldsymbol{A}(s)$ represents the matrix of covariates unique for each location s, and $\boldsymbol{\beta}$ denotes the vector of regression coefficients. In eq. (3.16) many different spatial covariates could be used. For example, if $\boldsymbol{A}(s)$ is set to be the latitude, longitude, and altitude for each location, the resulting regression model is obtained:

$$\alpha(s) = \beta_0 + \beta_1 \mathrm{lat} + \beta_2 \mathrm{lon} + \beta_3 \mathrm{alt}. \tag{3.17}$$

In geostatistics, a model like this one that uses only spatial coordinates for regression of the mean is known as the **response surface model**. The studies described later make exclusive use of response surfaces. However, the choice of covariates is not restricted to be only coordinates. For example, one could also choose as covariates the distance of a point to the nearest coast, mean decadal temperature at each position, etc.

As mentioned before, the stochastic process $\{e(s) : s \in \mathcal{S}\}$ in eq. (3.15) accounts for the spatial variability that is not explained by the mean α. The notion of pooling information in space enters the process $\{e(s)\}$ via the **covariance function**. Given the stochastic process $\{e(s)\}$, the covariance function is defined for two locations s and s' as

$$C(s, s') \equiv \mathrm{cov}(e(s), e(s')). \tag{3.18}$$

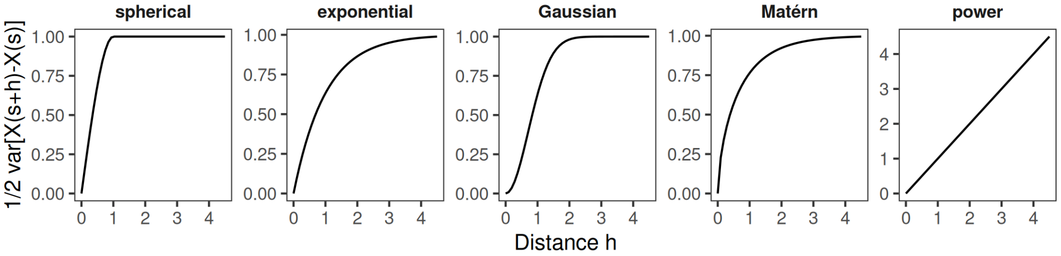

FIGURE 3.13: Example variogram for each of the five main families. For all variograms the following parameters where used: (Nugget = 0, Sill = 1, Range = 1). The Matérn variogram uses $k = 0.2$

For the simple model defined in eq. (3.15), two assumptions are widely made for the covariance function of the zero-mean process $\{e(s)\}$. These are the second-order stationarity and isotropy.

The assumption of **second-order stationarity** for the covariance function implies that the covariance does not depend on the location. That is, given any distance h,

$$C(s, s') = C(s + h, s + h') = \text{cov}(e(s + h), e(s' + h)).$$

The other assumption, **isotropy**, implies that covariance is a function of the distance h between locations (i.e., it is not a function of the direction). Isotropy then implies that

$$C(s, s + h) = C(h) = \text{cov}(e(s), e(s + h)).$$

While this assumption can be unrealistic for many rainfall fields, it is common to assume isotropy as a first approach.

The second-order stationary and isotropic zero-mean stochastic process $\{e(s) : s \in \mathcal{S}\}$ can be fully characterized by its covariance function $C(h)$. Note that second-order stationary covariance functions always decrease with increasing distance. The most common model in geostatistics to incorporate second-order stationary and isotropic stochastic process is to assume that $e(s)$ is a zero-mean Gaussian process. From eq. (3.3) it becomes apparent that doing so would mean that $X(s)$ is also a Gaussian process with mean $\alpha(s)$. We will see later that *kriging*, the main interpolation tool of geostatistics, is simply the application of this Gaussian process approach.

3.5.1 The (semi-)variogram

For environmental modeling, second-order stationarity can be rather strict. Thus, it is also common to assume **intrinsic stationarity**, a more relaxed assumption than second-order stationarity. Under this assumption, the variance (not the covariance!) of two different locations depends only on the distance h and not the location. In this case, the covariance function $C(\cdot, \cdot)$ is not defined; instead, similarity in space is given by the **(semi-)variogram**[5]. The (semi-)variogram $\gamma(h)$ is defined as

$$\gamma(h) = \frac{1}{2}\text{var}\left[e(s + h) - e(s)\right], \qquad (s, s + h) \in \mathcal{X}. \tag{3.19}$$

[5] The notation surrounding the term variogram/semivariogram is complicated. Some authors denote expression (3.19) as the semivariogram due to the 1/2 term in the formula. On the other hand, some authors call $\gamma(h)$ simply as the variogram (Bachmaier et al., 2011). For this thesis, I will exclusively use the term variogram, as defined in eq. (3.19).

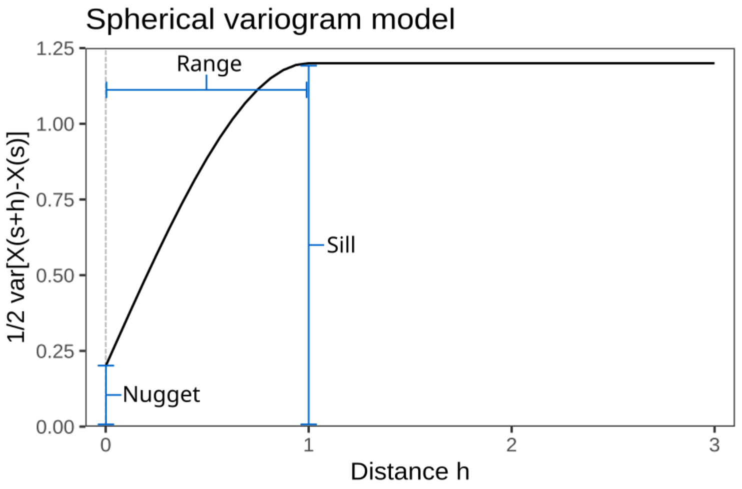

FIGURE 3.14: Example parametric variogram showing the three main characteristics: Nugget, Sill, and Range. For this particular variogram of the spherical family, nugget = 0.2, sill = 1, range = 1.

The quantity $1/2 \cdot \mathrm{var}\left[e(s+h) - e(s)\right]$ is sometimes known as **semivariance**, although some authors have objected to this term, preferring the term gammavariance (Bachmaier et al., 2011).

In the second-order stationary case, the variogram can be transformed into the covariance function via

$$\gamma(h) = C(0) - C(h),$$

where $C(0) = \mathrm{var}(e(s))$ represents an upper bound to the variogram. However, for the intrinsic stationary case, the covariance function does not exist; only the variogram does. This restriction means that transforming from the covariance function to the variogram is always possible, but not the inverse.

For statistical inference purposes, the variogram function can be expressed as different parametric families with well-defined parameters. These parametric families were designed to meet the necessary conditions to be valid covariance/variogram functions – just as parametric distributions were designed to be valid probability distributions. These functions are known as the **theoretical variogram**. Figure 3.13 shows the five most commonly used families of theoretical variograms (Spherical, Exponential, Gaussian, Matérn, and Power), described in Gelfand et al. (2010).

Figure 3.14 shows an example of a theoretical variogram with its three defining characteristics: the nugget, range, and sill. The **nugget** represents the variability inherent to $X(s)$ that is too small to be represented by the sampling interval (e.g. the measurement error of the device). The **range**, typically denoted by ρ, is the distance where the model "flattens out". This can be seen as a sort of measure of the distance at which spatial dependence is still relevant. Lastly, the **sill** is the upper limit of $\gamma(h)$ reached when the distance equals the range. The sill exists only for second-order stationary conditions, as the upper bound of the variogram is only defined for that case.

Once an appropiate parametric family for the theoretical variogram has been chosen for the problem at hand, the next step is to find the values of the variogram parameters that best describes the observed semivariance. Given a spatial dataset like the one described in section 3.4, the first step is to compute the semivariance for all $(N(N-1))/2$ possible pairs of $X(s_i)$. The resulting plot of all the semivariances

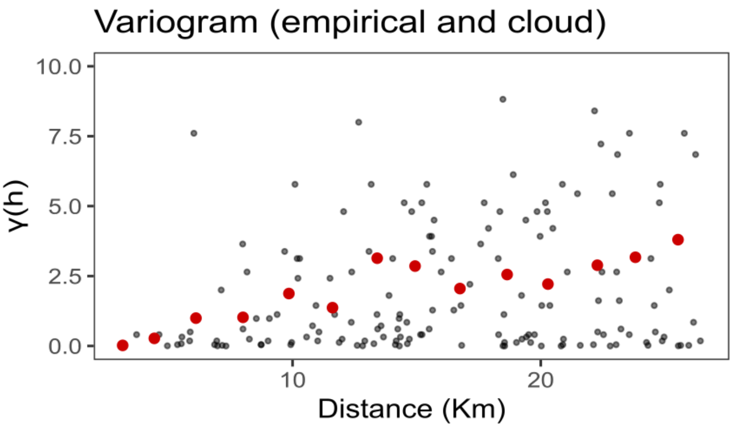

FIGURE 3.15: Example of a variogram cloud (black dots) and the empirical variogram (red dots). Data for this example was a dataset of 24 weather stations with hourly 2m temperature values in Mexico City (Source: aire.cdmx.gob.mx)

according to the distance h between locations is known as the **variogram cloud**. Figure 3.15 shows an example of a variogram cloud. The variogram cloud is an useful tool for initial exploration, but it commonly is too spread out to be able to be used to discern a theoretical model. This is why the cloud is then separated into bins according to h; an average value of the semivariance is then computed for each bin. The plot of the resulting binned means of the semivariance is known as the **empirical variogram**. Note that the operation of taking binned means assumes that γ is isotropic. The empirical variogram is the spatial analog of the histogram, and can be used to judge what parametric family of variograms should be used. An example of an empirical variogram is also seen in Fig. 3.15.

The fitted theoretical variogram acts as the covariance needed for the zero-mean Gaussian Process $\{e(s)\}$ of the model given in eq. (3.15). This function encodes the spatial dependence, so that the final model will borrow information in space to predict unobserved values. Now we need a procedure to go from the covariance of $\{e(s)\}$ to actual interpolation. Within the context of classical geostatistics, this is known as kriging.

3.5.2 Spatial interpolation: Kriging

One of the main goals of spatial models is to pool the existing information in space to interpolate to unobserved locations. In classical geostatistics, the base model is the one described in eq. (3.15), which divides $X(s)$ into a mean function and a zero-mean stochastic process. A natural choice is to set $e(s)$ to be a zero-mean Gaussian process. Thanks to the the additive property of GPs (eq. (3.3)), $X(s)$ becomes in that case a Gaussian process with mean $\alpha(s)$. The model is then given by

$$X(s) = \alpha(s) + e(s) \tag{3.20}$$

$$e(s) \sim \mathrm{GP}(0, \boldsymbol{K}), \tag{3.21}$$

where just as before, $\boldsymbol{K}$ is the covariance matrix, given by the covariance function. All that is needed now to interpolate is to find a way of obtaining the values of $e(s)$ for unobserved s.

The geostatistical model given by eq. (3.21) is called **universal kriging** (the same model is known as "Gaussian process regression" for non-spatial data). For universal

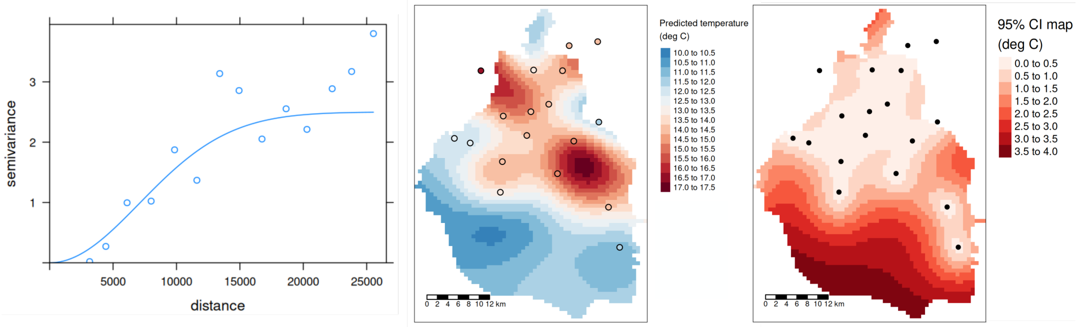

FIGURE 3.16: Example of predictions from kriging. Left: Empirical and Theoretical variogram. Center: Modeled temperature values using kriging. Right: 95% CIs for predictions. (Data source: aire.cdmx.gob.mx)

kriging the mean $\alpha(s)$ is modeled using the trend surface from eq. (3.17). Interpolation is then achieved with the following equation:

$$X(s) = \boldsymbol{A}(s)\boldsymbol{\beta} + \sum_{i=1}^{n} \lambda_i e(s_i), \tag{3.22}$$

where λ is the so-called vector of kriging weights, determined by the spatial dependence structure of the covariance function. These weights, as well as the β coefficients need to be fitted using the data. This is often done using the method of least squares.

Figure 3.16 shows an example of interpolated values for 2m air temperature using universal kriging. The observations (seen as points) consist of 24 air temperature measurements in the Mexico City metropolitan area, with an average distance of 16 km between stations. The variogram on the left shows a Gaussian theoretical variogram superposed with the empirical variogram, with a somewhat good fit (range = 10 km, sill = 2.5). The middle plot shows the predicted temperatures in a grid with resolution of 700 m, and the right plot shows the corresponding 95% CI derived from the variance of the estimation. Note that while the kriged temperatures gives values for regions in the south were no information exists, the uncertainty is rather large. Thus, one should be careful to interpolate outside of the region with information. The value of 10 km for the range parameter suggests that estimates outside of a 10 km radius from any observation should be taken with a lot of suspicion. Note that this example ignores the complex topography of Mexico City; therefore, this predicted temperature map should be seen only as informative for how kriging results look.

For modeling extreme rainfall in space, it is tempting to use classical geostatistics methods like universal kriging to, for example, interpolate the resulting return levels or the GEV parameters in space. However, as will be discussed in the next section, classical geostatistic methods are ill-fitted for extreme valued data. Therefore, an extension for extremes, called spatial extremes, needs to be introduced.

3.6 Spatial Extremes

The last section showed how the geostatistical methods pool information between stations to borrow the strength of spatial dependence, leading to improved estimates for unobserved locations. Unfortunately, the assumption of an underlying Gaussian Process renders most of these methods invalid for extreme-valued data like block maxima or threshold exceedances. The main problem is that the Gaussian distribution is not max-stable, making it a poor candidate for any EVT modeling. Furthermore, classical

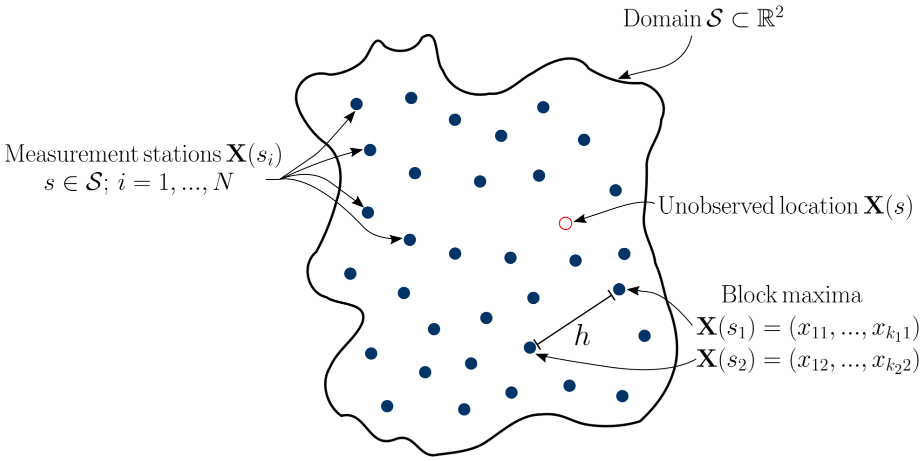
Domain $\mathcal{S} \subset \mathbb{R}^2$
Measurement stations $\mathbf{X}(s_i)$
$s \in \mathcal{S};\ i = 1, ..., N$
Unobserved location $\mathbf{X}(s)$
Block maxima
$\mathbf{X}(s_1) = (x_{11}, ..., x_{k_1 1})$
$\mathbf{X}(s_2) = (x_{12}, ..., x_{k_2 2})$
h

FIGURE 3.17: Typical data setting for an spatial extremes application.

geostatistics deal with central tendencies (like the mean), and little to no attention is given to the tail of the distribution, which is of central interest for extreme valued analysis. This last condition is also reflected in another important difference: geostatistical methods like kriging require only one observation per location, but for any meaningful EVT analysis, many observations from the same location are necessary. Nevertheless, we will see that under certain conditions, the central ideas from classical geostatistics (such as the covariance function or the variogram) can be extended to extremes. The resulting collection of methods that deal with extreme-valued data with a spatial component is usually classified under the umbrella term of **spatial extremes**.

Spatial extremes datasets typically consist of the measured series $\boldsymbol{X}(s_i)$, where s_i denotes the location of one of N-stations (e.g., rain gauges) inside the domain $\mathcal{S}$. Each measurement series x_{ki} contains $k = 1, ..., K_i$ block maxima, where the total number of K-maxima can differ between stations as a result of differing series lengths. Figure 3.17 shows an example of an imaginary spatial dataset. When modeling rainfall, the positions s_i represent the location of the rain gauges, which can be seen as fixed in time and space. A distance measure h can be defined for the different s_i locations, typically taken to be the euclidean distance. Using this kind of dataset, the central questions this time are:

1. How can we borrow strength from the spatial dependence to improve the marginal estimates of the GEV parameters for each location?

2. How do we use the existing observations to predict the GEV parameters of unobserved locations so that we account for the spatial dependence?

These questions generally deal with the joint distribution of the $\boldsymbol{X}(s)$ random vectors. This means spatial extremes deal with multivariate extreme valued distributions (i.e., multivariate max-stable distributions). Moreover, because any point s is assumed to have a marginal max-stable distribution, and the number of points $s \in \mathcal{S}$ is infinite, we require to extend the univariate methods of EVT to a setting with infinite dimensions. For this, the use of stochastic processes is ideal, as they operate on infinite dimensions.

A standard base assumption is that for every location s, $\boldsymbol{X}(s)$ is GEV distributed (eq. (3.6)), although this can vary depending on the specific method. To further simplify this problem, a widely used technique is to transform every margin to be unit-Fréchet distributed (i.e., GEV$(1, 1, 1)$). This transformation results in the simplified distribution function $G(z) = \exp(-1/z)$. There is no loss of generality in performing this transformation, as it is straightforward to transform the margins back to arbitrarily GEV-distributed margins. The transformation of the margins to unit-Fréchet is accomplished with the following equation:

$$z = \left[1 + \xi \left(\frac{x - \mu}{\sigma}\right)\right]_{+}^{1/\xi}, \tag{3.23}$$

where x is the original GEV-distributed data, and z is the unit-Fréchet distributed data.

Let $\boldsymbol{Z}(s) = \{Z_1, Z_2, ..., Z_D\}$ be a collection of random vectors corresponding to the unit-Fréchet transformed componentwise maxima $\boldsymbol{X}(s)$. Then, the limiting joint distribution of the $\boldsymbol{Z}(s)$ random vectors is given by:

$$\Pr[Z_1 \leq z_1, ..., Z_D \leq z_D] = \exp[-V(z_1, ..., z_D)],\ z_i > 0. \tag{3.24}$$

In this equation, $V(\cdot)$ is the **exponent measure**, defined by de Haan et al. (1977). The exponent measure has several constraints that ensure that the marginal distributions are unit-Fréchet and extend the max-stability property to higher dimensions. In principle, any function for $V(\cdot)$ that fulfills these conditions can be used to construct a valid max-stable multivariate distribution using eq. 3.24. Unfortunately, this means that no unique parametric form for $V(z)$ in the style of the GEV exists for multivariate distributions. However, several parametric forms have been proposed (Davison et al., 2015).

An alternative to the use of the exponent measure $V(z)$ is the use of the so-called **spectral measure**, typically denoted by $H(\omega)$. As is the case with the exponent measure, the spectral measure also determines the dependence structure of the random vector. This transformation can sometimes simplify the construction of valid functions (as described in the section on max-stable processes), but the same identification challenges remain. Mathematical details of both the exponent measure and spectral measure will be omitted, but the reader can consult de Haan et al. (1977) for more details.

Once a proper characterization of the dependence structure has been found (either by using $V(z)$ or $H(\omega)$), the two questions of spatial extremes can be solved using the joint distribution derived from the dependence measures. However, this method requires the transformation of the margins to unit-Fréchet. From eq. (3.23) it can be seen that this transformation requires the marginal distributions to be known a priori, as the GEV parameters are required inputs for the transformation. All of this suggests that the workflow of spatial extremes can be divided into two overarching goals:

1. estimate the marginal GEV parameters for each location to transform the data to unit-Fréchet, and

2. find an appropriate description of the spatial dependence of the transformed data using $V(z)$ or $H(\omega)$, which can be used to construct a joint distribution.

Regarding the two steps above, Cooley et al. (2012) make an essential distinction. They claim that the first step is akin to characterizing the overall tail variability in the domain, which comes from large-scale considerations. On the other hand, they see the second step as related to the local spatial effect, which results from large enough events that simultaneously hit several locations. Because this second step deals with the dependence left after the marginal transformation, they call it the

residual dependence. The residual dependence is what is captured by $V(z)$ or $H(\omega)$, and thus, is the central object of study for spatial extremes analyses.

An important observation is that, while the two steps described above appear to be sequential, they can be combined into a single overarching step by combining them in the likelihood. This is true for both the frequentist and Bayesian settings. This is explored in more detail in section 3.6.3.

The rest of this section is structured as follows: First, I discuss how the concept of similarity in the context of spatial extremes is measured. Secondly, a way to extend classical geostatistics to an infinite-dimensional setting with max-stability is discussed; this is where the prominent methodological powerhouse used in this thesis is introduced: the max-stable process. Additionally, a discussion of how to do inference for max-stable processes is given for the frequentist and Bayesian paradigms. The section then ends with a short discussion of alternative methods for modeling spatial extremes, namely the latent variable approach, copulas, and r-Pareto processes.

3.6.1 Extremal coefficient

Applying Tobler's first law of geography requires the definition of a similarity measure between the data from two locations. For classical geostatistics, this measure is the semivariance defined by the variogram (eq. (3.19)) for intrinsically stationary processes. However, these similarity measures do not work for extreme-valued datasets (i.e., block maxima or threshold exceedances), as the variogram focuses exclusively on the central tendency of the linear relationship between locations. For spatial extremes we require a similarity measure between the tails of the distribution. Therefore, we need to construct a function that assigns large (small) values when a large (small) value of a random variable co-occurs with a large (small) value of another random variable. Additionally, the measure should be valid for non-linear relationships and extendable to more than two dimensions.

For the latter requirement, several non-linear bivariate dependence measures exist (see Nelsen (2007) for more information). Examples include Kendall's Tau, Spearman's Rho, and Blomqvist's Beta. These three are measures of association (similar to Pearson's correlation coefficient) valid for non-linear relationships. However, while they help measuring the overall non-linear dependence between two random vectors, they do not focus on the tail of the distributions, which is the interest of EVT studies. Thus, we need a more specialized function to measure the so-called **tail dependence**.

As mentioned above, a complete characterization of the tail dependence (also known as the **extremal dependence**) between the components of a max-stable random vector is given by either the exponent measure function $V(z)$ or, equivalently, by a spectral measure like the Pickands dependence function $A(w)$. The problem is that direct computation of either of these functions is a formidable challenge, especially in high-dimensional settings. Therefore, a simpler summary coefficient that contains information about the extremal dependence without needing to fully specify it is preferred for most real-world applications.

Before discussing some of the widely used summary measures of extremal dependence, it is essential to distinguish between asymptotical tail dependence and independence. This is because the summary measures were designed to work only within one of the two regimes. Tail dependence/independence refers to the asymptotical tail behavior for regions in the far-right of the distributions: When the different random vectors still show a level of dependence for very large-valued regions, the data is said to be asymptotically dependent. On the contrary, when the data shows no discernible

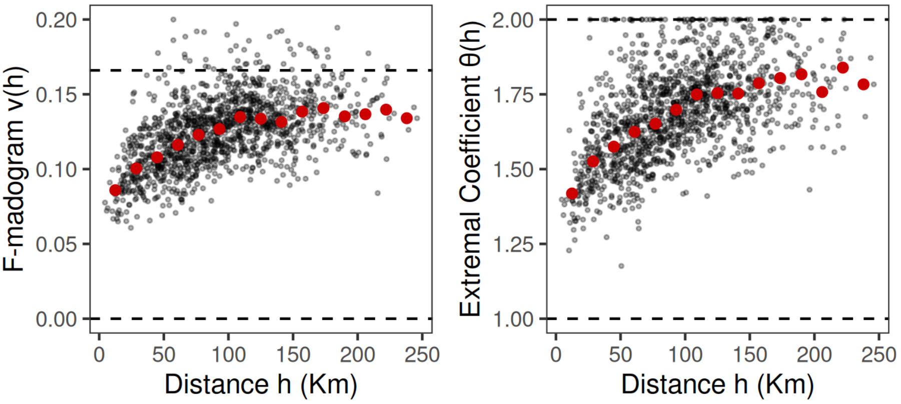

FIGURE 3.18: Comparison of the F-madogram (left) and the extremal coefficient (right) using the same dataset. The red dots represent the binned means from the underlying cloud. Dashed lines show the theoretical limit for both measures. Notice that the theoretical limit is violated several times for the F-madogram. The data source is explained in section 5.2.1

structure for very large-valued regions, it is said to be asymptotically independent. Using a summary measure of tail dependence for a mismatched asymptotical dependence regime will result in systematical errors. Most currently used applications of spatial extremes assume asymptotical tail dependence. Therefore, here I will exclusively focus on measures for the asymptotically dependent case. An alternative model for Spatial Extremes that can account for asymptotical independence is that of Wadsworth et al. (2019), which was applied to hourly rainfall data by Richards et al. (2021).

The first summary measure of tail dependence introduced here is very similar to the variogram: the **F-madogram** proposed by Cooley et al. (2006). For an isotropic and stationary stochastic process $\{Z(x)\}$, the F-madogram is defined by the following:

$$\nu_F(h) = \frac{1}{2}\mathrm{E}[F\{Z(x+h)\} - F\{Z(x)\}], \tag{3.25}$$

where $F(\cdot)$ denotes the cumulative distribution function, and $x, h \in \mathcal{X}$. The F-madogram is of particular usefulness for max-stable processes, defined in the next section. For simple max-stable processes, the F-madogram is well-defined, because the expected value $E[F\{Z(x)\}]$ is finite (which is not the case for the expected value $E[Z(x)]$ in the variogram). Plots of $\nu_F(h)$ against the distance h are very similar to plots of the variogram and have virtually the same interpretation. An example is shown in the left panel of Fig. 3.18. Furthermore, empirical estimates of the F-madogram can be obtained following the same procedure as the empirical variogram. $\nu_F(h)$ has lower and upper bounds of $[0, 1/6]$, indicating complete dependence and independence, respectively. However, these theoretical bounds can be violated, so that a correction is sometimes needed (Vettori et al., 2018).

An alternative summary measure of extremal dependence is the **extremal coefficient** θ proposed by Smith (1990b) [6]. The extremal coefficient θ is derived from the multivariate joint distribution, as $\Pr\{\max(X_1, ..., X_k) \leq z\} = F^{\theta}(x)$, where θ is the extremal coefficient. The theoretical bounds of θ are $1 \leq \theta \leq k$, where 1 denotes complete dependence and k complete independence. The extremal coefficient can be seen as “the effective number of independent random variables in the k-dimensional random vector” (Smith, 1990a; Cooley et al., 2012). Additionally, the extremal coefficient has a theoretical connection to Pickand’s dependence function, the exponent

[6] Not to be confused with the parameters θ of a distribution.

measure, and the F-madogram.

For spatial extremes applications, the extremal coefficient is usually restricted to the bivariate isotropic case, for which θ becomes a function only of the distance h between two points: $\theta = \theta(h)$. For the bivariate case, θ has a theoretical range of $1 \leq \theta \leq 2$, where 1 denotes complete dependence and 2 complete independence. The connection with Pickands' dependence function is given by $\theta = 2A(1/2)$. Furthermore, the F-madogram defined in eq. (3.25) can be transformed to $\theta(h)$ with the following:

$$\theta(h) = \frac{1 + 2\nu_F(h)}{1 - 2\nu(h)}. \tag{3.26}$$

The above equation suggests a straightforward method to estimate the value of $\theta(h)$ for an observed dataset. By estimating the empirical F-madogram, the empirical $\theta(h)$ can be derived using eq. (3.26). A comparison between the F-madogram and $\theta(h)$ is given in Fig. 3.18.

While the F-madogram is a valid method to estimate the extremal coefficient $\theta(h)$, it is not the only existing method for estimating $\theta(h)$. In a study comparing the different estimation methods for $A(w)$ (and by extension, θ), Marcon et al. (2017) found that the estimator proposed by Vettori et al. (2018) had better performance than the other methods. This method is based on the F-madogram but uses Bernstein-Bézier polynomials to ensure the resulting estimates fit within the constraints required by Pickands' dependence function. In fact, the issue of not correcting the empirical F-madogram is seen in the left plot of Fig. 3.18: Here, the empirical F-madogram contains points that violate the theoretical limit; this was corrected for $\theta(h)$ using the method proposed by Vettori et al. (2018) in the right plot. The estimation of the empirical extremal coefficient done in the subsequent studies of this thesis was done using the same method.

To summarize, the analog summary measure of similarity between two locations for spatial extremes is not given by the variogram $\gamma(h)$ but rather by the extremal coefficient $\theta(h)$. This coefficient can be estimated by the F-madogram, which behaves similarly to the variogram. We now explore a way to incorporate this measure into a model with max-stability for spatial datasets.

3.6.2 Max-Stable Processes

One of the main goals of spatial extremes methods is to account for the spatial "residual" dependence in the observed data. For non-extreme data, this was accomplished by proposing a zero-mean Gaussian process as the stochastic component of eq. (3.15) and then finding an appropriate covariance function that describes the dependence. For spatial extremes, we want to find a stochastic process that fulfills the max-stability property (section 3.3.1), for which the dependence between locations can then be characterized with an analog of the covariance function. This brings us to the concept of **Max-stable processes**.

Max-stable processes are extensions to infinite dimensions of finite-dimensional EVT theory models like the GPD or the GEV distribution. They arise as "the pointwise maxima taken over an infinite number of (appropriately rescaled) stochastic processes" (Ribatet, 2013). Let $X_1, X_2, ...$ be independent copies of a stochastic process $\{X(s) : s \in \mathcal{S}\}$ with continuous sample paths, where $\mathcal{S} \subset \mathbb{R}^d$ and $d > 1$. Assuming then that there exists continuous functions $a_n(x) > 0$ and $b_n(x) \in \mathbb{R}$ and that the limit is not-degenerate, a max-stable process $\{Z(s) : s \in \mathcal{S}\}$ is defined as:

$$Z(s) = \lim_{n \to +\infty} \frac{\max_{i=1}^{n} X_i(s) - b_n(s)}{a_n(s)}, \quad s \in \mathcal{S} \tag{3.27}$$

In the same manner as with the GEV distribution(eq. (3.6)), the sequence of processes $Z_1, Z_2, ...$ must possess the max-stability property, which means that taking the maxima of $Z(s)$ results in the same process $Z(s)$ as before. The max-stability property entails that any finite D-variate sample $[Z(s_1), ..., Z(s_D)]$ has a multivariate extreme-value distribution, meaning that the marginal distributions of this sample must be GEV distributed. While max-stable processes operate in infinite-dimensions, in practice the data will always have a finite number of dimensions. For this finite-dimensional case, the max-stable process $Z(s)$ can be seen as describing the limiting process of maxima from X_i i.i.d. random fields.

As mentioned above, the margins of a max-stable process are always GEV distributed. To simplify the use of max-stable processes the margins are usually transformed to the unit-Fréchet distribution $\mathrm{GEV}(1, 1, 1)$. A max-stable process whose margins are unit-Fréchet distributed is denoted as a **simple max-stable process**. Note that the transformation to unit-Fréchet given by eq. (3.23) requires that the values of the marginal GEV parameters be known beforehand. Within the context of modeling using max-stable processes, this means that an additional model for the variability of the parameters in space is needed to transform the data. This is typically achieved by using response surfaces like the one in eq. (3.17).

Admittedly, the usefulness of using max-stable processes for modeling spatial extremes is not readily apparent from their definition above. For this, assume that there exists some stochastic process $\{\boldsymbol{X}(s) : s \in \mathcal{S}\}$, where X represents the quantity observed (e.g., rainfall depth), $\mathcal{S} \subset \mathbb{R}^2$ represents the geographical domain, and s the spatial location. In this case, the problem stated in section 3.4 reduces to finding an appropriate model for the joint distribution of $\sup\{\boldsymbol{X}(s) : s \in \mathcal{S}\}$. The problem is that, unlike the univariate case where the maxima converge to a certain family of distributions, no single family exists for the multivariate case. However, the max-stable process $Z(x)$ arises as a justified model of $\sup\{\boldsymbol{X}(s) : s \in \mathcal{S}\}$ if we assume that it is a good candidate to model the partial maxima process. It must be noted that max-stable processes assume that asymptotic dependence exists, so they are not valid models for the asymptotically independent case. If we can construct a max-stable process that properly describes the partial maxima, we can use it to obtain the joint distribution.

Although well-defined, the definition given in eq. (3.27) for max-stable processes is not very useful for model construction, as it does not suggest ways to construct valid processes $\{Z(s)\}$. This problem was approached by de Haan (1984) and Schlather (2002), who proposed several canonical representations of max-stable processes. Of them, the **spectral representation** of de Haan (1984) is the most useful for this thesis. Consider the simple max-stable process $\{Z(s) : s \in \mathcal{S}\}$; For the spectral representation, such a process can be rewritten as:

$$Z(s) = \max_{i \geq 1} \zeta_i Y_i(s), \; s \in \mathcal{S}, \tag{3.28}$$

where $\zeta_i \in \Pi$ denotes the points $\{\zeta_i \geq 1\}$ of a Poisson process Π with intensity $\mathrm{d}\zeta/\zeta^2$ on $(0, \infty)$, and $\{Y(s) : s \in \mathcal{S}\}$ is a non-negative stochastic process with independent realizations $\{Y_i(s)\}_{i \in \mathbb{N}}$ and where $\mathrm{E}[Y(s)] = 1$. An important advantage of this representation is that it allows the straightforward construction of the joint distribution

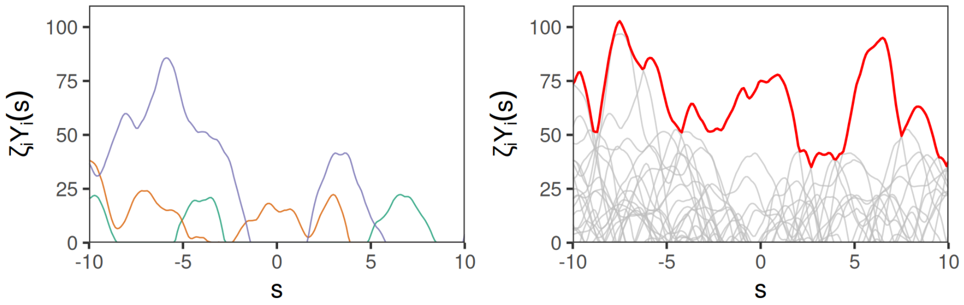

FIGURE 3.19: Visualization of the construction of a max-stable process using the spectral representation from eq. (3.28). The left plot shows three replications of $\zeta_i Y_i(s)$, where $Y_i(s)$ is a zero-mean Gaussian Process with Whittle-Matern covariance. The right plot shows thirty replications: a max-stable process is then the pointwise maximum over all s, denoted by the red line.

function using

$$\Pr[Z(s) \leq z(s), s \in \mathcal{S}] = \exp\left(-\mathrm{E}\left[\sup_{s\in\mathcal{S}}\left\{\frac{Y(s)}{z(s)}\right\}\right]\right). \tag{3.29}$$

Equations (3.28) and (3.29) show that different choices of the stochastic process $\{Y(s)\}$ will lead to different classes of max-stable processes (Cooley et al., 2012). Thus, a straightforward way to construct max-stable processes is to propose different types of stochastic processes for $Y(s)$. The properties of these proposals would then depend on domain-specific aspects.

To better understand the spectral representation of max-stable processes given in eq. (3.28), Smith (1990b) proposed the so-called *rainfall-storms* interpretation. This interpretation is a useful (although somewhat unrealistic) analogy with meteorology. The idea is that every point of ζ_i represents the overall intensity of a rainfall storm that impacts the region $\mathcal{S}$ with a spatial extent given by $\{Y_i(s)\}$. The total amount of rainfall for a storm i centered in s is then given by the product $\zeta_i Y_i(s)$. A max-stable process is given by the pointwise maxima taken over each point in $\mathcal{S}$ for an infinite number of storms. For example, the left plot of Fig. 3.19 shows an example of three "storms" impacting a certain 1-dimensional region. Note that while every storm is different, the overall spatial dependence of each storm is the same, given by the covariance function of the Gaussian Process $Y_i(s)$. The right plot shows many more "storms", for which the max-stable process (in red) is constructed by taking the pointwise maxima for every s. This procedure can be easily extended to two dimensions, giving a direct application of max-stable processes for spatial extremes.

An obvious first choice for the stochastic process $\{Y(s)\}$ in eq. (3.28) is a Gaussian process. Just as in classical geostatistics, the well-known properties of GPs make them easy to use and adapt to many problems. Furthermore, GPs are characterized by the covariance function, which encodes the information of spatial dependence. Thus, choosing a Gaussian process to construct a max-stable process is a straightforward way to include spatial dependence. In fact, the use of GPs as the basis of many parametric max-stable processes is why these processes are sometimes referred to as an extension to geostatistics, but for extremes.

Many different parametric families of max-stable processes exist, most of which use some type of Gaussian process for their construction using the spectral representation. The resulting max-stable processes are typically named after the author that proposed them. These include the Smith process (Smith, 1990b; Schlather, 2002), the Schlather process (Schlather, 2002), the Brown-Resnick process (Brown et al., 1977; Kabluchko et al., 2009), and the extremal-t process (Davison et al., 2012b; Opitz, 2013). Of the four, the Brown-Resnick max-stable process has been shown in many studies to be a good choice for modelling environmental extremes (Engelke et al., 2015; Thibaud et al., 2016; Asadi et al., 2015), and in particular, rainfall extremes (Davison et al., 2012b; Buhl et al., 2016; Davison et al., 2013; Cooley et al., 2012). We now make a brief description of the properties of the Brown-Resnick process, used in most of the studies of this thesis.

The Brown-Resnick max-stable process

First proposed by Brown et al. (1977), the **Brown-Resnick process** is based on a zero-mean Gaussian process with an assumption of intrinsic stationarity. While this process was one of the earliest proposals, it was not until the seminal work of Kabluchko et al. (2009) that its characterization was easy enough to handle for most applications. Following Kabluchko et al. (2009), the Brown-Resnick process is defined by

$$Z(x) = \max_{i\geq 1} \zeta_i \exp(Y_i(s) - \gamma(s)),\ s \in \mathcal{S}, \tag{3.30}$$

where $\{Y_i(s) : s \in \mathcal{S}\}$ are independent copies of a zero-mean Gaussian process with intrinsic stationarity, and γ is the variogram defined in eq. (3.19). The advantage of this process over the other GP-based processes is that the assumption of intrinsic stationarity allows the direct use of the variogram, extending much of classical geostatistics theory to extremes.

A widely used assumption for the variogram $\gamma(h)$ of the BR process is the following expression:

$$\gamma(h) = \left(\frac{h}{\rho}\right)^{\alpha}, \tag{3.31}$$

where h is the euclidean distance between two locations, $\rho > 0$ is the range parameter, and $0 \leq \alpha \leq 2$ is known as the smooth parameter. This variogram model assumes isotropy. However, alternative variogram models can be proposed to include anisotropy. From the variogram expression, it is clear that the dependence structure of the Brown-Resnick process is completely characterized by the parameters ρ and α. Therefore, these parameters are sometimes referred to as the dependence parameters.

An important aspect of the variogram $\gamma(h)$ is that h does not necessarily need to be the euclidean distance $h =\mid \sqrt{s^2 - s'^2} \mid$; h merely needs to be a monotonically non-decreasing function that describes a measure of distance between two points s and s'. For example, in the first study of this thesis, h is defined as a *temporal* distance between different durations.

The use of the variogram for Brown-Resnick processes gives way to a clear connection with the extremal coefficient $\theta(h)$. A theoretical estimate θ_{BR} is given by

$$\theta(h) = 2\Phi\{[\gamma(h)/2]^1/2\}, \tag{3.32}$$

where Φ is the standard Gaussian distribution function. Thus, after fitting the BR parameters, a simple model diagnostic is to compare the empirical estimates of $\theta(h)$ with the theoretical estimates given by eq. (3.32). The proposed model should not deviate too much from the empirical values.

The construction of a joint distribution function for the Brown-Resnick process is straightforward using eq. (3.29) for any number of dimensions D. The density function and the likelihood can then be obtained by differentiating the resulting distribution.

However, in practice, we restrict ourselves to the use of $D = 2$, which avoids the combinatorial explosion when having to differentiate the resulting distribution functions for high-dimensional orders. For example, using $D = 17$ results in 8.3×10^{10} combinations, a number that is intractable in most settings. Therefore, in this thesis, I restrict myself to the bivariate (i.e., two-dimensional) setting. For the bivariate setting, the distribution function of the Brown-Resnick process is given by:

$$\Pr[Z(s) \leq z_1, Z(s') \leq z_2] = \exp\left[-\frac{1}{z_1}\Phi\left(\frac{\sqrt{\gamma(h)}}{2} + \frac{1}{\sqrt{\gamma(h)}}\log\frac{z_2}{z_1}\right) - \frac{1}{z_2}\Phi\left(\frac{\sqrt{\gamma(h)}}{2} + \frac{1}{\sqrt{\gamma(h)}}\log\frac{z_1}{z_2}\right)\right]. \tag{3.33}$$

Here z follows a unit Fréchet distribution, Φ denotes the standard normal distribution function, h is defined as the distance between s and s', and γ is the variogram.

The density function is given by differentiating the bivariate function; an expression can be found in Tyralis et al. (2019).

3.6.3 Inference for max-stable processes

In principle, if N-stations are available in the dataset, inference for the model's parameters would be based on the corresponding N-dimensional distribution constructed from the max-stable process. However, as mentioned in the last section, the way to construct a distribution function for max-stable processes given by eq. (3.29) has many severe practical issues for modeling with these processes. MLE and Bayesian methods are based on the likelihood, which requires an expression for the density function. To get the density function from a max-stable process, differentiating the distribution function is necessary. This derivative is performed for all N dimensions and their combinations, which inevitably leads to a combinatorial explosion of terms when dealing even with moderately low-dimensional levels.

The lack of closed-form expressions for the joint likelihood of high-dimension cases motivates the use of the so-called **composite likelihood**. The composite likelihood is a type of misspecified likelihood that can be constructed from low-dimensional cases to be used for inference of the max-stable process. In recent studies, for example, the most common approach is to construct a composite likelihood from the 2-dimensional case. It should be noted that the development of asymptotic theories of MLE using composite likelihoods has led to most applications being based on the frequentist paradigm; however, some work with Bayesian inference has also been done in recent times. For a review of composite likelihood and their methods, check Sang (2016).

For working with max-stable processes in the bivariate case, the composite likelihood is given by the **pairwise likelihood**. For a dataset with N stations, the idea of the pairwise likelihood is to find all possible $N(N-1)/2$ pairs and use them to construct a likelihood function based on the bivariate distribution. In mathematical terms, the pairwise log-likelihood is given by:

$$\ell_P(\boldsymbol{\theta}) = \sum_{k=1}^{K}\sum_{i=1}^{N-1}\sum_{i'=i+1}^{N} \log f(x_k(s_i), x_k(s_{i'}) \mid \boldsymbol{\theta}), \tag{3.34}$$

where $f(\cdot)$ denotes the bivariate density of the max-stable process as given by the derivative of the distribution function, and (s, s') represents two different locations. The first sum is over the K i.i.d replicates (i.e., the different years/months), and the other two sums are over all the possible pairs of N-stations. Note that just like the normal likelihood, the log-version is preferred to work with, as this avoids the product of tiny quantities.

A potential problem surges when the different measuring stations have differing record lengths, leading to an imbalance of some locations having much more K-maxima than others [7]. The expression (3.34) for the log-pairwise likelihood does not directly indicate how to deal with this situation, but we propose as a solution to set the value of zero to the log-likelihood of any invalid pairs. Simply ignoring invalid pairs by assigning them a log-likelihood of zero avoids having to throw a considerable amount of data, but at the expense of potentially invalidating the theoretical framework of the composite likelihood. Therefore, in the studies for this thesis, this workaround was performed exclusively when the amount of missing data was judged to be inconsequential compared to the total number of stations.

It is tempting to treat the pairwise likelihood as if it was the full likelihood for inference: In fact, inference for bivariate max-stable processes is done by simply substituting the full likelihood with ℓ_P for MLE estimation. This substitution is not without merit, as MLE using composite likelihoods has been seen to yield estimators with good asymptotic properties (which will not be mentioned here). Nevertheless, it is important to remember that any composite likelihood constitutes a misspecified model, and as such, estimators obtained from composite likelihoods will typically result in a loss of statistical efficiency compared to its full likelihood counterpart. This does not mean that estimators based on the composite likelihood are useless, as several corrections can be used to counteract this effect. For example, the Open-Faced Sandwich correction proposed by Shaby (2014) extends the use of composite likelihoods for the Bayesian paradigm. Details about the OFS correction and its novel application to the second study described in this thesis are explained in section 5.2.3.

In contrast to the methods to approximate the prediction accuracy of the models explored in section 2.5, model selection is usually performed with the **Composite Likelihood Information Criterion** (CLIC). The CLIC is based on the expected KL-divergence between the true unknown model and the adopted misspecified model given by the composite likelihood. As with the AIC, the CLIC can be used to compare different models.

It is important to mention that the pairwise likelihood stemming from the bivariate density function constructed from eq. (3.29) is valid only for unit-Fréchet distributed margins. However, by transforming the margins to unit-Fréchet, we effectively ignore the spatial variation in the GEV parameters. This effect is undesirable, as finding the spatial variation of the GEV parameters commonly constitutes one of the goals of spatial analysis. To tackle this problem, two options exist: (i) finding the spatial variation of the GEV parameters beforehand by assuming the observations are iid, effectively ignoring the spatial dependence; or (ii) constructing a method of inference that finds the value of both the GEV parameters and the BR dependence parameters all-at-once. The latter method transfers some information from the spatial dependence to the marginal estimates. However, just how much information is transferred is not clear. The first two studies of this thesis delve more into this issue by fitting models with the all-at-once approach and comparing them to the iid option. As will be seen, the difference in the pointwise estimates is not very large, but the difference in the uncertainty is significant.

Introducing the required transformation from unit-Fréchet to an arbitrary GEV distribution of the margins for the pairwise likelihood requires a way to transform a random variable from one distribution to another. Let X be a series of random vectors

[7] This situation is exceedingly common when dealing with weather observations.

with probability density $p_x(x)$, and $Z = T(X)$ be a bijection mapping values of X to Z, where Z has probability density p_z. Both densities are related by

$$p_z(T(x)) = \frac{p_x(x)}{|J_T(x)|}, \tag{3.35}$$

where $|J_T(x)|$ is the determinant of the **Jacobian matrix** of T. The Jacobian matrix is given by the derivatives of each n-component of T:

$$J_T(x) = \begin{bmatrix} \frac{\partial T_1}{\partial x_1} & \cdots & \frac{\partial T_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial T_n}{\partial x_1} & \cdots & \frac{\partial T_n}{\partial x_n} \end{bmatrix}.$$

Therefore, the Jacobian matrix is necessary to transform a random vector from one distribution to another[8].

For the particular case of the Brown-Resnick process, the bijection $z = T(x)$ is given by eq. (3.23). This transforms the x GEV-distributed observations to unit-Fréchet ones (here represented by z). The bivariate density derived from eq. (3.33) is valid only for unit-Fréchet data. Thus, in order to use the GEV-distributed data in the bivariate density directly, we rearrange eq. (3.35) to be

$$f(x(s), x(s')) = f(T(x(s), x(s')))|J_T(x(s), x(s'))|,$$

where $f(\cdot,\cdot)$ is the bivariate density of the Brown-Resnick process. This expression lets us directly introduce the observed block-maxima in the pairwise log-likelihood ℓ_P, which in turn means that inference can be performed in the all-at-once setting.

To summarize this section, inference for max-stable processes is typically possible only for composite likelihoods like the pairwise likelihood, as the full likelihood is commonly intractable. The pairwise likelihood possesses favorable properties, so methods such as MLE or Bayesian inference can be applied by substituting the full likelihood with the composite likelihood. However, this usually comes at a cost associated with using a misspecified likelihood, so care must be used when interpreting the result. This is why several corrections exist for estimates obtained using composite likelihoods. Lastly, the density given by simple max-stable processes allows only unit-Fréchet margins to be used in the likelihood, but this can be extended to the full GEV distribution using the Jacobian transformation.

3.6.4 Simulation from max-stable processes

As seen in section 2.5, the goal of statistical modeling is sometimes to generate *simulated* values by randomly drawing from the resulting distribution. The idea is that the simulated values will also be distributed according to the model so that we can use them as future "predictions" of sorts. For the univariate case, simulating from a GEV distribution is relatively straightforward, for which many simulation techniques have been developed. However, the situation changes when simulating a draw from a max-stable multivariate distribution like the ones given by max-stable processes. Remember that closed-form analytical expressions for distributions resulting from max-stable processes are usually available only for bivariate distributions, and yet, we would like to simulate from a process that includes all dimensions.

[8] If p_x and p_z are probability functions for discrete random variables then no Jacobian term is needed.

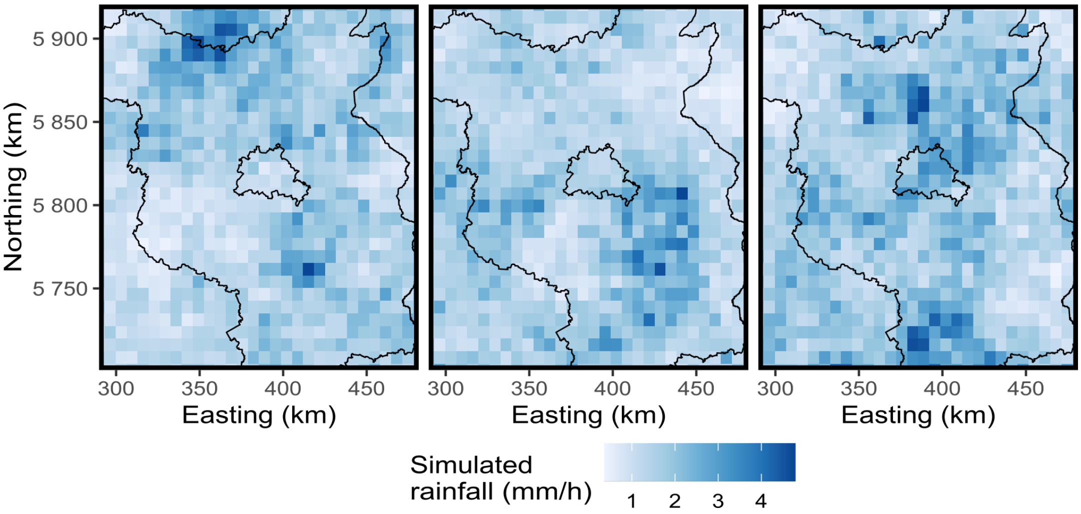

FIGURE 3.20: Three realizations from a Brown-Resnick max-stable process fitted using the data and methods described in section 5.2.1. Notice how the spatial dependence is kept for each realization.

Exact simulation from max-stable processes is usually not possible; however, some recently developed algorithms make it possible to sample a finite number of locations from a max-stable process (Oesting et al., 2016). For example, Dombry et al. (2016) proposed an algorithm for sampling a Brown-Resnick process via the spectral measure. This algorithm is currently implemented in the R-package `SpatialExtremes` (Ribatet, 2020).

Figure 3.20 shows three simulations from a Brown-Resnick max-stable process fitted using rainfall data. The spatial dependence of the BR process can be seen in the structures of each realization in the form of regions with increased intensity. Each individual realization will result in a different field. Note that this maps show a field of pointwise annual block maxima, so that the size of the events cannot be directly characterized, as the maxima for each point does not necessarily come from the same event. However, it does give an indication of possible sizes if one considers that many extreme rainfall events have a large spatial extent.

3.6.5 Alternative approaches for spatial extremes

So far in this section, we have only considered the use of max-stable processes to model the joint distribution of the $\boldsymbol{X}(s)$ block maxima. Max-stable processes can be seen as an extension of classical geostatistics and as such, share many of the advantages of geostatistical methods. However, they are not the only valid approach for modeling spatial extremes. For an exhaustive review of the other existing methods, see (Davison et al., 2012b; Cooley et al., 2012; Davison et al., 2015; Huser et al., 2022).

The main alternative to max-stable processes is using so-called **latent variable** models. A latent variable model is a multilevel model where the block maxima from different locations are all assumed to be independent and follow the same distribution family. In this setting, the margins follow the GEV distribution, whose parameters vary in space. In turn, the variability in space of the GEV parameters is modeled via an additional layer constructed using a Gaussian Process. Therefore, an individual GEV is given for each location, but at the same time information from all stations is used to inform the estimates of each location's parameters. An advantage of this method is that it requires much less computational demand than a max-stable process and is theoretically much simpler. In fact, this model type is optimal when the main goal of the analysis is to infer the marginal parameters. The problem with the latent variable approach comes from the independence assumption, which intrinsically forces the resulting likelihood to be misspecified, resulting in an underestimation of the uncertainty. Furthermore, simulations using the latent variable approach will be nonsensical, as the spatial structure will not be reflected in the simulated draws.

A further alternative for modeling spatial extremes with block maxima is the use of copulas. **Copulas** are functions that describe the dependence between several random vectors, independent of their marginal distribution. However, it is well-known that copulas have a poor performance when modeling extremes, as they usually cannot capture the tail dependence adequately. The exception is the use of extreme-value copulas; however, these kinds of copulas constitute what are effectively max-stable processes.

Finally, some alternatives to the block maxima approach exist for spatial extremes. An extension of PoT methods is given by the **r-Pareto** process, described in Fondeville et al. (2018) and Fondeville et al. (2020). This kind of model is optimal for modeling the size of events, which cannot be discerned when using block maxima.

3.7 Summary

The spatial modeling of extreme rainfall requires the extension of classical geostatistics to incorporate max-stable distributions and account for the extremal dependence structure. This is given by the methods of spatial extremes, which extend the theoretical basis of univariate EVT to a multivariate setting. The extension of EVT to multivariate settings results in max-stable processes. In the univariate case, block maxima are described exclusively by the GEV distribution. However, for the multivariate case, no single parametric family exists, so no single form for a max-stable process exists. Nevertheless, the spectral form of max-stable processes provides a straightforward way to construct them using different types of Gaussian processes. Of the different possibilities, the Brown-Resnick process has been proven to be successful at modeling rainfall maxima. The Brown-Resnick process encodes the spatial dependence into the variogram $\gamma(h)$, which can be transformed into the summary measure of extremal dependence known as the extremal coefficient. Thus, Brown-Resnick max-stable processes are proposed as a theoretically-sound method for modeling extreme rainfall in the following sections.

Part II

Developments in the statistical modeling of extreme rainfall

4 Evaluating the Performance of a Max-Stable Process for Estimating Intensity-Duration-Frequency Curves

Publication details

Jurado, Oscar E., Ulrich, Jana, Scheibel, Marc, & Rust, Henning W. (2020). **Evaluating the Performance of a Max-Stable Process for Estimating Intensity-Duration-Frequency Curves.** Water, 12(12), 3314. https://doi.org/10.3390/w12123314

Author contributions

Conceptualization, O.E.J. and H.W.R.; Data curation, M.S.; Formal analysis, O.E.J. and H.W.R.; Funding acquisition, H.W.R.; Investigation, O.E.J.; Methodology, O.E.J. and J.U.; Project administration, H.W.R.; Resources, H.W.R.; Software, O.E.J. and J.U.; Supervision, H.W.R.; Validation, O.E.J.; Visualization, O.E.J.; Writing—original draft, O.E.J.; Writing—review & editing, J.U., M.S. and H.W.R.

Abstract

To explicitly account for asymptotic dependence between rainfall intensity maxima of different accumulation duration, a recent development for estimating Intensity-Duration-Frequency (IDF) curves involves the use of a max-stable process. In our study, we aimed to estimate the impact on the performance of the return levels resulting from an IDF model that accounts for such asymptotical dependence. To investigate this impact, we compared the performance of the return level estimates of two IDF models using the quantile skill index (QSI). One IDF model is based on a max-stable process assuming asymptotic dependence; the other is a simplified (or reduced) duration-dependent GEV model assuming asymptotic independence. The resulting QSI shows that the overall performance of the two models is very similar, with the max-stable model slightly outperforming the other model for short durations ($d \leq 10\,\mathrm{h}$). From a simulation study, we conclude that max-stable processes are worth considering for IDF curve estimation when focusing on short durations if the model's asymptotic dependence can be assumed to be properly captured.

4.1 Introduction

Much research has been recently done on the application of multivariate methods to estimate Intensity-Duration-Frequency (IDF) curves. IDF curves are a popular tool among hydrologists to estimate exceedance probabilities of extreme rainfall events with different durations. In broad terms, IDF curves model a relationship between intensities of extreme rainfall events and their frequencies (i.e., return periods) as a function of event duration. A challenge in estimating IDF curves is how to deal with the simultaneous modeling of intensities for different durations, in particular, how to account for the possible dependence that could arise between intensities of different durations.

Initially, the conventional approach to model IDF curves was based on univariate extreme value theory (EVT) models. Early work on the topic estimates extreme value distributions individually for several fixed durations and subsequently fits an empirical relation to quantiles (return levels) as a function of duration (Chow, 1953; Aparicio, 1997; García-Bartual et al., 2001; Monjo, 2016). This approach is prone to inconsistencies as the natural ordering of quantiles is not guaranteed to be preserved over all durations (in other words: quantiles cross). To address this problem, Koutsoyiannis et al. (1998) suggested a consistent extreme value model for intensities as a function of duration with location and scale being functions of duration. Later on, a couple of studies implemented methods from Bayesian statistics for the univariate relationship between intensity and duration. Lehmann et al. (2016) formulated a Bayesian Hierarchical Model (BHM) based on the model from Koutsoyiannis et al. (1998), and Van de Vyver (2018) presented a multi-scale model using Bayesian inference. More recently, Ritschel et al. (2017) used this model to characterize stochastic precipitation models, and Ulrich et al. (2020) proposed the addition of spatial covariates to the model from Koutsoyiannis et al. (1998), extending the work of Fischer et al. (2017) using spatial covariates to model daily precipitation maxima. These studies make an assumption of stationarity, which may not be valid under a changing climate. Some recent studies have focused on tackling this issue with univariate methods to construct consistent IDF curves in a nonstationary setting. Some examples include the work of Padulano et al. (2019) using the storm index method, as well as those of Ganguli et al. (2017) and Ganguli et al. (2019) comparing the estimates from a stationary and nonstationary method.

Many of the previous univariate models assume that rainfall intensities are independent for different durations, thus simplifying the modeling efforts. However, the estimation of an extreme value distribution as a function of durations brings along the problem of dependence of extremes associated with different durations, as longer duration series are always aggregated from series of shorter durations. An important consequence of this way of aggregating is that there exists no “single” event for any given duration. As an example, the values of the 15 minute duration events are not events that lasted exactly 15 minutes, but rather the “largest” 15 minute long average values from longer events.

In recent years there has been widespread use of multivariate EVT methods for modeling IDF curves, which allow the explicit modeling of dependence structures that could not be captured by the univariate approach. Essentially, a multivariate extreme value distribution (MEVD) is fitted to extreme precipitation data, with the marginal distributions being frequently modeled by a univariate extreme value distribution. Simultaneously, the dependence structure is described with methods such as max-stable processes or copulas. An early example of this was proposed by Muller et al. (2008) who, alongside the independence likelihood (analogous to the univariate approach), proposed the use of a so-called trivariate likelihood for three durations: 1, 24, and 72 h. In this study, the dependence between 24 and 72 h was modeled with a bivariate extreme distribution from the logistic family. Afterward, Van de Vyver (2015) investigated the use of the trivariate likelihood by calculating the parameters' uncertainty using a Bayesian approach. He found that, while the resulting posterior distributions from the trivariate likelihood were narrower than the independence likelihood ones, its limitations were too strict to recommend its use over the independence likelihood. An early example of modeling IDF curves using copulas is the study of Singh et al. (2007), who used a Frank Archimedean copula to estimate the IDF relationship in a bivariate setting.

Some of the most recent advances have been a result of studies that attempted to model IDF curves within a spatial setting. These approaches take advantage of the methods developed for modeling so-called spatial extremes (Davison et al., 2013). For example, Stephenson et al. (2016) implemented a spatial max-stable process for IDF estimation; they estimated IDF curves in a spatial setting by incorporating a Bayesian Hierarchical Model in every station with a max-stable process. While their approach was able to capture the spatial dependence, they had to limit their scope to assume that the rainfall maxima were independent for different durations. Later on, Tyralis et al. (2019) proposed the use of a max-stable process to estimate IDF curves for a single station in a way that the asymptotic dependence between rainfall intensity maxima of different durations was explicitly modeled, extending the spatial methodology by proposing a so-called duration space instead of geographical coordinates. Remarkably, their proposed max-stable process was able to explicitly account for the asymptotic dependence between intensities for different durations.

Although Tyralis et al. (2019) demonstrated the feasibility of a duration-dependent max-stable process to estimate IDF curves, they did not investigate their performance compared to, e.g., univariate EVT methods. Their results showed that both the univariate and multivariate approaches adequately approximated the empirical quantiles, where the max-stable approach resulted in more conservative (i.e., higher) intensities for large quantiles (i.e., longer return periods). In our study, we aimed to build upon the results of Tyralis et al. (2019) by estimating the impact on performance when accounting for the asymptotical dependence between rainfall intensity aggregated over different durations. We expected that, whenever the asymptotical dependence between durations is high and the max-stable approach is able to capture the "strength" of such dependence for the estimated intensities, the model performance should be significantly higher than for a model assuming independence.

The present paper introduces a scheme to evaluate the impact of the asymptotic dependence between rainfall intensities aggregated over different durations on the performance of EVT-based IDF models. This involves a comparison of skill between two IDF models: one accounting for asymptotic dependence between durations, the other assuming independence. The comparison shows that accounting for the dependence between rainfall intensity aggregated over different durations slightly improves the point estimates for long return periods, and particularly for "short" durations ($d \leq 10\,\text{h}$) usually associated with convective phenomena. However, this comes at the price of increased complexity of modeling the asymptotic dependence.

4.2 Methods and Data

Our study involves numerical experiments to estimate the relative performance of IDF curves modeled with two approaches: one based on a max-stable process to account for asymptotic dependence (henceforth named as the MS-GEV approach), and another one based on the assumption of independence using the reduced d-GEV model (henceforth named as the rd-GEV approach). By doing so, we aimed to estimate how the performance of IDF curves is affected by considering (or ignoring) the asymptotic dependence between rainfall intensity maxima for different durations.

We performed the study in two broad steps. First, we conducted a case study using data from 6 rain-gauge stations. This data was used to estimate the respective IDF-model parameters from both the MS-GEV and rd-GEV approaches. We compared the performance of both approaches using a measure of skill. In the second step, we introduced synthetic data with known levels of dependence to estimate the effect that the level of dependence has on the resulting estimations. We used the synthetic data to estimate and compare their performance again. Finally, we compared the results from both steps to determine how the performance was affected by the asymptotic dependence between durations. The two different methods used for estimating IDF curves are explained in more detail in the following section. Subsequently, the methods used for verification are described, and, finally, the observation data and the synthetic data are presented.

4.2.1 Estimation of IDF Curves

Let $\zeta_d(t)$ be the instantaneous rainfall intensity series integrated over a time window of length d, where d is an arbitrary time duration (which commonly ranges from the measurement interval to 72–120 h). Given $\zeta_d(t)$, we obtain the series of the maximum annual average rainfall intensity for each value of d as

$$i(d) = \max_{y^- < t < y^+} \{\zeta_d(t)\} \qquad y = (1, ..., n), \tag{4.1}$$

where n is the total number of observation years, and y^-, y^+ are the beginning and end of the yth year, respectively. As a rule, k durations $d_j, j = 1, ..., k$ are simultaneously used when constructing $i(d)$, with values from the measurement interval to 120 h (depending on the application). The resulting k series can be thought of as a realization of a random variable $I(d)$. Notice that, in Equation (4.1), d is not a random variable but a parameter for the intensity, as noted by Koutsoyiannis et al. (1998).

The construction of $i(d)$ from a single duration series (e.g., hourly precipitation sums) generates a statistical dependence between $i(d_1)$ and $i(d_2)$ corresponding to different aggregation durations $d_1 \neq d_2$. For example, $i(d = 2\,\text{h})$ and $i(d = 3\,\text{h})$ show a very high dependence (that is, when one of them has what is considered to be a high value, the other intensity also has a high value). However, as the gap in aggregation duration between the values grows, this dependence diminishes: $i(d = 2)$ and $i(d = 24)$ show almost complete independence. Nadarajah et al. (1998) proposed a scheme to work with this type of random variable, which they denoted as ordered random variables. Some authors have linked this concept to the different physical processes that result in different time scales for precipitation events. For example, Muller et al. (2008) claimed that the independence between the 1 h and 24-h events were due to the 1-h events being a result of convective (local) motions, while the 24-h event was related to synoptic phenomena. In this paper, we take a closer look at how this dependence affects the estimation of IDF curves down the road.

Following the block-maxima approach, e.g., (Coles, 2001), a GEV can be fitted to each k-series of $i(d_j)$. Then, the return levels $z_{d,T}$ associated with return periods $T = 1/p$ can be calculated for each duration used for the fit d_j, where p is denoting the non-exceedance probability with values usually in the range corresponding to upper extreme quantiles. The idea behind IDF curves is to describe the return level $z_{d,T}$ for arbitrary durations d based on the sample $i(d_j)$ in a meaningful way. This typically involves a parametric form of $z_{d,T}$ as a function of d. The choice of the model used for estimating IDF curves depends on the choice of parametric form of $z_{d,T}$.

For this study, we employed two different approaches for the parametric form of $z_{d,T}$ as a function of d. For the model that assumes asymptotical independence for $i(d)$ for different d, we follow the duration dependent GEV model (d-GEV) of Ritschel et al. (2017). Based on Koutsoyiannis et al. (1998), the d-GEV model of Reference Ritschel et al. (2017) estimates a GEV simultaneously from annual maxima associated with various durations, thus conceiving the GEV as a function of duration. The d-GEV yields consistent quantiles $z_{d,T}$, which cannot cross by definition. The other approach is based on a max-stable process for modeling the relationship of $z_{d,T}$ that accounts for the asymptotic dependence between durations.

Using the Duration-Dependent GEV

Following Ritschel et al. (2017) and Ulrich et al. (2020), we used the duration dependent GEV (d-GEV) to model $i(d)$ with the distribution

$$G(x) = \exp\left[-\left(1+\xi\left(\frac{x}{\sigma(d)}-\tilde{\mu}\right)\right)^{-1/\xi}\right], \tag{4.2}$$

where $\sigma(d) = \sigma_0/(d+\nu)^\eta$ is the duration dependent scale parameter, and $\tilde{\mu} = \mu(d)/\sigma(d)$ is the modified location parameter.

Given the estimated $(\mu(d), \sigma(d), \xi)$ parameters, it is straightforward to calculate the return level $z_{d,T}$ for any arbitrary duration using

$$z_{d,T} = \mu(d) + \frac{\sigma(d)}{\xi}\left[\left(-\log\left(1-\frac{1}{T}\right)\right)^{-\xi} - 1\right]. \tag{4.3}$$

To compare the resulting $z_{d,T}$ of this approach with those estimated using the MS-GEV approach, we set the parameter $\nu = 0$. This parameter is related to sub-hourly duration values ($0 < d < 1$), which we do not consider here. Therefore, the dependence of location and scale parameter on duration follows

$$\mu(d) = \tilde{\mu}\sigma_0 d^{-\eta}, \tag{4.4}$$

$$\sigma(d) = \sigma_0 d^{-\eta}. \tag{4.5}$$

This results in a model with four parameters to be estimated, namely $\{\tilde{\mu}, \sigma_0, \xi, \eta\}$. We call this modified distribution the *reduced* d-GEV or rd-GEV for short. The parameters of the d-GEV distribution are estimated by maximizing the likelihood as implemented in the R-package IDF (Ulrich et al., 2019). Equation (4.3) is used to get intensities for arbitrary durations ($d \geq 1$).

Using a Max-Stable Process

Max-stable processes are extensions to infinite dimensions of finite-dimensional extreme value theory models (i.e., extremes of random variables or vectors). They arise as "the pointwise maxima taken over an infinite number of (appropriately rescaled) stochastic processes" (Ribatet, 2013). Let $\{X(x) : x \in \chi\}$ be a stochastic process, where χ is a compact subset of $\mathbb{R}^d, d \geq 1$, and $\{Z(x) : x \in \chi\}$ be a max-stable stochastic process. Following de Haan (1984), if there exist continuous functions $a_n(x) > 0$ and $b_n(x) \in \mathbb{R}$, and provided that the limit is non-degenerate, the process $Z(x)$ can be defined as

$$Z(x) = \lim_{n \to +\infty} \frac{\max_{i=1}^{n} X_i(x) - b_n(x)}{a_n(x)}, \qquad x \in \chi \,. \tag{4.6}$$

The max-stable process $Z(\cdot)$ describes the limiting process of maxima from the X_i IID random fields (Zheng et al., 2015). The use of this max-stable process for modeling spatial extremes is justified when, based on n independent replicates, and, if n is large enough, we assume that $Z(x)$ is a good candidate for modeling the partial maxima process $\{\max_{i=1,\ldots,n} X_i(x) : x \in \chi\}$ (Dey et al., 2016).

One of the main advantages of using a max-stable process is that it provides a flexible way of modeling the dependence structure between the X_i IID random fields. If we assume that $\chi \subset \mathbb{R}^2$ represents a geographical catchment, we can think that for multiple points ($x \in \chi$), the marginal distributions are jointly modeled via the max-stable process, resulting in continuous functions of the GEV parameters $\mu(x), \sigma(x), \xi(x)$ for each margin.

In order to implement a max-stable model for estimating IDF curves, we followed the framework proposed by Tyralis et al. (2019). This approach (MS-GEV) employs the Brown-Resnick process, a frequently used parametric family of max-stable processes for modeling environmental extremes (Engelke et al., 2015; Thibaud et al., 2016; Asadi et al., 2015). Previous studies have shown the applicability of the Brown-Resnick process for extreme rainfall applications (Davison et al., 2012b; Buhl et al., 2016; Davison et al., 2013; Cooley et al., 2012). A central proposition of our current approach is to define a continuous variable $\bar{i}(d)$ in a one-dimensional space, where each "location" is one of the durations d. This in contrast to other applications of max-stable processes, where the variable of interest is commonly defined in a two-dimensional (e.g., latitude and longitude) space.

For any max-stable process, the $X(s_i)$ marginals are generalized extreme value (GEV) distributed, with distribution function:

$$G(i) = \exp\left\{ - \left[\left(1 + \xi \frac{i - \mu}{\sigma} \right)_+^{-1/\xi} \right] \right\} , \tag{4.7}$$

where μ, σ, ξ are the location, scale and shape parameters, and $x_+ = \max(0, x)$. Following de Haan (1984) (with the adaptation for $d > 0$), when the limiting process $\{\bar{i}(d) : d > 0\}$ is non-degenerate, a simple max-stable process can be constructed by its so-called spectral characterization, which is a representation of the max-stable process in the frequency domain (Ribatet, 2013). In the spectral characterization, the max-stable process is given by choice of the stochastic process ($X_i(d)$ in Equation (4.21)).

Such max-stable processes are called *simple* as the margins $\bar{z}(d)$ are unit Fréchet distributed (i.e., $\mu = \sigma = \xi = 1$). The use of unit Fréchet marginals is standard, as the max-stable process theory is based on the assumption that the marginals have a common, convenient max-stable distribution. There is no loss of generality in assuming that the limiting process $\{\bar{i}(d) : d > 0\}$ has unit Fréchet margins, as it is straightforward to transform such margins into arbitrarily GEV-distributed ones and vice versa.

For this study, we used the bivariate form of the Brown-Resnick process given by Kabluchko et al. (2009) (see Appendix 4.A) as the stochastic process $X_i(d)$. In this form, the dependence is a function of the semivariogram γ, defined as

$$\gamma^2(h) = 2\left(\frac{h}{\rho}\right)^{\alpha}, \qquad \rho > 0, \qquad 0 < \alpha \leq 2\,, \tag{4.8}$$

where α and ρ are, respectively, the smooth and range parameters of the semivariogram, and h represents a measure of the *distance* between two durations. Tyralis et al. (2019) calculated this distance as the euclidean distance

$$h_e = |d_j - d_i|, \tag{4.9}$$

where the indices i and j denote different durations d_i in hours, and $j \neq i$. However, this measure does not account for the non-linearity of the distance between durations for events of increasing magnitude. For example, consider that an event of 4 h compared to one of 2 h ($h_e = 2$) is already twice as long, while an event of 50 h compared to one of 48 h ($h_e = 2$) is only 1.04 times the second one.

To address this issue, we explored the use of a distance measure based on a logarithm following Van de Vyver et al. (2018). This distance is defined as

$$h_l = \log_2(d_j) - \log_2(d_i) = \log_2\left(\frac{d_j}{d_i}\right), \tag{4.10}$$

where i and j denote different durations d_i in hours, and $j > i$.

We compare the resulting pairwise extremal coefficients from both distance measures to discern which one results in a more appropiate fit for the semivariogram of Equation (4.8).

The bivariate form of the Brown-Resnick max-stable process described in Equation (4.22) is valid only for unit Fréchet marginals. Therefore, the series of yearly rainfall intensity $\bar{i}(d)$ requires an appropriate transformation (see Tyralis et al. (2019) for details) to be unit Fréchet distributed, using the relationship

$$\bar{z}(d) = (1 + \xi(\bar{i}(d) - \mu(d)/\sigma(d))_{+}^{1/\xi}\,. \tag{4.11}$$

To link $\bar{i}(d)$ to $\bar{z}(d)$ for all durations, we used the response surfaces for the GEV parameters (Tyralis et al., 2019):

$$\mu(d) = \mu_0 d^c\,, \tag{4.12}$$

$$\sigma(d) = \sigma_0 d^c\,. \tag{4.13}$$

These response surfaces follow the constraints given by Equations (4.18)–(4.20), and describe a function $\Psi(d)$ for the parameters of the marginals of $\bar{i}(d)$. Equations (4.12) and (4.13) are equivalent to Equations (4.4) and (4.5) in the rd-GEV model. The response surfaces allow to use all durations simultaneously when estimating the max-stable process parameters.

Taking the response surfaces into account, the parameters that we need to estimate for calculating IDF curves using the Brown-Resnick model are six: $[\rho, \alpha, \mu_0, \sigma_0, \xi_0, c]$. This is accomplished via the maximum likelihood estimate of the pairwise likelihood given in Equation (4.23). The estimation of the parameters is done using the `R` package `SpatialExtremes` (Ribatet, 2020).

Of particular usefulness for our study, is a measure to summarize the strength of the asymptotical dependence modeled by the Brown-Resnick max-stable process. A well-known measure is the extremal coefficient θ, which for the bivariate case, can take values of $1 \leq \theta \leq 2$. The value of θ decreases when the dependence between the two margins increases. When $\theta = 1$ the two margins are completely dependent, and when $\theta = 2$ they are independent.

For the dependence structure of the Brown-Resnick max-stable process described by Equation (4.8), the extremal coefficient is a function only of the distance between durations: $\theta = \theta(h)$. Given the semivariogram γ for a Brown-Resnick max-stable process, with Φ denoting the standard normal distribution function, the extremal coefficient is given by

$$\theta(h)_{BR} = 2\Phi(\gamma(h)/2)^{1/2}\,. \tag{4.14}$$

For comparison purposes, we also calculate the extremal coefficient nonparametrically ($\hat{\theta}_{emp}$). For our study, we used the method proposed by Marcon et al. (2017), which was found by Vettori et al. (2018) to generally perform better than other nonparametric estimators.

To summarize: To estimate IDF curves with a max-stable process, we (i) transform our block-maxima data $i(d)$ into unit Fréchet using Equation (4.11) with the response surfaces given by Equations (4.12)–(4.13); then (ii) estimate the parameters via the maximum likelihood estimates of the pairwise likelihood given by Equation (4.23); and, finally, (iii) calculate the intensity for any arbitrary duration d and return period T from Equation (4.3). We perform all the computations within the `R` language (R Core Team, 2020). The data and code is available as supplementary information.

4.2.2 Verification and Model Comparison

As a performance measure, we use the quantile score (QS) (Bentzien et al., 2014). This allows us to evaluate predictions of $z_{d,T}$ estimated from IDF-curves in terms of quantiles (i.e., return periods). For D durations, N years, and a given return period T the QS is defined as

$$\mathrm{QS}_T = \frac{1}{ND}\sum_{d=1}^{D}\sum_{n=1}^{N}\rho_T\left(i_n(d) - z_{d,T}\right), \tag{4.15}$$

where $i_n(d)$ is the observed block maxima, $z_{d,T}$ is the corresponding intensity from the model (using Equation (4.3)), and $\rho_T(u)$ is the so-called check function

$$\rho_T(u) = \begin{cases}(1-1/T)\,u & u \geq 0\\ (-1/T)\,u & u < 0;\end{cases} \tag{4.16}$$

thus, for this particular application, $\rho_T(u) = \rho_T(i_n(d) - z_{d,T})$.

The QS is always positive and reaches an optimal value at zero. To compare the performance of a model with a reference, we use the Quantile Skill Index (QSI) (Ulrich et al., 2020) derived from the Quantile Skill Score $QSS = 1 - QS_{\text{model}}/QS_{\text{ref}}$ (see Wilks (2011) for more details on skill scores), defined as

$$\mathrm{QSI} = \begin{cases}1 - \frac{\mathrm{QS}_{\text{model}}}{\mathrm{QS}_{\text{ref}}}, & \text{if } \mathrm{QS}_{\text{model}} < \mathrm{QS}_{\text{ref}}\\ -\left(1 - \frac{\mathrm{QS}_{\text{ref}}}{\mathrm{QS}_{\text{model}}}\right), & \text{if } \mathrm{QS}_{\text{model}} \geq \mathrm{QS}_{\text{ref}}\end{cases}. \tag{4.17}$$

For our study, $\mathrm{QS}_{\text{model}}$ is the score for the MS-GEV approach, and $\mathrm{QS}_{\text{reference}}$ is the score for the rd-GEV approach. Positive (negative) values of the QSI indicate a gain (loss) of skill for the MS-GEV approach compared to the rd-GEV one. The advantage of using the QSI over QSS is that negative values have a more meaningful interpretation.

To get a robust estimation of the prediction error for the QSI, we applied 10-fold cross-validation to estimate the QS (Hastie et al., 2009). The QSI is then obtained using the mean cross-validated QS, averaged over all cross-validation folds, for each model. We used return periods $T = (5, 10, 20, 40, 100)$ years. However, the results from the 100 year return period should be interpreted with caution, as the data used for parameter estimation consists of much shorter series of approximately 40 years. The QSI has to be interpreted with care for return periods much larger than the length of the time series (e.g., $T > 40$ years). The lack of observations for this region could result in a really high uncertainty of the value of the QS, and, therefore, of the QSI.

4.2.3 Data

Two different datasets are used for our study. The first one is a synthetic dataset generated for the simulation study. The second one is the block maxima from six rain gauge stations located in the Wupper Catchment (West Germany). We describe both datasets in the following section.

Synthetic Data

We generate synthetic datasets with varying levels of dependence to investigate the models' performance in estimating IDF curves. We designed three synthetic datasets that simulate rainfall block maxima aggregated over different durations with increasing levels of dependence. For each dataset, we simulate values from a Brown-Resnick simple max-stable process with known dependence parameters using the `R`-package `SpatialExtremes`. For the marginal d-GEV distribution, we used a set of parameters characteristic of those d-GEV distributions fitted from the stations in the observational dataset. Then, to fulfill the constraints of annual rainfall maxima averaged over durations d_i, $i = 1, ..., k$, we transformed the initial simulated data from having unit-Fréchet margins to GEV margins that follow the constraints given by (Nadarajah et al., 1998; Koutsoyiannis et al., 1998)

$$\zeta(d_i) = \zeta_0\,, \tag{4.18}$$

$$\sigma(d_i)(d_i/d_j) \leq \sigma(d_j) \leq \sigma(d_i), \quad d_i \leq d_j \quad \forall i, j, \quad \text{and} \tag{4.19}$$

$$\mu(d_i)(d_i/d_j) \leq \mu(d_j) \leq \mu(d_i), \quad d_i \leq d_j \quad \forall i, j\,. \tag{4.20}$$

This transformation uses the response surface described in Equations (4.12) and (4.13). We simulate 40 values for each dataset (representing 40 years) for $d = (1, 3, ..., 119, 120)$ h. Following Zheng et al. (2015), three sets of dependence parameters were used: $(\rho = 1, \alpha = 1)$ for weak dependence, $(\rho = 0.5, \alpha = 0.5)$ for moderate dependence, and $(\rho = 0.5, \alpha = 0.2)$ for strong dependence. For each parameter set, we generate 1000 realizations. An issue we encountered is that the nature of the simulated data allowed for rainfall intensity series that did not strictly follow the constrains given by Equations (4.18)–(4.20), however, we considered this to happen at a frequency that would not affect the final result of the study.

Observations

We used six rain gauge stations from the Wupper Catchment in Germany (Figure 4.1). All the stations have hourly values of accumulated precipitation height for the period 1979–2016. The stations are all within an elevation range of 250 m, and horizontally the shortest and longest distance between each station is 7 Km and 34 Km, respectively. We chose this dataset as it has a large number of years with high-frequency (hourly) measurements. Furthermore, the stations range from the Bergisches Land to the Upper Rhine Plain and therefore represent very well the different altitudes of the catchment. Figure 4.1 shows the distribution of rainfall maxima for this period.

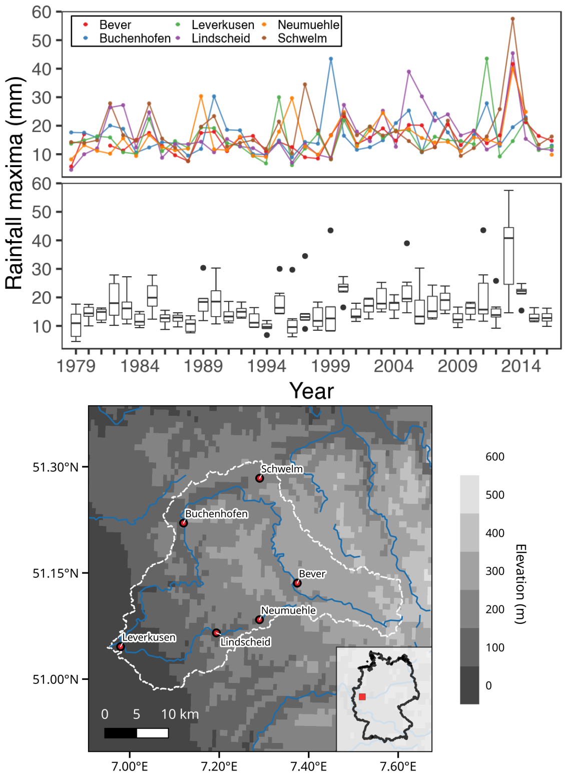

FIGURE 4.1: **Left**: (Lower panel) Distribution of the annual rainfall maxima at a 1-h accumulation duration plotted against time. Each boxplot shows the distribution of the pooled maxima from the six stations used for the case study in the Wupper catchment region. (Upper panel) Time series of the annual rainfall maxima, showing the values of each station as a different color. **Right**: Map of the Wupper catchment (dashed line) showing the location of all 6 stations; the lower-right corner shows the location of the catchment within Germany.

We obtain the annual block maxima $i(d_j)$ of the averaged rainfall intensity $\zeta_d(t)$ over the time window d_j for each station using Equation (4.1). For estimation purposes, we used durations $d = (1, 3, ..., 119, 120)$ h. The decision for the cut-off value of 120 h was based on previous studies on IDF curve estimation (Tyralis et al., 2019; Stephenson et al., 2016). By visual inspection of the corresponding Quantile Quantile (QQ)-plot, we ensure good agreement of the resulting $i(d_j)$ block maxima for all 6 stations with the GEV distribution. A small subset of the QQ-plots can be seen in Figure 4.10.

4.3 Results

We present the results for the case study in the Wupper region of Germany first, followed by the results of the simulation study.

4.3.1 Case Study

Structure of the Extremal Dependence

Figures 4.2 and 4.3 show a comparison of the pairwise extremal coefficient θ derived from the parameters of the MS-GEV approach (Equation (4.14), red line) and from a nonparametric estimate (dots) to assess how well the MS-GEV approach captures the observed asymptotic dependence. We used different distance measures h for each plot. Figure 4.2 uses the euclidean distance h_e (Equation (4.9)), and Figure 4.3 uses the log-distance h_l (Equation (4.10)). Additionally, the different colors show the lower distance d_i used for each duration pair (d_i, d_j), where $i < j$, and $(i > 0, j > 0)$.

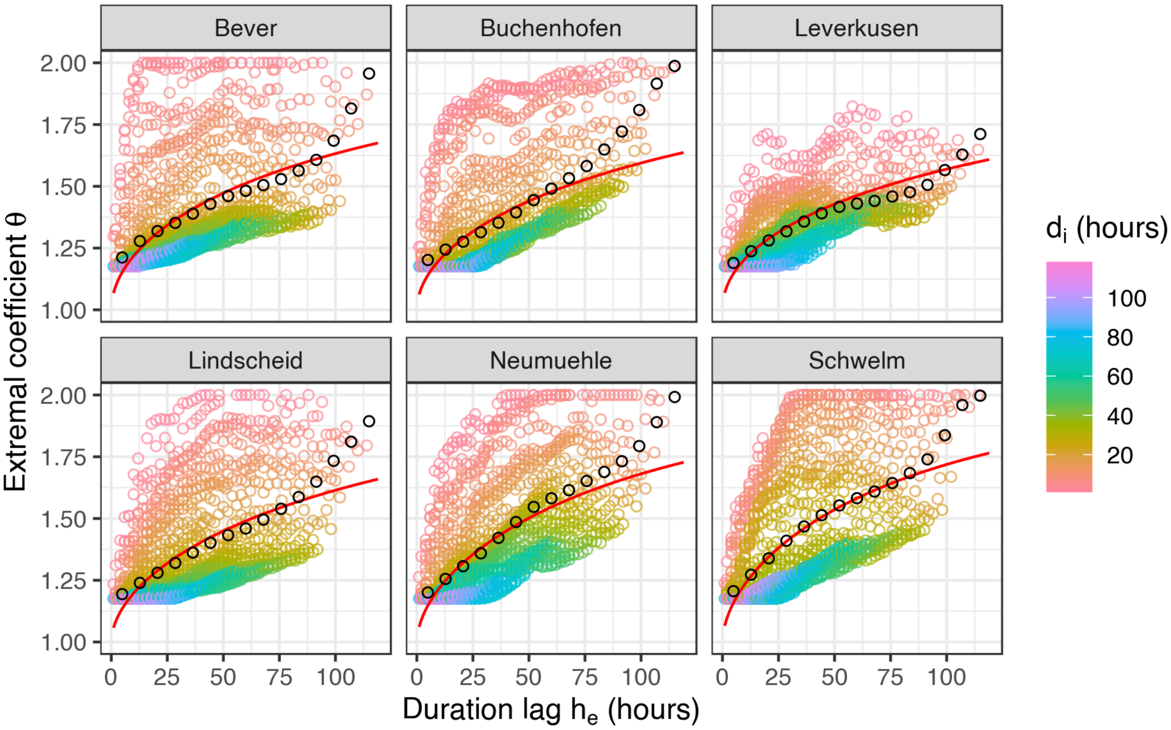

FIGURE 4.2: Nonparametric (dots = $\hat{\theta}_{emp}$) and parametric (solid line = θ_{BR}) estimates for the pairwise extremal coefficient θ using the euclidean distance h_e. The estimated nonparametric mean of θ for each duration lag bin is shown as black dots. Each color represents the lower duration d_i used for each duration pair.

For the euclidean distance h_e (Figure 4.2), the values of θ_{BR} are close to the binned means of $\hat{\theta}_{emp}$ (black circles) with respect to the scattering of the individual empirical estimates (color circles) for all stations. Nevertheless, the empirical point clouds show a remarkably high variability of θ around the mean value. To name one example, in station Buchenhofen, the distance of $h_e = 25\,\text{h}$ has a range of very different values of θ, spanning from 1.2 to 1.8. Thus, using a model that approximates the empirical mean-binned values in this case does not appear to be a meaningful representation of the overall variability of the dependence between durations.

Furthermore, each set of duration pairs with a fixed lower duration d_i in Figure 4.2 (represented by different colors) seems to follow a different path as the distance h_e grows. In particular, for duration pairs with a short lower duration ($d_i \leq 10\,\text{h}$), the value of θ grows much faster than duration pairs with longer lower durations ($d_i > 10\,\text{h}$). This suggests that several different regimes of dependence coexist, which are not only a function of the distance h_e, but also of the magnitude of the durations used. This is a transgression of the assumption that the extremal coefficient, and by extension, the semivariogram model of Equation (4.8) is isotropic (i.e., γ should only be a function of h). It is thus not evident that using a dependence model for the MS-GEV approach based on the euclidean distance h_e adequately captures the asymptotical dependence between durations seen on the data used for this study.

The resulting extremal coefficient using the log-distance h_l is shown in Figure 4.3. The point cloud shows a remarkably lower variability around the binned means than those of Figure 4.2. Furthermore, the empirical values of $\hat{\theta}_{emp}$ appear to follow the same regime, suggesting that θ is isotropic when using the log-distance. The shape of the point cloud seems to be appropriately captured by θ_{BR} for duration ratios $d_j/d_i \lesssim 6$ (i.e., when the upper duration d_j is around six times the lower duration d_i). However, the parametric model deviates from the empirical estimates as h_l increases. With the exception of station Leverkusen, the parametric model consistently overestimates the strength of the asymptotical dependence for the pairs with a duration ratio $d_j/d_i > 6$.

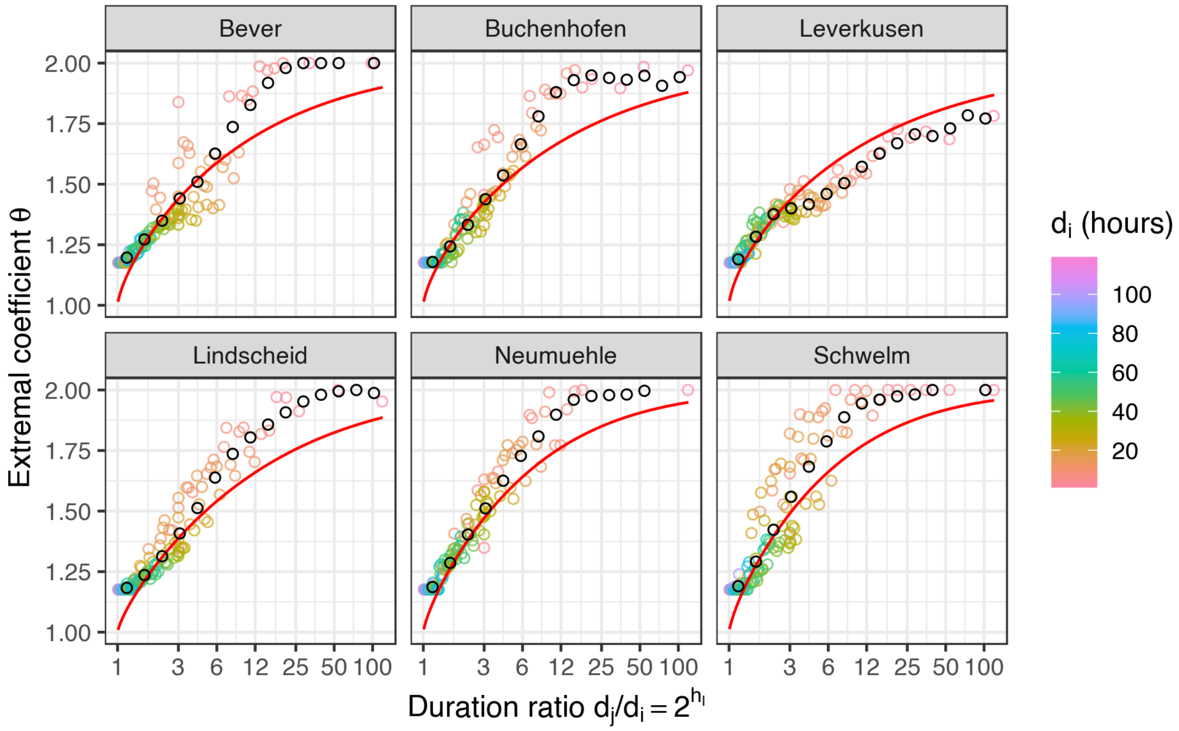

FIGURE 4.3: Nonparametric (dots = $\hat{\theta}_{emp}$) and parametric (solid line = θ_{BR}) estimates for the pairwise extremal coefficient θ using the logarithmic distance h_l. For ease of interpretation, the values of the log-distance h_l in the x-axis were transformed to the duration ratio d_j/d_i. The estimated nonparametric mean of θ for each duration ratio bin is shown as black dots. Each color represents the lower duration d_i used for each duration pair. Notice the difference in the variability of the point clouds when compared to those of Figure 4.2.

In light of the above results shown by Figures 4.2 and 4.3, we decided to use the log-distance h_l when estimating the semivariogram of the MS-GEV approach in all of the following calculations.

Estimation of IDF Curves

Table 4.1 shows the parameter estimates for the MS-GEV approach for all Wupper catchment stations, estimated from durations $d = (1, 3, ..., 119, 120)$ h. The estimates for the range parameter ρ are reasonably consistent across all stations. Their value of ~ 2 suggests that, for all stations, the rainfall intensities of different durations become asymptotically independent when their ratio d_j/d_i is larger than $(2^2 = 4)$.

TABLE 4.1: Parameter estimates from the MS-GEV approach for stations in the Wupper catchment using durations $d = (1, 3, ..., 119, 120)$ h.

Station	α	ρ	μ_0	σ_0	ξ_0	c
Bever	1.42	2.09	13.32	2.84	0.03	−0.58
Buchenhofen	1.39	2.22	13.38	3.03	0.02	−0.63
Leverkusen	1.32	2.19	10.74	2.34	0.05	−0.64
Lindscheid	1.54	2.44	13.69	3.23	0.06	−0.64
Neumuehle	1.54	1.84	13.52	2.74	0.04	−0.60
Schwelm	1.53	1.74	13.07	2.77	0.05	−0.62

Figure 4.4 shows the IDF curves following from Equation (4.3) for the MS-GEV approach (solid lines) and compares them to the IDF curves based on the rd-GEV approach (dashed lines). Return levels from the MS-GEV are for the most part consistently higher, which is in agreement with Tyralis et al. (2019).

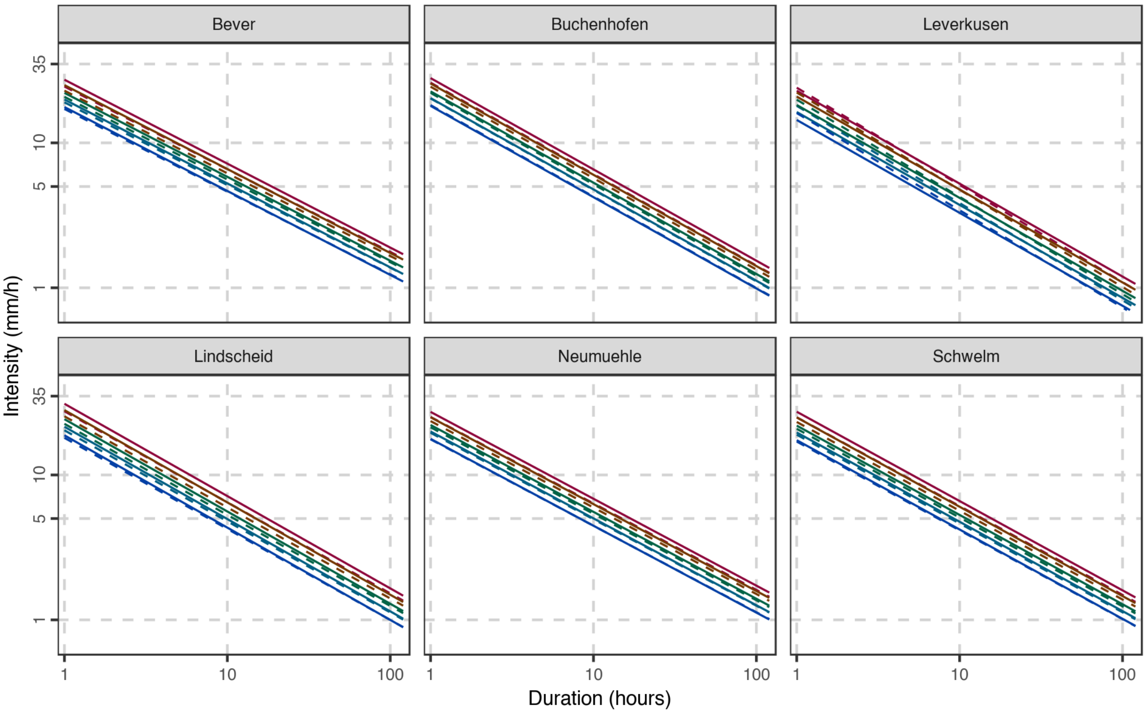

FIGURE 4.4: Comparison of IDF curves for the MS-GEV (**solid line**) and rd-GEV approach (**dashed line**) for all stations. Different colors represent different return periods; from bottom to top: (5, 10, 20, 40, 100) years.

To compare the results of the IDF curves using the euclidean distance h_e instead of the log-distance h_l, a plot comparing the resulting 100-year return level of both distances is shown in Appendix 4.A.

Performance Averaged Over All Durations

Figure 4.5 shows the cross-validated QSI evaluating the performance of the MS-GEV approach compared to the rd-GEV one. Similar behavior can be seen for all stations. For short return periods, the QSI is close to zero (denoting that both models are equally good), increasing towards longer periods. Station Lindscheid profits most from the MS-GEV with a 20% increase in skill for the 100-year return level. Only for a few points is skill negative, mostly for the shorter return periods.

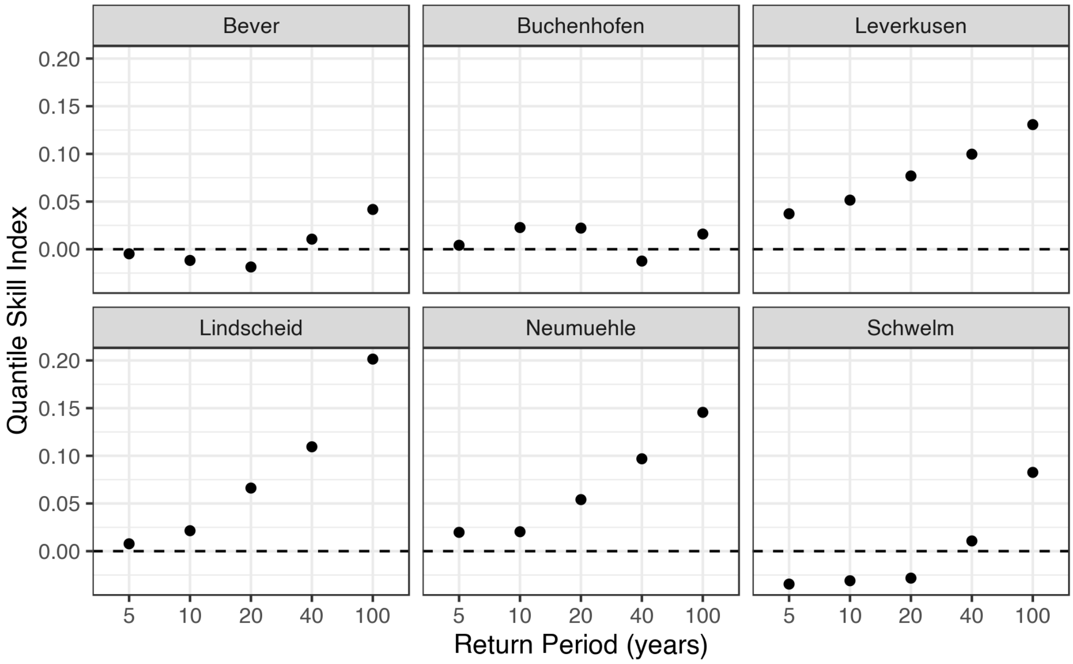

FIGURE 4.5: Quantile Skill Index comparing the MS-GEV versus the rd-GEV approach for all stations in the Wuppertal catchment. Positive values favor the MS-GEV approach.

Performance for Individual Durations

For a detailed comparison, we show the QSI conditioned on duration in Figure 4.6. The QSI varies appreciably over different durations for a given return period. For short durations ($d < 10\,\mathrm{h}$), the QSI is mostly positive for all return periods; for long durations $d > 100\,\mathrm{h}$, it is mostly negative. Gauge Lindscheid exhibits positive skill for many more combinations of durations ($d < 50\,\mathrm{h}$) and return periods $T > 40$ years; station Bever is the only station not showing a positive skill for very short durations, showing a general loss of skill for all but the shortest durations.

4.3.2 Simulation Study

We studied the effect of the level of dependence on the performance of the MS-GEV approach. To this end, we used synthetic data with known dependence parameters (ρ and α in Equation 4.8) and estimate the performance for various levels of dependence. Figure 4.7 shows the cross-validated QSI obtained using Equation (4.17), using the averaged QS values over all $d = (1, 2, ..., 120)\,\mathrm{h}$ durations for 1000 replications as box-whisker plots. Similar to the case study, the distance h used for the semivariogram of the Brown-Resnick process was the log-distance h_l (Equation 4.10).

The variation of skill among the replicates increases with increasing return period. The median is consistently positive (MS-GEV superior to rd-GEV) but below 0.05, suggesting that the dependence does not impact strongly on the performance results. The strong level of dependence leads to a slightly higher median skill but also to a considerable larger variance.

Figure 4.8 shows the QSI calculated with the model and reference QS averaged over all replicates for individual durations. The results are similar to the pooled QSI over all durations of Figure 4.7: The QSI increases with return period and dependence, staying below QSI = 0.1 for a 100 year return period and the strongly dependent series. This again suggests that the level of dependence has little impact on the performance of the return levels from the IDF curves.

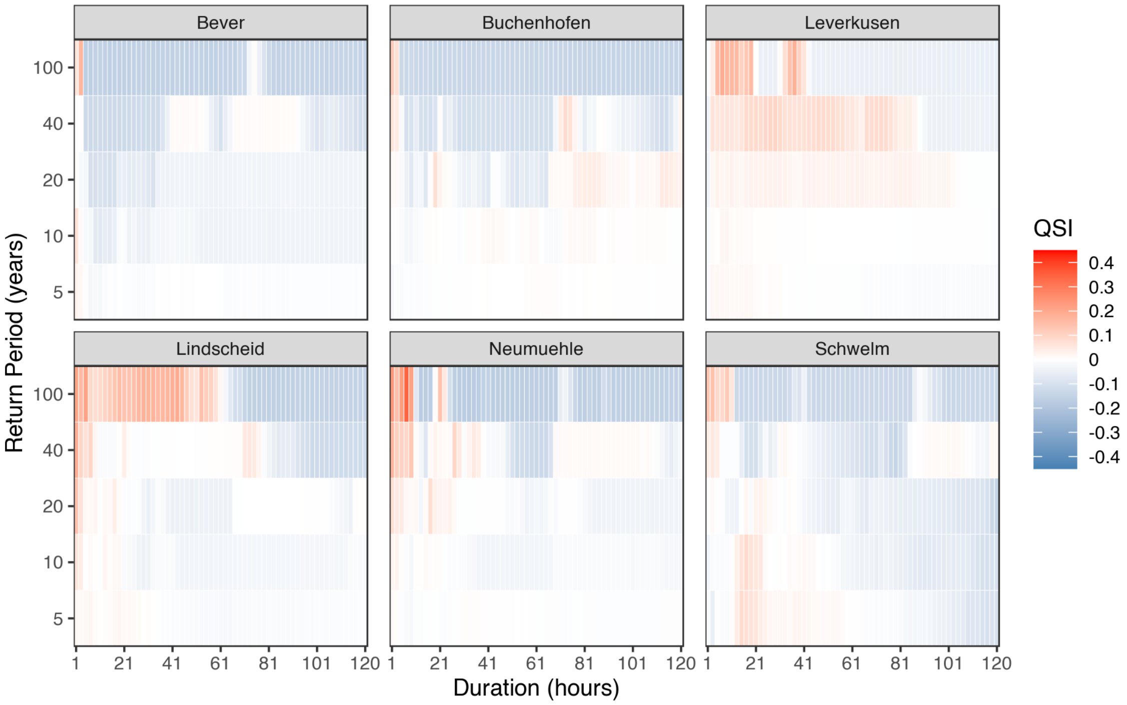

FIGURE 4.6: Quantile Skill Index conditioned on duration comparing the MS-GEV (using log-distance h_l) versus the rd-GEV approach for all stations in the Wuppertal catchment. Positive values favor the MS-GEV approach for different durations.

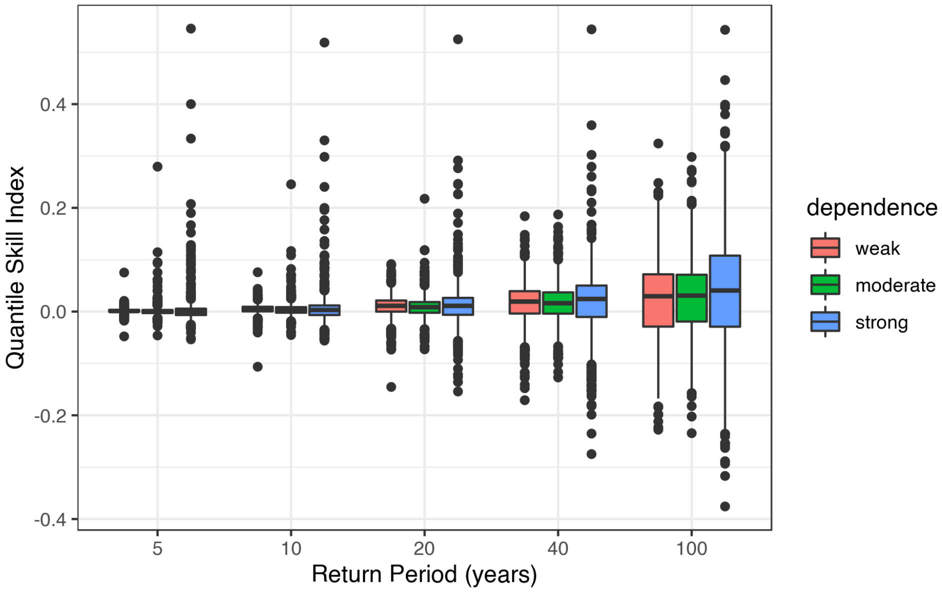

FIGURE 4.7: Quantile Skill Index calculated from QS averaged for all durations $d = (1, 2, .., 120)$ h, comparing the MS-GEV versus the rd-GEV approach as a function of simulated data's dependence parameter. Positive values of the QSI favor the MS-GEV approach. Each boxplot represents the distribution from the results of 1000 simulations.

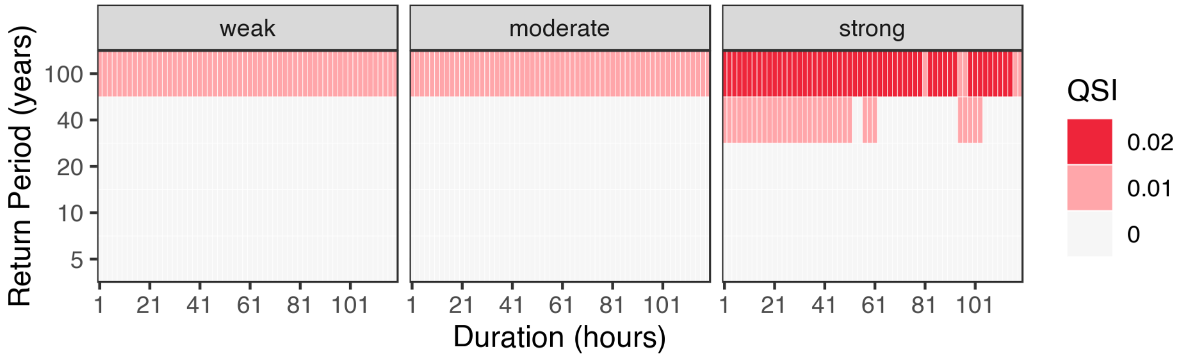

FIGURE 4.8: Quantile Skill Index, as in Figure 4.7, but showing the quantile score index (QSI) as a function of return period and duration.

4.4 Discussion

In this study, we obtained a measure of relative performance for return level estimates of IDF curves for various durations involving a max-stable process that allows for asymptotic dependence between durations (MS-IDF), compared to a model that assumes independence (rd-GEV). To do so, we built upon the previous study of Tyralis et al. (2019), who focused on the theoretical basis of using max-stable processes for modeling IDF-curves and did not investigate the consequences in terms of model performance.

To investigate the possible impact of the asymptotical dependence, we evaluated and compared the performance of estimating IDF curves with two different approaches: i) using a max-stable process to describe the dependence between rainfall intensity maxima for different durations and ii) assuming independent maxima. We evaluated individual performance based on a score that allows us to focus on the tail of the distribution, namely the quantile score (QS). The comparison between the MS-GEV and rd-GEV approach was carried out with the quantile score index (QSI), an index based on the QS skill score. The QSI enables us to quantify a gain/loss in performance when accounting for asymptotic dependence in return level estimation.

The results of Figure 4.2 showed that the resulting pairwise extremal coefficient was non-isotropic when using the euclidean distance h_e as the measure h in the semivariogram of the Brown-Resnick process for the MS-GEV approach. Thus, in contrast to the approach of Tyralis et al. (2019), we explored the use of a logarithmic distance measure instead of an euclidean one for the semivariogram of the Brown-Resnick process. Figure 4.3 shows that this was a reasonable choice, with the resulting parametric extremal coefficient properly capturing the variability of the empirical extremal coefficient around its binned means.

A simulation study suggests a minor advantage of the MS-GEV approach over the rd-GEV, particularly for long return periods (large quantiles). This advantage increases with the strength of the dependence, but remains low (QSI ≤ 0.1) even for the strongest level of dependence. A complementary case study for six gauges in the Wupper catchment (Germany) corroborates a general advantage for long return periods when averaging the performance measure over all durations. A detailed investigation of performance conditioned on durations shows, for our case study, that this advantage results mostly from short ($d \lesssim 10\,\mathrm{h}$) and, in some cases, from intermediate ($20\,\mathrm{h} \lesssim d \lesssim 50\,\mathrm{h}$) durations, depending, however, on the specific station.

The presented findings support the idea of Tyralis et al. (2019) that max-stable processes are valuable models for IDF curve estimation. The simulation and case study results indicate that an increase in skill with the MS-GEV approach is mostly found for large quantiles and short to medium durations. This effect might be related to the fact that shorter durations have a larger number of pairs than the longer durations for the pairwise likelihood. This means that the longer durations could be underrepresented in the current likelihood expression, leading to a better fit of the model for the shorter durations for the MS-GEV approach. However, extreme events of short durations have usually a higher impact on society than those of long durations. Therefore, we believe that this underrepresentation does not necessarily detract value from using the MS-GEV approach. When focusing on large quantiles and short durations, it might be worth taking the added computational expense and model complexity in exchange for increased skill in the return levels for such events.

In addition, the simulation study demonstrated that the level of dependence had a modest impact on the overall performance and variability of the MS-GEV approach. These findings contradict those of Zheng et al. (2015), who found that the dependence strength did not influence the performance of the return level estimates. However, our study focused on a different dependence structure than that of Zheng et al. (2015), who used a spatial approach, in contrast to our duration space. This contradiction may be associated with the particular form of the asymptotic dependence for the different durations of $i(d)$, which stems from the nature of $i(d)$ as random ordered variables. As previously studied by (Nadarajah et al., 1998; Nadarajah et al., 2019), when two random variables are intrinsically ordered, each margin's distribution is affected, something that has to be taken into account. While the response surfaces described in Equations (4.12)–(4.13) take this ordering into account, the dependence structure given in Equation (4.8) does not, a factor that could explain the contradiction with previous studies.

The parametric estimate of the extremal coefficient θ_{BR} from the MS-GEV approach shown in Figure 4.3 appears to be a reasonable fit for the empirical values of θ when the ratio of the duration pair is around ($d_i/d_j \lesssim 6$). However, as the upper duration gets around six times larger than the lower duration, the model consistently overestimates the strength of the dependence. Figure 4.3 also showed that duration pairs (d_i, d_j) involving short lower durations $d_i \leq 10\,\mathrm{h}$ had an extremal coefficient that approached independence (i.e., $\theta = 2$) faster than those with longer lower durations as h_l increased. This suggests that the dependence between rainfall maxima of different aggregation durations is short-ranged, in particular, for the shorter durations.

For an operational use of this approach, uncertainty estimates for the quantiles (IDF curves) need to be incorporated. In this regard, Mélèse et al. (2018) and Ganguli et al. (2017) showed that a Bayesian Hierarchical Model approach resulted in reliable credibility intervals for IDF curves. For the rd-GEV approach, we investigate a bootstrap-based method for estimating the uncertainty of d-GEV based return levels in a different study (Ulrich et al., 2020). We also limited our study to using data with hourly frequency, resulting in the value of the (sub-hourly) ν parameter of the d-GEV to be artificially set to zero (what we called the rd-GEV). Further studies would benefit from using sub-hourly frequencies, allowing ν to vary freely. Moreover, we assumed that the stations' data was stationary, ignoring the possible effects of climate change. Several studies have shown that accounting for nonstationarity has a measurable effect on the return levels estimates (Ganguli et al., 2017; Ganguli et al., 2019; Padulano et al., 2019).

The method presented in this study is a straightforward and practical manner of estimating IDF return level performance based on a max-stable process. The results for the case study encourage to investigate its performance within a larger geographical setting. Furthermore, it seems worth implementing more flexible functions describing the variability of GEV parameters with duration as used for the response surface, e.g., an additional parameter accounting for different behavior, particularly for short durations, as suggested in Koutsoyiannis et al. (1998). Another critical issue for future studies is to explore how different dependence structures could impact the performance of the estimation. Furthermore, a comparison with the recent developments in the use of covariates for the rd-GEV approach could result in better skill for such an approach compared with our current MS-GEV one (Ulrich et al., 2020).

4.5 Conclusions

Our findings indicate that the use of models that allow for the asymptotic dependence between rainfall maxima of different durations when estimating IDF curves can lead to moderately better return level estimates, particularly for long return periods (100 years, generally of considerable interest) and short durations ($d \leq 10\,\mathrm{h}$). However, the former comes at the expense of the added complexity of modeling the asymptotic dependence. Furthermore, this asymptotical dependence seems to be short-ranged for the short durations. We, therefore, recommend the use of the simpler univariate-EVT methods assuming independence between durations for a single station when the main goal is obtaining return levels for a wide range of short and long durations from IDF-curves.

Funding information: The research was partially supported by Deutsche Forschungsgemeinschaft (DFG) through grant CRC 1114 "Scaling Cascades in Complex Systems," Project A01 "Coupling a multiscale stochastic precipitation model to large scale flow." O.J. acknowledges support by the Mexican National Council for Science and Technology (CONACyT) and the German Academic Exchange Service (DAAD). J.U. and O.J. acknowledge support by Deutsche Forschungsgemeinschaft (DFG) within the research training program *NatRiskChange* (GRK 2043/1).

Acknowledgments: We would like to acknowledge the Wupperverband (Untere Lichtenplatzer Str. 100, 42289 Wuppertal, North Rhine-Westphalia, Germany) for providing the precipitation time series. We also acknowledge the contribution of Hans Van de Vyver on the use of the log-distance for the variogram. The publication of this article was funded by Freie Universität Berlin.

Appendices

4.A Inference from the Brown-Resnick Max-Stable Process

Consider a stochastic process $\{X(d) : d \in \chi\}$, where χ is a compact subset of $\mathbb{R}^D, D \geq 1$, and a Poisson process Π with intensity $d\zeta/\zeta^2$ on $(0, \infty)$. Let $X_i(d)$ be independent realizations of a process $X(d)$ with $E[X(d)] = 1$, and let $\zeta_i \in \Pi$ be points of the Poisson process. A simple max-stable process is then given by

$$Z(d) = \max_{i \geq 1} \zeta_i X_i(d),\ d \in \chi\,. \tag{4.21}$$

Smith (1990b) proposed a useful analogy to interpret this kind of max-stable process as the so-called *rainfall-storms* interpretation. In this interpretation, ζ represents the overall intensity of a rainfall storm that impacts the region χ, and $\zeta X(d)$ corresponds to the total amount of rainfall for the storm centered at position d. A max-stable process would then be the pointwise maxima (taken over each point in χ) over an infinite number of storms.

For the Brown-Resnick process, we follow the proposal from Kabluchko et al. (2009), where $X_i(d) = \exp(e_i(d) - \frac{1}{2}\sigma^2(d))$. Here, $e_i(d)$ is a Gaussian process with stationary increments and semivariogram $\gamma(h) = \frac{1}{2}\mathrm{Var}(e_i(d+h) - e_i(d))$. The bivariate distribution function for the Brown-Resnick process (Kabluchko et al., 2009; Davison et al., 2013), is

$$\Pr[Z(d_1) \leq z_1, Z(d_2) \leq z_2] = \exp\left[-\frac{1}{z_1}\Phi\left(\frac{\sqrt{\gamma(h)}}{2} + \frac{1}{\sqrt{\gamma(h)}}\log\frac{z_2}{z_1}\right) - \frac{1}{z_2}\Phi\left(\frac{\sqrt{\gamma(h)}}{2} + \frac{1}{\sqrt{\gamma(h)}}\log\frac{z_1}{z_2}\right)\right], \tag{4.22}$$

where $\bar{z}$ follows a unit Fréchet distribution, Φ denotes the standard normal distribution function, h is a measure of the "distance" between duration pairs (d_i, d_j) (given in this study by Equation 4.10), and the semivariogram γ is defined in Equation (4.8).

For inference purposes, we applied the commonly used pairwise likelihood proposed by Padoan et al. (2010), given by

$$L(\psi|i_1(d_1), ..., i_n(d_k)) = \sum_{t=1}^{n}\sum_{j=1}^{k-1}\sum_{j'=j+1}^{k} \log f(i_t(d_j), i_t(d_{j'})|\psi)\,, \tag{4.23}$$

where $\psi = [\mu, \sigma, \xi, \rho, \alpha]$ represents the parameters to estimate. Here each term $f(i_k(d_j), i_k(d_{j'})|\psi)$ is the (appropriately transformed) bivariate density function derived from Equation (4.22) for observed maxima $\bar{i}(d)$ at durations d_j and $d_{j'}$. Note that the first three parameters in ψ are the univariate parameters of the GEV distribution, unique for each duration, while the last two parameters of ψ are the parameters of the Brown-Resnick process, which model the asymptotic dependence.

4.B Comparison of 100-year Return Level between Euclidean and Log-Distance for MS-GEV Approach

Figure 4.9 shows a comparison of the resulting 100-year return level intensity resulting from the MS-GEV approach using the euclidean distance h_e and the log-distance h_l. To facilitate the comparison, it shows the ratio $q_hl(0.99)/q_he(0.99)$, where $q(0.99)$ is the quantile corresponding to the probability of 0.99, that is, the corresponding 100-year return level.

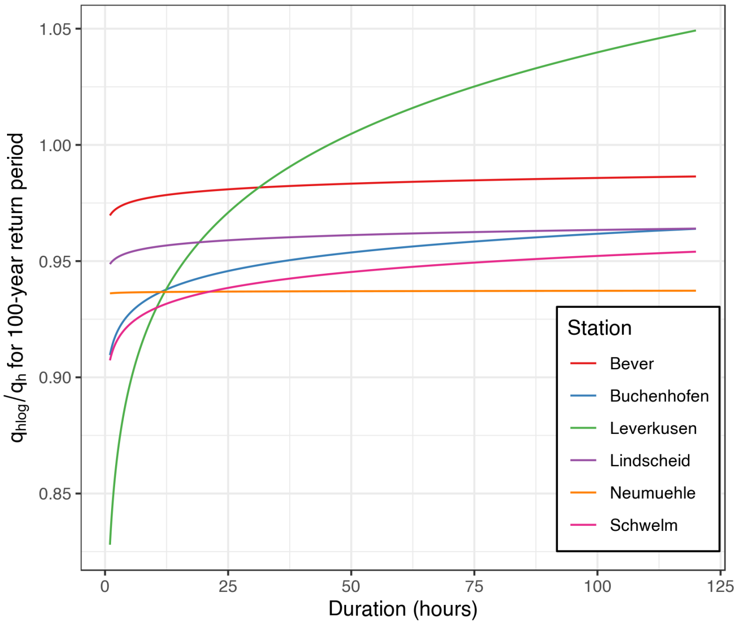

FIGURE 4.9: Comparison of the 100-year return level intensity for the MS-GEV approach using the euclidean distance h_e and the log-distance h_l.

Surprisingly, although the variability of θ around its mean is remarkably different when using h_l instead of h_e, the resulting 100-year return levels are very similar for both distances. The single exception was for station Leverkusen, which is the only station of the catchment located in the Upper Rhine Plain; the change observed for this stations was however still relatively small. However, this result is only accounting for the point estimates of the return levels. As seen in Figures 4.2 and 4.3, the variability of the extremal dependence is much higher when using the euclidean distance h_e than the log-distance h_l. Thus, we believe that the resulting uncertainty for the MS-GEV approach should be lower when using h_l instead of h_e. Nevertheless, as mentioned in the limitations of our study, we did not perform any estimation of the uncertainty.

4.C QQ-Plots for Selected Stations and Durations

Figure 4.10 shows the QQ-plots for validation of the marginal fits of the GEV distribution for four intensity maxima series $\bar{\imath}(d)$, with $d = (1, 3, 48, 72)$ h, for three stations (Bever, Leverkusen, and Neumuehle.). The closer that the points are aligned to the identity line, the better the fit of the GEV distribution.

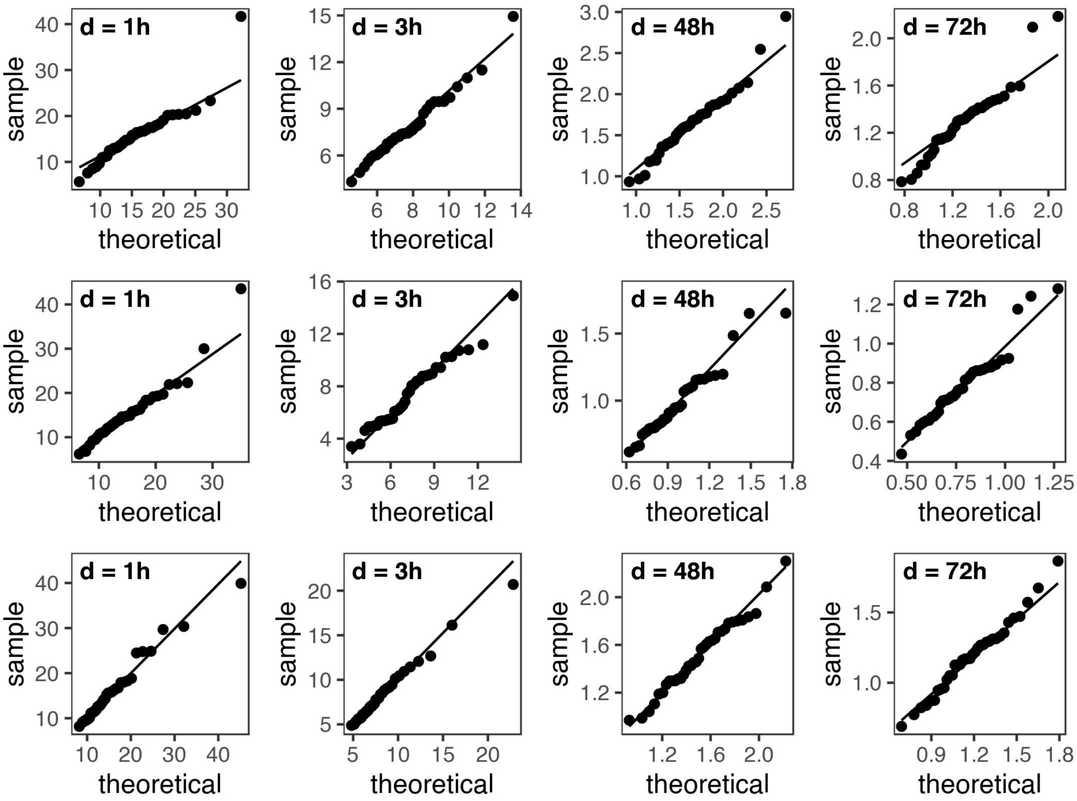

FIGURE 4.10: QQ plots for model checking of the marginal distributions for three stations: Bever (**top row**), Leverkusen (**middle row**), and Neumuehle (**bottom row**). The duration used for the accumulation of the rainfall maxima is indicated in each plot.

5 Implications of Modeling Seasonal Differences in the Extremal Dependence of Rainfall Maxima

Publication details

Jurado, Oscar E., Oesting, Marco, & Rust, Henning W. (2022). **Implications of Modeling Seasonal Differences in the Extremal Dependence of Rainfall Maxima.** Submitted to *Stochastic Environmental Research and Risk Assessment*. Preprint available in http://arxiv.org/abs/2207.03993

Author contributions

Conceptualization: OEJ, MO, HR; Methodology: OEJ, MO, HR; Formal analysis and investigation: OEJ; Software - OEJ; Visualization: OEJ; Writing - original draft preparation: OEJ; Writing - review and editing: MO, HR; Funding acquisition: HR; Resources: HR; Supervision: MO, HR.

The text for this chapter contains the original version of the manuscript submitted to the SERRA journal. A revised version of this manuscript that includes changes suggested by two reviewers was submitted for final publication.

Abstract

For modeling extreme rainfall, the widely used Brown-Resnick max-stable model extends the concept of the variogram to suit block maxima, allowing the explicit modeling of the extremal dependence shown by the spatial data. This extremal dependence stems from the geometrical characteristics of the observed rainfall, which is associated with different meteorological processes and is usually considered to be constant when designing the model for a study. However, depending on the region, this dependence can change throughout the year, as the prevailing meteorological conditions that drive the rainfall generation process change with the season. Therefore, this study analyzes the impact of the seasonal change in extremal dependence for the modeling of annual block maxima in the Berlin-Brandenburg region. For this study, two seasons were considered as proxies for different dominant meteorological conditions: summer for convective rainfall and winter for frontal/stratiform rainfall. Using maxima from both seasons, we compared the skill of a linear model with spatial covariates (that assumed spatial independence) with the skill of a Brown-Resnick max-stable model. This comparison showed a considerable difference between seasons, with the isotropic Brown-Resnick model showing considerable loss of skill for the winter maxima. We conclude that the assumptions commonly made when using the Brown-Resnick model are appropriate for modeling summer (i.e., convective) events, but further work should be done for modeling other types of precipitation regimes.

5.1 Introduction

The statistical modeling of extreme precipitation is essential for designing public hydrological infrastructure and urban planning worldwide (Durrans, 2010). This approach typically combines observed information from past events with models from Extreme Value Theory (EVT) to give a probabilistic estimate of the magnitude and frequency of future extreme precipitation events (Coles, 2001). Information about past events usually comes from ground observations (e.g., rain gauges), operated mainly by local weather services. For a typical EVT application, information from rain gauges is used to fit the parameters of a max-stable distribution (such as the Generalized Extreme Value (GEV) distribution), from which information on the magnitude and frequency of events in the far-right tail of the distribution can be elicited. The ultimate goal of EVT analyses is then to provide adequate estimates of these estimates along with their uncertainties. These estimates are commonly communicated to decision-makers either in the form of return periods for certain return levels (i.e., “1-in-n years event”) or as a more general quantity like the probability of exceedance and risk of failure over a given design life period (Serinaldi, 2015; Rootzén et al., 2013).

A common problem when modeling extreme rainfall is that no observations exist in many locations where information from statistical modelling of extreme events would be useful. However, on many occasions, observations exist near unobserved locations. This setting is the same as in Geostatistics, except that the focus is on extremes and max-stable distributions in this case. This problem has given way to different EVT models that allow interpolation of estimates to unobserved locations, usually englobed within the term “Spatial Extremes”. Spatial Extremes models follow a very similar theoretical background to the methods of Geostatistics and can be thought of as extensions of Geostatistics, but for extremes (Davison et al., 2012a).

Most Spatial Extremes and Geostatistical models use the so-called first law of Geography: “everything is related to everything else, but near things are more related than distant things.” (Tobler, 1970). In other words, there exists a particular covariance function that depends on the distance between points with observations. Spatial models use the observations from the different locations to fit a covariance function that describes how much two or more variables change as a function of some distance metric. Thus, covariance functions describe the spatial dependence between the observed locations. In the case of Spatial Extremes, the corresponding analog to the covariance function (e.g., the tail-dependence function) is combined with an appropriate model for extremes to fit a joint distribution for the different locations and, in some cases, to also obtain the estimates of the marginal parameters in each location. Interpolation to unobserved locations is then achieved by combining the fitted tail-dependence function with the fitted model.

When dealing with block maxima stemming from observations fixed in space (e.g., rain gauges), a commonly used spatial extremes model is a max-stable process (Davison et al., 2015). Max-stable processes are an extension to infinite dimensions of univariate EVT models for block maxima (Padoan et al., 2010). Unlike univariate EVT models, there does not exist a single parametric family of max-stable processes to which block maxima always converge. Nevertheless, diverse parametric families with different tail-dependence functions have been proposed. For the spatial modeling of extreme precipitation, a commonly used family of max-stable processes is the Brown-Resnick family (Le et al., 2018; Davison et al., 2012b; Buhl et al., 2016). Brown-Resnick models are based on Gaussian processes with a tail dependence function that includes the geostatistical concept of the (semi-)variogram. Assuming that the underlying Gaussian process possesses stationary increments (i.e., it is only a function of the distance between different stations), the spatial dependence structure can be modeled exclusively with the variogram.

In previous studies using Brown-Resnick models for extreme rainfall, the focus has been on maxima that stem from summer events. This choice is typically justified as rainfall events in summer are usually the events with the largest magnitude and, thus, the ones with the most significant impact. Furthermore, these events are usually associated with convective activity, which for the study region is predominant in summer (Berg et al., 2013). Nevertheless, little work has been done to model extreme rainfall resulting from other types of events, such as stratiform ones. These events are relevant, as they could be the dominant types in other regions of the planet or of interest to different stakeholders. An essential aspect of our study is that these events differ significantly in terms of spatial and temporal extent, which likely leads to different spatial dependence structures, creating a need to research and improve our understanding of modeling extreme rainfall for maxima that originates from different types of events.

The present study aims to investigate how the extremal dependence changes for different rainfall-generating mechanisms and how this change influences the estimation of return levels down the line. The modeling of the extremal dependence is done via a Brown-Resnick max-stable process that accounts for the spatial variability of precipitation maxima in the Berlin-Brandenburg region. Instead of taking annual block maxima, we obtain semmi-annual block maxima from two seasons: winter and summer. We hypothesize that summer block maxima come mainly from convective events, while winter block maxima come from slow-moving storms that lead to stratiform and frontal events. This choice is justified based on the results of (Ulrich et al., 2021), who for the Wuppertal region in Germany found that convective events dominated in the summer months, while stratiform/frontal events dominated in the winter months. By using the semi-annual from the two seasons to fit the Brown-Resnick model, we estimate how the dependence changes with different rainfall-generating mechanisms.Moreover, we selected two temporal scales for each season to investigate the impact of processes with different time scales. We fit a Bayesian distributional linear model that assumes independence in space as a reference to our spatial model to discern the effects of the change in dependence on the estimated return levels. This reference model is compared with the spatial model within a verification framework.

This study is organized as follows: first, we present a review of the different types of rainfall-generating mechanisms that dominate in our study region. Then, we present the EVT methods we used to model extreme rainfall. Afterward, we introduce a verification framework to compare the different models, from which we present the results to determine whether a considerable change in the extremal dependence was observed and their consequences on the reported return levels.

5.1.1 Extremal dependence in rainfall data

Rainfall is the result of complex processes and interactions in the hydro- and atmosphere, involving processes from a wide range of temporal and spatial scales. In particular, for the midlatitude region, rainfall characteristics are heavily associated with the synoptic weather situation present when the event happened. Walther et al. (2006) classifies rainfall events as either frontal or convective based on synoptic-scale considerations. The synoptic-scale is usually defined as a length scale of around 2000 km to 20000 km, involving events that last from days to weeks. The distinction between synoptic or convective rainfall is relevant for the statistical modeling efforts, as

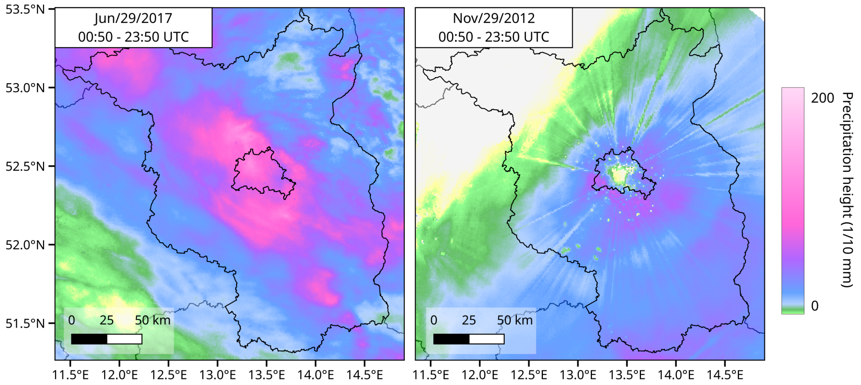
Jun/29/2017
00:50 - 23:50 UTC
Nov/29/2012
00:50 - 23:50 UTC
53.5°N
53.0°N
52.5°N
52.0°N
51.5°N
0 25 50 km
0 25 50 km
11.5°E 12.0°E 12.5°E 13.0°E 13.5°E 14.0°E 14.5°E
11.5°E 12.0°E 12.5°E 13.0°E 13.5°E 14.0°E 14.5°E
200
0
Precipitation height (1/10 mm)

FIGURE 5.1: Map showing daily accumulated precipitation height for two extreme precipitation events chosen arbitrarily for demonstration). Left: A summer convective event. Right: A winter frontal event. Data comes from the RADOLAN database made available from the DWD.

rainfall events associated with fronts (i.e., synoptic-scale) have very different temporal and spatial characteristics than those associated with isolated convective (typically mesoscale) events. By way of illustration, Orlanski (1975) characterizes thunderstorms as events lasting from half an hour up to a few hours covering areas of several km^2, and frontal events as having a lifetime of more than a day with a spatial spread of hundreds of km. These spatiotemporal characteristics can also influence the magnitude and the timing of extreme rainfall events; for example, Bohnenstengel et al. (2011) found that for a 25 km x 25 km region to the southeast of Berlin, extreme precipitation events occur more often in times of convective events than during times with frontal precipitation.

In the midlatitudes region, the type of dominant rainfall-generating mechanism changes during the year. For example, Berg et al. (2013) found that synoptic observations of convective events dominated during the summer seasons in four stations across Germany. In contrast, they found that most rainfall in the winter months resulted from stratiform clouds (commonly associated with frontal events). Thus, we predict that when looking at seasonal block maxima for a study region in Germany, summer maxima will originate mainly from convective events, while winter maxima will primarily originate from frontal ones. An example of this can be seen in Fig. 5.1, which shows the daily precipitation height in the Berlin-Brandenburg region for a convective event in summer (left) against that of a frontal/stratiform event in winter (right). For most stations in the domain, the semi-annual block maxima in the two corresponding seasons were attained for these two particular events, meaning they can be seen as extreme events. Extremal dependence in space arises when an extreme event is large enough to impact several rain gauges simultaneously. Therefore, the extremal dependence heavily depends on the spatial characteristics presented by the rainfall generating mechanisms. Thus, if these mechanisms change seasonally during the year, we expect the dependence structure to also change throughout the year.

5.2 Methods and Data

In this study, we perform the statistical modeling of extreme rainfall using the block maxima approach. This approach is based on the Fisher-Tippett-Gnedenko theorem, which states that under mild conditions block maxima of a sufficiently large block

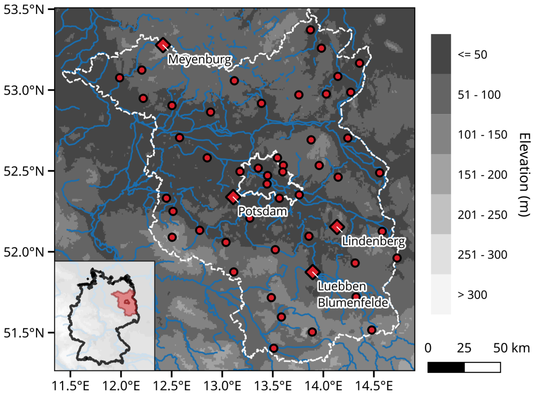

FIGURE 5.2: Map showing the location inside the Berlin-Brandenburg area of the DWD weather stations included in this study (red dots). The lower left inset shows the location of the study domain within Germany. Diamonds show the reference stations used to showcase results.

length of independently and identically distributed random variables can be approximately modeled by the generalized extreme value (GEV) distribution. When dealing with rainfall, block lengths of one month have been proven to be long enough to assure convergence to the GEV distribution (Fischer et al., 2017). Yearly block maxima can then be used to fit the parameters of the GEV for each rain gauge individually, resulting in the “zeroth-order” approach to modeling extreme rainfall in space. This pointwise approach, however, does not pool any information across stations and, therefore, cannot predict values for ungauged sites. Prediction of ungauged sites can be achieved by extending the pointwise GEV approach to include spatial covariates, which pools information from different locations, typically resulting in reduced uncertainties for the estimated parameters of the GEV (Ulrich et al., 2020). Nevertheless, this second approach ignores the spatial dependence in the data, resulting in a misspecified likelihood that consistently underestimates the uncertainty of the estimates. Our study extends this approach by using a max-stable process to include spatial dependence.

5.2.1 Data

We used accumulated hourly and daily precipitation height measurements (in mm) from 53 stations belonging to the German Meteorological Service (DWD) in the Berlin-Brandenburg region of Germany (Fig. 5.2). The data was acquired through the German Meteorological Service (DWD) Open Data Server using the `R`-package `rdwd` (Boessenkool, 2021). The stations were chosen to include only those that contained measurements with both hourly and daily periods. This choice reduced the available number of stations with daily measurements from 300 to 53. Reducing the total number of stations was considered necessary to ensure the fairness of the comparisons with results using stations with hourly measurements and to lower the computational burden needed to fit the models. The average distance between all station pairs was approx. 95 km, with a range of $[4, 245]$ km. The raw data contains further information about the type of precipitation measured (liquid or solid), but for the purpose of this study no discrimination was done with regard to precipitation type.

Two different periods were considered for this study: from 1970-2020 for the daily observations and 2004-2020 for the hourly observations. These periods were chosen in order to minimize the number of invalid pairs when using the pairwise likelihood (see Appendix 5.A).

At the location $s_j \in \mathcal{S}$, where $\mathcal{S}$ represents the geographical domain and $j = 1, ..., n$ is an index denoting the rain gauge, the data contains the accumulated rainfall values $(r_{d,1}(s_j), ..., r_{d,k_j}(s_j))$ in mm, where $d \in \{1, 24\}$ is an index for the duration of the considered precipitation events, namely, hourly or daily. Different gauges can have different lengths for the measurement period, so that k_j depends on the location s_j. The accumulated rainfall values were transformed to the average hourly/daily intensity values $(\zeta_{d,1}(s_j), ..., \zeta_{d,k_i}(s_j))$ in mm/h.

Following (Koutsoyiannis et al., 1998), the average hourly intensity data $\zeta_{d=1,\tau}(s)$ (where τ represents the time (in hours) of the observation) were aggregated to create the 12-hour accumulated precipitation intensity time series $\zeta_{d=12,\tau}(s)$ (in mm/h). This aggregation was necessary because a visual inspection of the pairwise extremal coefficient resulting from the hourly series strongly suggested that the data was asymptotically independent, which violates a major assumption for using max-stable processes. The lowest aggregation duration that did not show asymptotic independence was 12 hours. The 12-hour aggregated series is obtained using

$$\zeta_{d=12,\tau}(s) = \frac{1}{12} \sum_{i=0}^{11} \zeta_{d=1,\tau-i}(s), \tag{5.1}$$

which can be seen as a moving average with a time window of 12 hours. The aggregation described in Eq. (5.1) was done using the package `IDF` (Ulrich et al., 2020).

The 12-hour $\boldsymbol{\zeta}_{12}(s_j)$ and daily $\boldsymbol{\zeta}_{24}(s_j)$ average precipitation intensity series are then used to get four series of semi-annual block maxima series $(i^l_{d,t=1}(s_j), ..., i^l_{d,t=N_j}(s_j))$. In this case, the index t can be seen as indicating the year. These four series result from combining the two durations $d \in \{12, 24\}$ and the two seasons $l \in \{\text{sum}, \text{win}\}$ using the corresponding abbreviations for sumer and winter, respectively. The semi-annual block maxima were obtained using

$$i^l_{d,t}(s) = \max_{l^-_t < \tau < l^+_t} \zeta_{d,\tau}(s), \tag{5.2}$$

where l^-_t and l^+_t correspond to the beginning and end of either winter or summer for each year t. For this work, we consider summer as May, June, July, and August; winter is considered to be the months of January, February, November, and December. In order to avoid having winter block maxima that come from disconnected months, we shifted the $\boldsymbol{\zeta}_d(s)$ values of November and December to the following year, making the four winter months of any given calendar year come from the same "meteorological" winter. Note that for each instance of τ within the same type of season, i.e. summer or winter, and for each fixed duration $d \in \{12, 24\}$, we perceive $\zeta_{d,\tau}(\cdot)$ as independent realizations of some stochastic process $\{X(s) : s \in \mathcal{S}\}$ which will be the justification for the use of GEV distributions and max-stable processes for modeling the distribution of $i^l_{d,t}(s)$ below.

Figure 5.3 shows the temporal distribution of the semmi-annual block maxima for summer, i.e. $\boldsymbol{i}^{\text{sum}}_{12}(s)$ and $\boldsymbol{i}^{\text{sum}}_{24}(s)$, and winter, i.e. $(\boldsymbol{i}^{\text{win}}_{12}(s)$ and $\boldsymbol{i}^{\text{win}}_{24}(s))$ over the 53 stations. From Figure 5.3, it is apparent that the magnitude of the maxima changes depending on the season, with consistently larger values for summer events. The final length of the daily series is 50 years, while for the 12-hour series, it is 26 years.

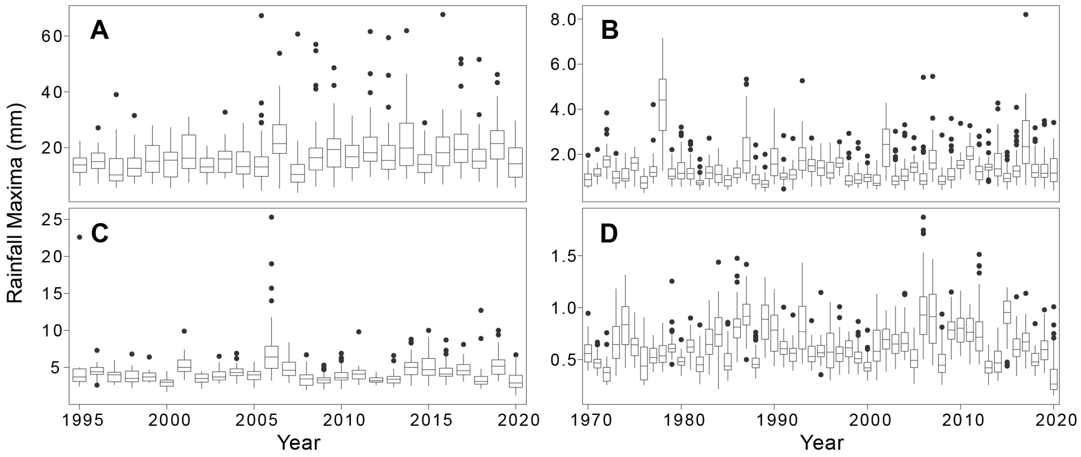

FIGURE 5.3: Boxplots showing the distribution of the rainfall semi-annual block maxima for the 53 stations included in this study. (A) 12-hour summer maxima, (B) daily summer maxima, (C) 12-hour winter maxima, and (D) daily winter maxima.

5.2.2 Characterizing extremal dependence

To explore how the bivariate extremal dependence changes for the block maxima derived from the summer and winter seasons, we used an estimate of the empirical pairwise extremal coefficient $\theta(s_j, s_{j'})$, which is a summary measure of dependence of a random two-dimensional vector $(X(s_j), X(s_{j'}))$ (Coles, 2001; Ribatet et al., 2016). The pairwise extremal coefficient can take values in the rage $[1, 2]$, where 1 denotes complete dependence and 2 asymptotic independence.

For each pair $(s_j, s_{j'})$ of locations of gauged stations, we estimate the empirical extremal coefficient $\hat{\theta}_{\mathrm{NP}}(s_j, s_{j'})$ using the non-parametric method proposed by Marcon et al. (2017), which Vettori et al. (2018) found to have the best overall performance compared to other empirical estimators. The estimation of $\hat{\theta}_{\mathrm{NP}}(s_j, s_{j'})$ is done with the R-package ExtremalDep (Boris et al., 2021). This method requires the specification of a polynomial order, for which a graphical analysis (not shown) found that a fixed value of $k = 20$ yielded the most appropriate values of $\hat{\theta}_{\mathrm{NP}}(s_j, s_{j'})$ for the different $\boldsymbol{i}_d^l(s)$ series.

5.2.3 Modeling of extreme rainfall

We follow a two-step approach to model the $\boldsymbol{i}_d^l(s)$ series. In the first step, we model the marginal distribution of the pooled data from all stations by including spatial covariates within a Bayesian distributional model (DM). For the second step, we extend the model of the first step with a max-stable process, allowing the model to capture the so-called "residual dependence" left from the first-step that arises from the extremal dependence (Cooley et al., 2012). We then compare the models from both steps using a forecast verification framework to study how the extremal dependence influences the estimates of the model parameters. We consider the BDM approach to act as a "control" compared to the max-stable process approach, allowing us to explore how the seasonal difference in the extremal dependence affects the estimates down the line.

Estimations made within the framework of extreme value statistics are usually made with small data samples, as extreme events are by definition rare. The small sample size, in turn, leads to high uncertainty of all estimates, a problem compounded by the fact that most applications of EVT focus on the very far right of the distribution, where estimates already have high levels of uncertainty. Therefore, any EVT study must include information about the uncertainty that can be easily interpreted and adapted for the final-user applications. Uncertainty in this study is exclusively obtained using Bayesian methods for inference, which allow a straightforward and intuitive interpretation of their values.

The following sections explore the two approaches used for this study: First, the approach that includes spatial covariates but assumes independence in space (henceforth denoted as the DM approach), and second, the approach that uses a Brown-Resnick max-stable process to account for the spatial dependence (henceforth denoted as the BR approach).

Using a Bayesian distributional model

A simple but effective approach to model the variability of extreme rainfall in space is to pool information from all stations in the study region and assume that all values are independent and identically distributed. This approach assumes that observations at each station are independent of those at any other station. Instead, information is pooled from different stations using spatial covariates, such as the position of each station, as a predictor within a model. The resulting model can then characterize extremal behavior at unobserved locations simply by using their position in the covariates.

In this study, we use an analog of Vector Generalized Linear Models known in the Bayesian literature as Distributional Models (DMs), or sometimes, as Bayesian distributional regression (Umlauf et al., 2018). Distributional models allow for the simultaneous linear modeling of all distributional parameters. This is in contrast to standard GLMs, where only the location parameter is modeled. Furthermore, like VLGMs, DMs allow the use of distributions from outside the exponential family, such as the GEV distribution. Extending a GLM to be a Bayesian DM is straightforward, as one requires only to add the additional log-likelihood contribution from the additional parameter models in the MCMC steps. Using these models, we can incorporate spatial covariates into linear models for every parameter of the marginal distributions.

For every rain gauge j located at s_j, the block maxima $\boldsymbol{i}^l_d(s_j) = (i^l_{d,1}(s_j), ..., i^l_{d,N_j}(s_j))$ are assumed to be i.i.d. and, as the Fisher-Tippett-Gnedenko Theorem for block maxima suggests, follow the Generalized Extreme Value (GEV) distribution, which following (Coles, 2001) is given by

$$G(x) = \begin{cases} \exp\left[-\left(1+\xi\frac{x-\mu}{\sigma}\right)_+^{-1/\xi}\right] & \xi \neq 0, \\ \exp\left[-\frac{x-\mu}{\sigma}\right] & \xi = 0, \end{cases} \tag{5.3}$$

where $\mu \in \mathbb{R}, \sigma > 0, \xi \in \mathbb{R}$ are the location, scale, and shape parameters, respectively, and $x_+ = \max(0, x)$. This assumption is verified for all stations using Quantile-Quantile plots (not shown).

We then follow (Fischer et al., 2017) and describe the spatial variation of location μ and scale σ using a linear combination of Legendre polynomials of longitude and latitude as covariates. Legendre polynomials form a set of orthogonal basis functions on $[-1, 1]$, ensuring that their evaluations at the covariates – normalized to that interval – will be linearly independent. Our model is restricted only to the northing and easting coordinates, ignoring the altitude. Thus, we are left with the distributional model

$$\mu(s) = \beta_0^{\mu} + \sum_{j=1}^{J} \beta_{j,x}^{\mu} P_j(x') + \sum_{k=1}^{K} \beta_{k,y}^{\mu} P_k(y'), \tag{5.4}$$

$$\log(\sigma(s)) = \beta_0^{\sigma} + \sum_{j=1}^{J} \beta_{j,x}^{\sigma} P_j(x') + \sum_{k=1}^{K} \beta_{k,y}^{\sigma} P_k(y'), \tag{5.5}$$

$$\xi = \xi\,, \tag{5.6}$$

where $s = (x', y')$, and a logarithmic link function is used for the scale parameter σ to ensure positivity. $P_i(\cdot)$ denotes the ith order Legendre Polynomial. We transform the coordinates from longitude and latitude to *Universal Transverse Mercator* (UTM) x and y coordinates (UTM zone 33N) so that the distances between stations are measured in meters instead of degrees, simplifying the analysis. The (x, y) coordinates are then shifted and scaled to the (x', y') coordinates within the $[-1, 1] \times [-1, 1]$ square in order to compute the respective Legendre Polynomials.

The shape parameter ξ is left constant throughout the domain, as other studies have found that this parameter is complicated to estimate properly and can strongly impact the model's performance (Cooley et al., 2012).

Model Selection The linear model in Eqs. (5.4) and (5.5) requires an order for the Legendre Polynomials to be specified. The order is chosen within the model selection framework using the Widely Applicable Information Criteria (WAIC) (Vehtari et al., 2017a). A total of 140 possible combinations of up to order $P_5(\cdot)$ were fitted, and the model with the lowest WAIC value was chosen. Furthermore, a regularizing prior (detailed below) was used to lower the risk of overfitting.

Using a max-stable process

For the second step of our study, we expanded the model for the marginal distribution presented in section 5.2.3 by a simple max-stable process. The latter was chosen to capture the extremal dependence in the rainfall maxima. Max-stable processes are extensions to infinite dimensions of finite-dimensional Extreme Value Theory models, arising as "the pointwise maxima taken over an infinite number of (appropriately rescaled) stochastic processes" (Ribatet, 2013).

More precisely, let $X(s)$ be a random variable representing the daily precipitation height at site $s \in \mathcal{S}$ (for some fixed duration d and season l); that is $\{X(s) : \ s \in \mathcal{S}\}$ is a stochastic process modeling the precipitation at each site in the spatial domain $\mathcal{S}$. If we have i.i.d. replicates $\{X_i(s) : \ s \in \mathcal{S}\}$ of the process such as precipitation heights for different days within the same season, then as already discussed, under mild conditions, the Fisher-Tippett-Gnedenko Theorem states that, for each site s and sufficiently large n, the distribution of $\max_{i=1,\dots,n} X_i(s)$ may be approximated by a GEV distribution with spatially varying parameters $\mu(s), \sigma(s), \xi(s)$. Assuming that not only the marginal distributions, but also the spatial dependence structures converge, by a spatial extension of the Fisher-Tippett-Gnedenko theorem the block maxima process $\{\max_{i=1,\dots,n} X_i(s) : \ s \in \mathcal{S}\}$ can be approximated by a max-stable process $\{Z'(s) : \ s \in \mathcal{S}\}$, given that n is large enough (Ribatet et al., 2016). Thus, max-stable processes does not only allow for arbitrary GEV marginal distributions $Z'(s) \sim \mathrm{GEV}(\mu(s), \sigma(s), \xi(s))$, but also provide a flexible way of modeling the dependence structure of the maxima of the X_i random fields.

Consequently, we will assume that the semi-annual block maxima $\boldsymbol{i}_d^l(s_j)$ form realizations of a max-stable process $\{Z'(s):\ s \in \mathcal{S}\}$ at the gauged sites $s_j \in \mathcal{S}$. It is common in extreme value theory to transform the original block maxima data $\boldsymbol{i}_d^l(s_j)$ into standardized maxima $\boldsymbol{z}_d^l(s_j)$ following unit Fréchet marginal distributions (i.e., the case where $\mu = \sigma = \xi = 1$ in Eq. 5.3). Transformation of the margins to the unit Fréchet distribution does not affect the dependence structure. This transformation is performed via the relationship

$$z_{d,t}^l(s) = \left[1 + \xi\left(i_{d,t}^l(s) - \frac{\mu^l(s)}{\sigma^l(s)}\right)\right]_+^{1/\xi}. \tag{5.7}$$

As this transformation can be easily reversed, it then allows us to focus on the max-stable process $\{Z(s):\ s \in \mathcal{S}\}$ without any loss of generality. For such standardized max-stable processes, a variety of parametric submodels has been developed including the popular Brown-Resnick max-stable process model (Kabluchko et al., 2009).

In our study, the marginal standardization requires the specification of response surfaces for $\mu^l(s)$ and $\sigma^l(s)$ to link $\boldsymbol{z}_d^l(s)$ to $\boldsymbol{i}_d^l(s)$. We chose the response surfaces to have the same expressions as the model resulting from the model selection of Eqs. (5.4) and (5.5), with the shape parameter assumed to be constant over the entire domain.

In theory, max-stable process models can be used to model the joint distribution of all semi-annual block maxima $(i_{d,1}^l(s_j), \ldots, i_{d,N_j}^l(s_j))$, $j = 1, \ldots, 53$, and their standardized analogues $(z_{d,1}^l(s_j), \ldots, z_{d,N_j}^l(s_j))$, $j = 1, \ldots, 53$, respectively. In practice, however, the resulting likelihood terms are intractable for even relatively low-dimensional settings. This is why a common strategy is to restrict the process to the bivariate case, where the distribution functions and their corresponding densities are well-known. The bivariate joint probability for the rescaled maxima $Pr\{Z_d^l(s) \leq z_1, Z_d^l(s+\boldsymbol{h}) \leq z_2\}$ is then modeled using the bivariate distribution of the Brown-Resnick max-stable process model (Kabluchko et al. (2009), see appendix 5.A). For the Brown-Resnick model, the extremal spatial dependence is a function only of the variogram γ, which, with a slight abuse of notation, for this study has the following theoretical model:

$$\gamma(s, s+\boldsymbol{h}) = \gamma(\boldsymbol{h}) = \left(\frac{\|\boldsymbol{h}\|}{\rho}\right)^{\alpha}, \tag{5.8}$$

Here, $\|\boldsymbol{h}\|$ is the Euclidean distance between the two locations considered, ρ is the range parameter, and α is the smoothness parameter. The range parameter ρ can be seen as the distance for which the dependence is still effective and takes values ($\rho > 0$). The smooth parameter α has no straightforward interpretation and is constrained to be $\alpha \in [0, 2]$. For this study, we restrict the variogram to be isotropic and stationary, i.e. $\gamma(s, s+\boldsymbol{h})$ depends on $\|\boldsymbol{h}\|$ only. The two Brown-Resnick parameters (α, ρ) contain the information regarding the pairwise dependence structure. To compare the Brown-Resnick dependence estimated from the model with the empirical dependence shown by the data, we also obtain a parametric estimate of the bivariate extremal coefficient $\theta(s_j, s_{j'})$ using the (ρ, α) parameters of the Brown-Resnick model. This parametric estimate $\theta_{\mathrm{BR}}(s_j, s_{j'})$ is computed from the Brown-Resnick variogram $\gamma(h)$ obtained in Eq. (5.8) using the following relationship:

$$\theta_{\mathrm{BR}}(s_j, s_{j'}) = 2\Phi(\gamma(\|s_j - s_{j'}\|)/2)^{1/2}), \tag{5.9}$$

where Φ represents the standard normal distribution function.

Statistical Inference

The estimation of the posterior distribution of the $(\beta_0, \beta_{i,P})$ coefficients for the DM approach and the response surfaces, as well as for the BR dependence parameters (α, ρ), was carried out using Bayesian inference. Given a random variable or vector Y and a probabilistic distribution function $G(\phi)$ such that one assumes that $Y \sim G(\phi)$ (where ϕ represents the distributional parameters), Bayesian inference assumes that the parameters ϕ also follow a probability distribution. The quantity of interest is the so-called *posterior distribution* of probable values for ϕ given observations y from the random variable Y, which is obtained using Bayes' rule: $p(\phi \mid y) \propto p(y \mid \phi)p(\phi)$. The uncertainty of the estimates is then directly obtained from the posterior distribution $p(\phi \mid y)$. Furthermore, the likelihood $p(y \mid \phi)$ is derived from the model, and has the same mathematical expression as the likelihood used for MLE methods. Finally, the so-called prior $p(\phi)$ includes the information known about the parameters ϕ *before* observing the data y. For studies involving extremes, the choice of $p(\phi)$ is of particular importance, as the small size of the data sample typically results in a strong influence of the prior over the posterior. Stephenson (2016) provides current strategies to choose appropriate priors when performing inference of the GEV distribution.

For the inference of the parameters in this study, we used a Markov Chain Monte Carlo (MCMC) sampling scheme. MCMC sampling requires that the right-hand side of Bayes' rule is known up to a multiplicative constant, for which it is enough to know the expression for the likelihood $p(Y \mid \phi)$ and the prior distribution $p(\phi)$.

The likelihood term of the DM approach given by Eqs. (5.4)-(5.6) is directly obtained from the GEV distribution (Eq. (5.3)). For the BR approach, using the full likelihood is unfeasible as the data's high dimensionality made the full likelihood intractable; we chose instead to use the pairwise likelihood from (Padoan et al., 2010) (see Appendix 5.A for details). The expression for the pairwise likelihood of the Brown-Resnick model included both the marginal and dependence parameters so that each MCMC step updated the value of all $\phi = \{\rho, \alpha, \beta_0^\mu, \beta_0^\sigma, \beta_0^\xi, \beta_{i,P}^\psi\}$ parameters simultaneously, where $\beta_{i,P}^\psi$ denotes all potential relevant coefficients for $\psi = \{\mu, \sigma\}$ aside from their intercepts β_0^μ or β_0^σ.

The last step to perform Bayesian inference is to propose a prior distribution for all parameters. For the DM approach, this includes the three intercepts $(\beta_0^\mu, \beta_0^\sigma, \beta_0^\xi)$ and all the possible coefficients $\beta_{i,P}^\psi$, where $\psi \in \{\mu, \sigma\}$.

The covariates were recentered around zero so that the value of the intercepts can be interpreted as the value when all other covariates are set to their mean values. Based on the study of (Fischer et al., 2017), who did a similar analysis for the same region, we use the following priors for the location and scale intercepts:

$$\beta_0^\mu \sim \text{Normal}(1.54, 0.6166) \tag{5.10}$$

$$\beta_0^\sigma \sim \text{Normal}(0.4166, 0.4166) \tag{5.11}$$

The prior for the shape parameter ξ is a rescaled Beta-distribution $\beta_0^\xi \sim \text{Beta}(2,2)$ that has support in $[-0.5, 0.5]$. This choice was made as this prior has already been used by Dyrrdal et al. (2015) and also in the operational application used by MeteoSwiss (Fukutome et al., 2018). For the $\beta_{i,P}^\psi$, we use the prior $\beta_{i,P}^\psi \sim \text{Student} - \text{t}(2, 0, 1)$, which is a regularizing prior (Kruschke, 2014), preventing overfitting.

For the BR approach, the priors for the marginal response surfaces were the same as those used for the DM approach described above. Using the same priors for the two models was done to simplify the comparison between them. Additionally, the prior for the range parameter ρ and the smooth parameter α were elicited from typical values of these parameters in other studies (Zheng et al., 2015; Stephenson et al., 2016) and were chosen to be $p(\rho)$ = Normal$(30000, 7000)$ and $p(\alpha)$ = Exponential(2.5), respectively. The scales of the parameters for $p(\rho)$ are in meters.

MCMC sampling was performed using the software `Stan` (Stan Development Team, 2022). A total of 4 chains with 2500 post-warmup samples per chain using 1000 samples as warmup was used. A visual analysis of the ridge and trace plots was performed for all models to detect issues with MCMC chain convergence.

A known issue when using pairwise likelihoods for Bayesian inference is that the resulting posterior distributions will severely underestimate the spread of the distribution ((Ribatet et al., 2012; Ribatet et al., 2016; Chan et al., 2017)). The underestimation occurs because the pairwise likelihood over-uses the data by including each location in $n/2$ terms of the objective function rather than just one, as would be the case with the full likelihood, resulting in a likelihood function that is far too sharply peaked (Ribatet et al., 2012). While this issue does not severely affect the overall median of the posterior distribution (Chan et al., 2017), the estimated credible intervals of the parameters will be strongly underestimated. To tackle this issue, we applied the *Open Faced-Sandwich* (OFS) correction proposed by (Shaby, 2014) to all posterior MCMC samples from the Brown-Resnick model. The OFS-corrected samples produce credible intervals that have proper coverage values. However, it is worth noting that while the resulting posterior samples fulfill the desired coverage properties, they are no longer truly Bayesian. Appendix 5.C shows a comparison between the raw MCMC samples and the OFS-corrected ones.

Prediction of return levels

Once a posterior distribution of the marginal GEV parameters is obtained from the MCMC samples, it is straightforward to calculate $q_p(s)$ quantile levels for any probability p of non-exceedance (i.e., return levels) via the quantile function of the GEV distribution

$$q_p(s) = \begin{cases} \mu + \frac{\sigma}{\xi}[(-\log p)^{-\xi} - 1] & \xi \neq 0, \\ \mu - \sigma \log(-\log p) & \xi = 0\,. \end{cases} \tag{5.12}$$

For each one of the S MCMC sampled parameter values, we calculate a value of $q_p(s)$ with probability p. This results in a distribution of S return levels. We report the median of these return levels as the estimated return level. Their uncertainty is calculated as the 2.5% and 97.5% quantiles of the S return levels, forming 95% credibility intervals. Note that the resulting return levels no longer stem from a GEV distribution but rather from a mixture of many GEV distributions.

5.2.4 Verification and Model Comparison

We use the quantile score (QS) (Bentzien et al., 2014) as a measure of accuracy for both the marginal and the Brown-Resnick models. Given a series for a single rain gauge of semi-annual block-maxima observations $(i^l_{d,1}(s_j), ..., i^l_{d,N_j}(s_j))$ with N_j years of data for the j-th gauge and the corresponding prediction for the quantile level $q^l_{p,d}(s_j)$ with probability p for the same location s_j, duration d and season l, the QS is defined as:

$$QS^l_{p,d} = \frac{1}{N}\sum_{t=1}^{N} \rho_p(i^l_{d,t}(s_j) - q^l_{p,d}(s_j)); \tag{5.13}$$

$$\text{where} \quad \rho_p(u) = [|u| + (2p-1)u]/2. \tag{5.14}$$

The QS is always positive and reaches an optimal value at zero. We obtain the QS values for both the marginal and the Brown-Resnick model for probability levels of $p = (0.9, 0.95, 0.98, 0.99)$, corresponding to return periods of $(10, 20, 50, 100)$ years.

To compare the performance of two models, Ulrich et al. (2020) defined the Quantile Skill Index (QSI), a measure derived from the Quantile Skill Score QSS (cf. Wilks, 2011, for an introduction to skill scores). Given the QS for a model to be tested ($\mathrm{QS}_{\mathrm{model}}$) and the QS for a reference model ($\mathrm{QS}_{\mathrm{ref}}$), the QSI is defined as

$$\mathrm{QSI} = \begin{cases} 1 - \frac{\mathrm{QS}_{\mathrm{model}}}{\mathrm{QS}_{\mathrm{ref}}}, & \text{if } \mathrm{QS}_{\mathrm{model}} < \mathrm{QS}_{\mathrm{ref}} \\ -\left(1 - \frac{\mathrm{QS}_{\mathrm{ref}}}{\mathrm{QS}_{\mathrm{model}}}\right), & \text{if } \mathrm{QS}_{\mathrm{model}} \geq \mathrm{QS}_{\mathrm{ref}} \end{cases}. \tag{5.15}$$

Positive (negative) values of the QSI indicate a gain (loss) of skill for the tested model over the reference. The advantage of the QSI over the QSS is that the interpretation of negative or positive values is equivalent (which is not the case for skill scores). For this study, the tested model is the Brown-Resnick max-stable process model, and the reference model is the marginal distributional model.

To get an estimation of the out-of-sample performance for the QSI, we applied 10-fold cross-validation in space to estimate the QS values. The folds were constructed such that in each one, 90% of the stations were used for training the model and the remaining 10% for validation. Each station appears in a given validation set once and only once. This specific cross-validation scheme gives an estimate of how good the model is at predicting values at ungauged sites, and it does not give any information on the model's skill at predicting future observations. Considering the sizeable computational load needed to perform MCMC sampling for all 8 models for all 10 folds, we opted to use maximum likelihood instead of Bayesian inference for this step. Using MLE instead of full Bayesian inference was considered a fair assessment as we are only interested in point estimates of return levels when calculating the QS using Eq. (5.14). A separate analysis (not shown) revealed that the QS point estimates obtained from maximum likelihood were almost always very similar to the median QS values obtained from the full posterior distribution.

5.3 Results

5.3.1 Extremal dependence

The estimated bivariate extremal coefficient $\hat{\theta}_{\mathrm{NP}}(s_j, s_{j'})$ for the $\boldsymbol{i}_{12}^{(\mathrm{sum,win})}(s_j)$ and the $\boldsymbol{i}_{24}^{(\mathrm{sum,win})}(s_j)$ block maxima series is shown in Fig. 5.4. The main feature is that winter maxima (blue) consistently show lower average values (i.e., higher extremal dependence) until a distance of around $h = 150$ km. For distances $h > 150$km this relationship is inverted. Furthermore, the average distance where the pairwise maxima still show asymptotical dependence is shown to be lower than 150 km. The difference between seasons is larger for the 12-hour series, possibly reflecting differences in the rainfall generating mechanisms at this timescale compared to the 24-hour series.

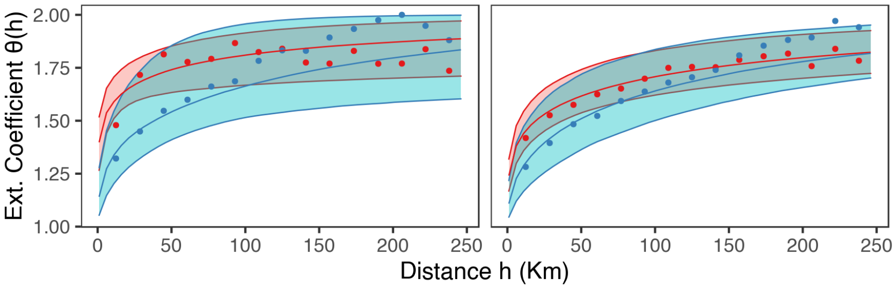

FIGURE 5.4: Empirical values of the extremal coefficient $\hat{\theta}_{\text{NP}}(s_j, s_{j'})$ (dots) and estimated values from the resulting Brown-Resnick variogram $(\theta_{\text{BR}}(s_j, s_{j'})$ (solid lines, shaded regions represent the 50% CI). Colors represent the season: blue for winter and red for summer. The left panel shows results from the 12-hourly data; the right panel shows results for the daily data. $\theta(s_j, s_{j'}) \in [1, 2]$, where one is complete dependence and two is complete independence.

Estimates of the extremal coefficient based on the Brown-Resnick model $(\hat{\theta}_{\text{BR}}(s_j, s_{j'}))$ are also shown in Fig. 5.4 as the solid lines with the shaded regions representing 50% credibility intervals. We compare the values from the Brown-Resnick model to the empirical $\hat{\theta}_{\text{NP}}(s_j, s_{j'})$ to get an idea of how well the BR approach captures the pairwise dependence shown by the data. For the 12-hour series (left), this comparison shows that for winter, the model consistently overestimates the strength of the dependence for $h > 100$, while for summer, the average dependence is properly captured for $h \lesssim 150$ km; for greater distances, the dependence is underestimates. The overestimation in the winter model can also be seen for the daily series (right); however, it is much less pronounced, with most of the average $\hat{\theta}_{\text{NP}}(s_j, s_{j'})$ falling inside of the 50% CIs. In both time series, the 50% CIs are larger for winter; this could suggest that the extremal dependence for winter is more complex than for summer, resulting in the winter model exploring a greater range of values for the dependence parameters. Additionally, the daily series shows less variability than the 12-hour series. This difference in variability may be due to the increased length of observations for the daily series.

An initial inspection of the $\hat{\theta}_{\text{NP}}(s_j, s_{j'})$ values would suggest that the data shows asymptotic dependence for all series for distances up to $h \leq 150$ km. Therefore, the assumption of asymptotic dependence necessary for using a max-stable process, should be justified. Further discussion about this topic can be found in Appendix 5.D.

5.3.2 Model building

Model selection

The procedure to choose the orders for the Legendre Polynomials of Eqs. (5.4)-(5.6) results in the models described in Tab. 5.1. Basic prior and posterior predictive checks were performed to detect any misspecification issues; some examples for the reference stations described below can be found in Appendix 5.E. A visual analysis of the prior and posterior checks did not detect any severe issues.

5.3.3 Parameter estimates

Parameter estimates for the dependence and shape parameters are reported in Tab. 5.2 as median and 95% credibility intervals for each parameter. A distinct difference can be seen in the value of the range parameter ρ (in meters) between summer and winter,

TABLE 5.1: Maximum chosen orders of Legendre Polynomials for the distributional model in Eqs. (5.4)-(5.5)

	μ	σ
summer (24h)	2	1
winter (24h)	3	2
summer (12h)	2	1
winter (12h)	3	2

.

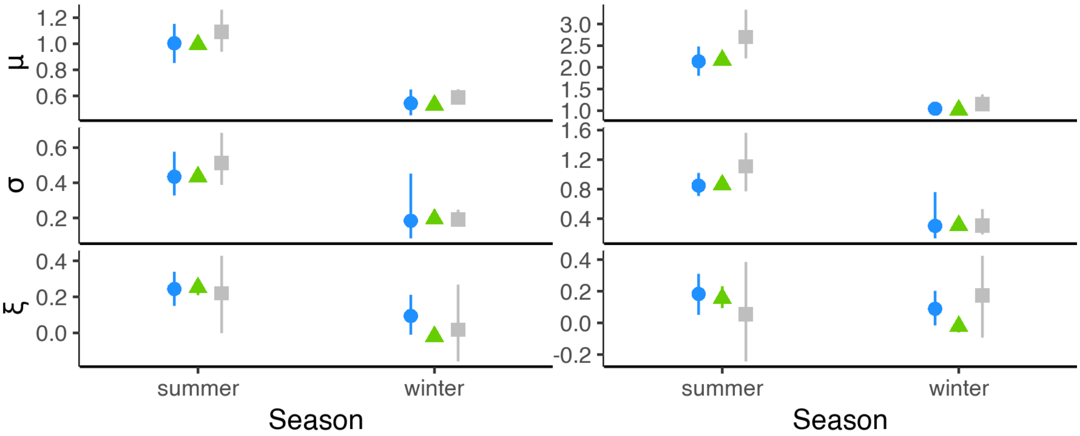

FIGURE 5.5: Estimated values of the location μ, scale σ, and shape ξ parameters of the GEV distribution for station Potsdam. The symbols' shape indicates the model used for estimation: circle (blue) = BR, triangle (green) = DM, square (gray) = pointwise GEV.

as the value in winter is always significantly larger than for summer, regardless of the time scale. This result is consistent with the behavior of the extremal coefficient seen in Fig. 5.4, and it may indicate that the rainfall events leading to the block maxima in winter are, on average, larger than those in summer. Furthermore, the shape parameter shows a difference for winter and summer, regardless of the time scale.

TABLE 5.2: Bayesian estimates of the Brown-Resnick max-stable model parameters. Posterior medians are reported along with their 95% credible interval limits on either side as (lower,median,upper). The coefficients corresponding to the Legendre Polynomials were omitted from this table.

	ρ	α	ξ
12h (s)	413,4896,11997	0.17,0.40,0.64	0.05,0.18,0.31
12h (w)	3596,43870,104192	0.31,0.78,1.24	-0.01,0.08,0.20
24h (s)	4722,22993,43936	0.39,0.54,0.69	0.15,0.24,0.33
24h (w)	6801,53228,109265	0.57,0.83,1.12	-0.01,0.09,0.21

5.3.4 Marginal parameters and return levels

Four reference stations were chosen to illustrate the differences in the marginal GEV parameters and return levels from the DM and the BR models (respective locations of the reference stations are given by red diamonds in Fig. 5.2). We chose the two stations with the longest time series, which are closely surrounded by other stations (Potsdam and Lindenberg), a station with a long time series that is isolated from other stations (Meyenburg), and a station with a short time series which is surrounded by

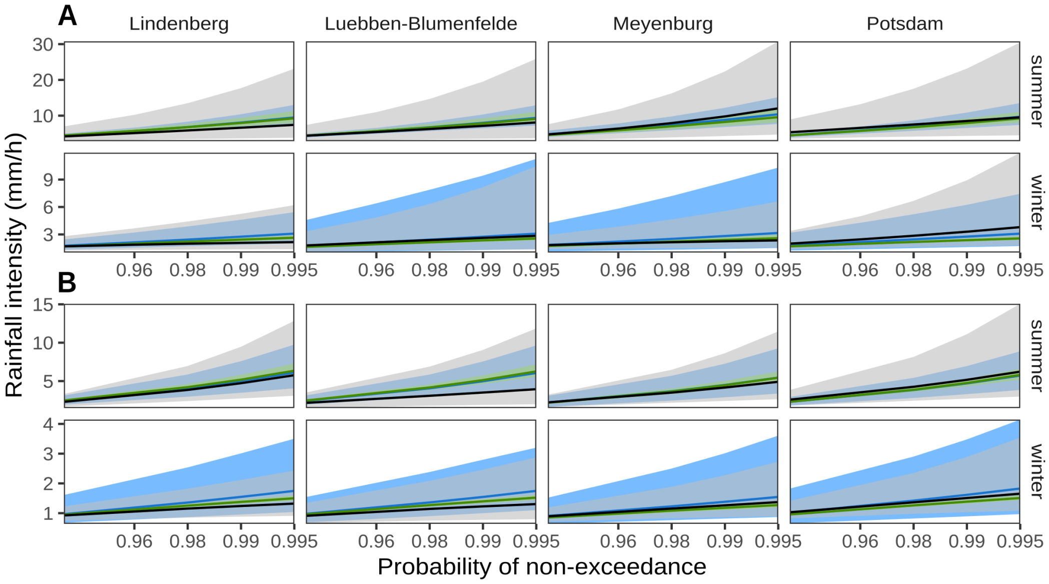

FIGURE 5.6: Return level of precipitation intensity (mm/h). Color denotes the model used: Blue for the BR model, green for the DM model, and gray for the pointwise GEV. Shaded regions represent pointwise 95% credibility intervals. (A) 12-hour data, (B) daily data. For reference, the probabilities of non-exceedance $p = (0.96, 0.98, 0.99, 0.995)$ correspond to the $(25, 50, 100, 200)$ year return periods, respectively.

other stations (Luebben-Blumensfelde). Figures 5.5-5.6 show the GEV parameter estimates and the resulting return levels with 95% credibility intervals, respectively. Furthermore, pointwise GEV estimates and their resulting return levels with 95% credibility intervals were added for reference; these estimates were obtained using the same priors for the intercepts described in section 5.2.3.

Concerning the GEV parameters, Fig. 5.5 shows that the pointwise estimates (taken here as the median value of the posterior distributions) are similar for the DM and the BR models. This similarity was expected, as the marginal parameters are only vaguely affected by the spatial dependence through their incorporation in the likelihood term of Eq. (5.16). However, when comparing the models, a pattern concerning the uncertainty of the estimated parameters (taken here to be the 95% credibility intervals) is visible. For summer (and in the case of ξ, also winter), the highest uncertainty is always seen for the pointwise GEV model, followed by the BR model and the DM model, which consistently show the smallest uncertainty. In contrast, the largest uncertainty for location and scale in winter can be seen for the BR model, followed by the pointwise GEV and the distributional models. This phenomenon can be observed in other stations (not shown). We infer that the uncertainty estimated for the marginal parameters is strongly affected by the underlying spatial dependence, which changes according to the rainfall-generating mechanisms dominant in the respective season.

To further delve into the last point, Fig. 5.6 presents how the return levels for different non-exceedance probabilities for the BR and DM approaches differ. As before, we compare different seasons and two different durations. The median return level is generally similar across the different models, with increasing differences for larger probabilities of non-exceedance. In contrast, the uncertainty is noticeably different for each model, which is consistent with the results of the GEV parameters. In summer, the uncertainty is always largest for the pointwise GEV model, followed in order by the BR and the DM models. This changes in winter, when the uncertainty is largest

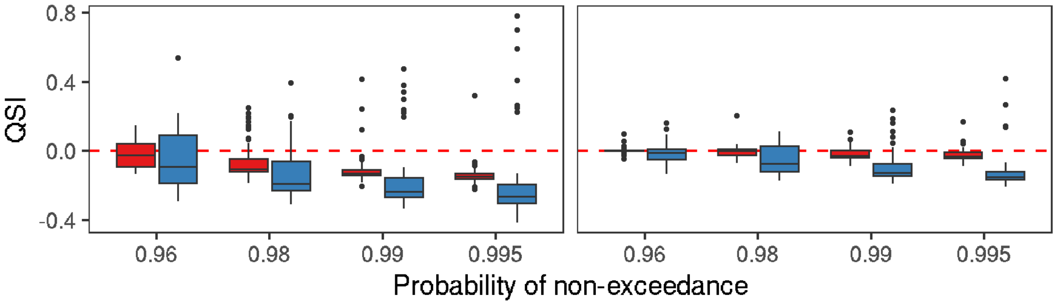

FIGURE 5.7: Boxplots showing the distribution of the Quantile Skill Index for all stations for 12-hourly (left) and daily (right) data. The colors indicate the season. Positive (negative) values indicate an increase (decrease) in skill for the BR approach compared to the DM one.

typically for the BR model, with a few exceptions. Surprisingly, it would appear that the inclusion of the max-stable dependence on the model for winter resulted in an overall increase in the uncertainty, even when compared to the pointwise model that contains no information about other stations. This result may be associated with a loss of skill for the BR model when modeling block maxima in winter, an aspect that will be explored in the next section.

5.3.5 Model comparison

We now explore how the seasonal differences in the extremal dependence affect the accuracy of the return levels estimates using the BR model. We use the DM approach as reference in the QSI to assess how much the dependence influences the return level estimates. Positive (negative) QSI values mean that the predicted return levels for ungauged sites have better (worse) QS values for the dependent BR model than for the independent DM one. For this study, our main focus is on the QSI difference between seasons, as we believe this arises from a change in the extremal dependence when analyzing the semi-annual block maxima from different meteorological regimes.

Figure 5.7 depicts the distribution of the cross-validated QSI values over all stations. The 12-hourly data shows an overall average loss of skill for the winter and summer models when using the BR model. This loss of skill increases as the non-exceedance probability increases, with the winter model showing substantially lower average QSI values than the summer model. Furthermore, the variability in QSI values is noticeably larger for winter than summer; in fact, the highest QSI value is always found within an outlier of the winter model. For the daily data, the winter model shows the same decrease in skill with increasing non-exceedance probabilities with high variability; however, in this case, the summer model consistently shows QSI values close to zero with very low variability. Finally, a noteworthy difference between average QSI values can be observed between summer and winter for both periods. This difference increases with the non-exceedance probability, but it remains constant between the 12-hourly and daily periods.

To further explore the difference in QSI values for both seasons and durations, Figure 5.8 shows the spatial distribution of QSI values, i.e. values for every station. No apparent pattern is visible from the different configurations, as QSI values appear to be largely random. Nevertheless, the lowest QSI appear mostly at stations close to the domain's border, suggesting that the BR model performs better when a station is surrounded on all sides by other stations. This effect is admittedly not very reliable, as stations with very low values of QSI can also be found within the middle of the

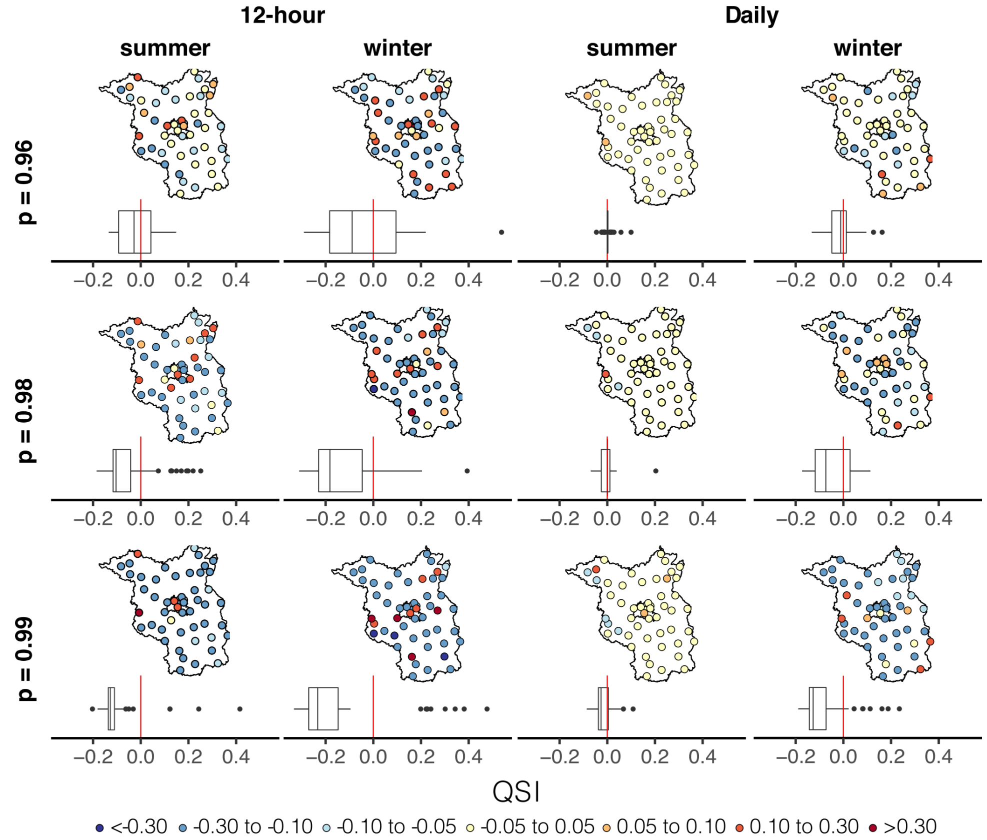
12-hour
Daily
summer
winter
summer
winter
p = 0.96
p = 0.98
p = 0.99
−0.2 0.0 0.2 0.4
QSI
● <-0.30 ● -0.30 to -0.10 ● -0.10 to -0.05 ● -0.05 to 0.05 ● 0.05 to 0.10 ● 0.10 to 0.30 ● >0.30

FIGURE 5.8: Spatial distribution within Berlin-Brandenburg (solid line) of QSI values for the 12-hourly (leftmost two columns) and daily (rightmost two columns) data. Boxplots below each map show the distribution of QSI values as in Fig. 5.7

domain. A closer inspection of the difference between seasons reveals a subtle pattern: similar QSI values seem to cluster in summer, while the distribution is predominantly random in winter. This change could be attributed to the difference in the rainfall generating processes, as will be discussed in the following.

5.4 Discussion

The results described in the last section provide compelling evidence that the extremal dependence shown by the data changes sufficiently enough to have a noticeable effect in the resulting marginal estimates down the line when using a model capable of capturing such dependence (the BR model in our case). This difference was mainly observed when comparing different marginal quantities from two seasons: the estimated GEV parameters (with their respective return levels) and the cross-validated Quantile Score (QS), an out-of-sample performance measure for the predicted return levels for ungauged sites.

The observed difference in marginal estimates when using a spatial model is consistent with previous extreme rainfall studies; Stephenson et al. (2016), for example, reports that the incorporation of the max-stable process dependence led to an overall shift towards heavier tailed marginal distributions across their study location. They also found that the uncertainty for the estimated marginal quantities was larger for the max-stable process model than for the independent model, which is in good agreement with our results for summer maxima estimated return levels. Additionally, the spatial distribution of the QSI falls in line with Le et al. (2018), who found that return levels estimated from a max-stable process presented noticeable differences in their spatial distributions compared to an unconditional model. However, these studies did not estimate the impact of this difference on the model performance. Our study then provides insight into the operational use of Brown-Resnick max-stable models by first examining how different types of rainfall-generating mechanisms affect the marginal estimates and then applying a model validation framework for the out-of-sample ungauged model accuracy.

A comparison of the uncertainty in the GEV parameters and the return levels showed that when modeling summer maxima, the uncertainty resulting from the BR model appeared to be a middle point between the DM and the pointwise GEV models. However, the significant reduction in uncertainty for the DM model compared to the pointwise GEV model signals that this model underestimates the natural variability in the rainfall data. Thus, it seems plausible that the larger uncertainty seen in the BR model is a more accurate representation of such variability. In contrast, when dealing with winter maxima, the uncertainty obtained by the BR model appears to have been consistently overestimated compared to the pointwise approach. From our results, it is not completely clear why this is the case, but it may be speculated that the isotropic Brown-Resnick dependence model was misspecified for the extremal dependence structure in the winter data. A possible source of error is the assumption that the dependence structure is isotropic, which might be a better approximation for convective events than for synoptic/mixed events that occur in winter. On the other hand, the larger values estimated for the range and smooth parameters indicate that the dependence is stronger for winter than in summer; however, these parameters do not say anything about the isotropic/anisotropic structure. This larger dependence in winter could be attributed to frontal events being generally larger and more elongated than convective events. Thus, more stations are simultaneously affected by the same event, increasing the dependence. Figure 5.9 in appendix 5.B reports how many unique events resulted in block maxima being chosen from the daily series in winter and summer. This table supports the idea that the events are larger in winter, as the number of unique events is consistently lower in winter than in summer. However, Fig. 5.4 also reveals a surprising increase in dependence for distances larger than 120 km; this may suggest that some underlying weather patterns from a larger scale than the convective scale influence the dependence.

Our findings report that the Brown-Resnick model is mostly as good as the unconditional DM model when modeling summer block maxima, whereas the BR model presents a remarkable loss in skill compared to the DM model when modeling winter block maxima. It is worth noting that past studies have primarily focused on summer maxima, as the convective nature of the rainfall-generating mechanisms in this season typically leads to the annual maxima events to occur in summer. Our findings suggest that the isotropic Brown-Resnick dependence model is a proper first approximation when dealing with block maxima resulting from convective events. On the other hand, the loss in skill for the winter maxima model provides further evidence that this model is misspecified when dealing with either synoptic, stratiform, or a mixture of synoptic/convective events.

We acknowledge potential limitations to this study. An important question for future studies is to determine the effect of anisotropy in the results, which, as discussed above, is expected to have an important role in modeling the spatial dependence for synoptic events. Furthermore, previous studies have shown that rain gauge networks are typically too scarce to resolve convective cells properly (Lengfeld et al., 2019). Thus, in order to get a better representation of the spatial dependence, future work should make use of radar networks to complement rain gauge data. A significant limitation of our work was the use of the pairwise likelihood instead of the full likelihood of the Brown-Resnick model within the Bayesian framework. While some of the most known issues with this approach were tackled by using the Open-Faced Sandwich approach of (Shaby, 2014), it would be beneficial instead to use a full-likelihood approach such as that of (Dombry et al., 2017). Furthermore, due to the high computational demand of performing Cross-Validation within a Bayesian setting, the QS and QSI results reported in the results come from a maximum likelihood estimation. Moreover, we assumed that the data was stationary, ignoring the possible effects of climate change. The effect of this non-stationarity on the extremal dependence should be explored in further studies, as it has been shown that accounting for non-stationarity results in a measurable effect on the return level estimates (Ganguli et al., 2017). Our study indirectly classified precipitation types based on dominant types for different seasons. Further studies should use a direct classification of event types, which would avoid the mixing of convective and frontal events in winter. Some work in classifying extreme events already exists, for example, that of Lengfeld et al. (2021). Furthermore, the use of max-stable processes requires that the data present asymptotic tail dependence, an assumption that does not hold for aggregation durations lower than 12 hours. For a more in-depth study of convective events shorter durations would be needed; in this case, a more flexible model that can capture both asymptotic tail dependence and independence would be needed, such as the one proposed by Wadsworth et al. (2019), which was applied to hourly rainfall data by Richards et al. (2021).

This study indicates that different rainfall mechanisms can strongly influence the spatial dependence presented by the block maxima. This change in the dependence structure can, in turn, result in significant misspecification of the model if not accounted for properly. Thus, it is essential to understand the types of rainfall-generating mechanisms in the domain of study when using max-stable models.

Appendices

5.A Inference from the Brown-Resnick max-stable process

Inference is done using the pairwise log-likelihood (Padoan et al., 2010), which for our study is

$$L(\phi \mid i^l_{d,1}(s_1), ..., i^l_{d,N}(s_J)) = \sum_{t=1}^{N} \sum_{j=1}^{J-1} \sum_{j'=j+1}^{J} \log f(i^l_{d,t}(s_j), i^l_{d,t}(s_{j'}) \mid \phi), \tag{5.16}$$

where $\phi = \{\rho, \alpha, \beta_0^{\mu}, \beta_0^{\sigma}, \beta_0^{\xi}, \beta_{i,P}^{\psi}\}$ represents the parameters to estimate, $i^l_{d,t}(s_j)$ is the observed semi-annual block maxima for the duration d and season l at location s_j for year t, and each term $f(\cdot,\cdot)$ is the appropriately transformed bivariate density function derived from the bivariate distribution function for the Brown-Resnick process given by

$$\Pr[Z(s_1) \leq z_1, Z(s_2) \leq z_2] =$$
$$\exp\left[-\frac{1}{z_1}\Phi\left(\frac{\sqrt{\gamma(h)}}{2} + \frac{1}{\sqrt{\gamma(h)}}\log\frac{z_2}{z_1}\right) - \frac{1}{z_2}\Phi\left(\frac{\sqrt{\gamma(h)}}{2} + \frac{1}{\sqrt{\gamma(h)}}\log\frac{z_1}{z_2}\right)\right]. \tag{5.17}$$

Here z follows a unit Fréchet distribution, Φ denotes the standard normal distribution function, h is defined as the euclidean distance between s_1 and s_2, and the variogram γ is defined in Eq. 5.8. In equation (5.16) it is assumed that the number of years N is equal for all J-stations. However, this is not the case, as some stations have longer records than others. We took N to be the one from the station with the longest records, and whenever a station did not have data for the t-th year, we made the corresponding term in the log-likelihood to be zero. However, the time period used for all stations was chosen to minimize the number of paired stations with no data.

5.B Number of events per year

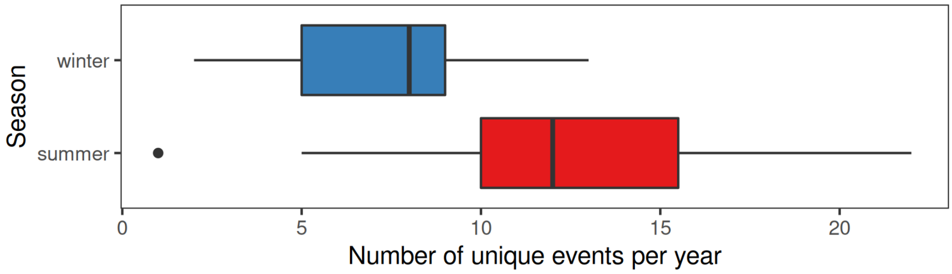

FIGURE 5.9: Boxplots showing the difference in unique events for the different seasons studied. Only the 24 hour data is shown.

Figure 5.9 shows how many unique events that resulted in block maxima were seen from the daily $i^{(\text{sum,win})}_{24,l}$ series. Overall, the number of unique events in summer is larger than in winter.

5.C OFS correction for Bayesian Inference using composite likelihood

A comparison between the uncorrected raw samples from the MCMC sampling using Stan with the pairwise likelihood of Eq. (5.16) and the corresponding samples corrected using the Open-Faced Sandwich (OFS) correction from Shaby (2014) is shown in Fig. 5.10. It can be seen that the uncorrected samples grossly underestimate the uncertainty shown by the 95% credibility intervals. On the other hand, the OFS corrected samples keep the same median but "stretch" the resulting uncertainty so that the desired 95% coverage of the intervals is achieved.

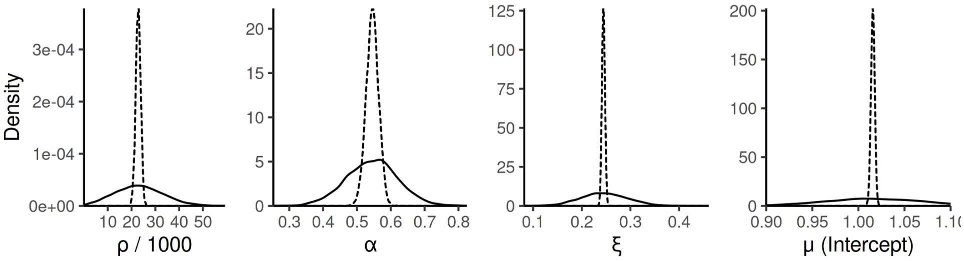

FIGURE 5.10: Density plots for the raw MCMC samples (dashed line) and the resulting OFS corrected samples (continuous lines) for 4 selected parameters from the daily summer results.

5.D Analysis of asymptotic dependence using extremal coefficient

Figure 5.11 shows the distribution of bootstraped samples for $\hat{\theta}_{\mathrm{NP}}(s_j, s_{j'})$, where the estimation method is the same as the one used for Fig. 5.4. The bootstraped samples provide an estimate of the uncertainty that allows us to judge the asymptotic dependence conditions present in the data.

The figure shows that for the daily series, both seasons show a value of the extremal coefficient below 1.75 for $h \leq 150$ km, suggesting that the data is asymptotically dependent at least for this distance. After 150 km, the coefficient goes close to 2, but not immediately. On the other hand, the situation is different between summer and winter for the hourly frequency. Here, the uncertainty is much larger, which could be a reflection of the smaller number of years. Furthermore, while the winter series behaves similar to the daily winter series having reasonably strong dependence for distances up to 150 km, the hourly summer series tends very quickly to lower dependence levels. This again suggests that the events in summer are typically smaller in size that those in winter. Asymptotical dependence can be reasonably suggested for the hourly winter data, but for hourly summer, one could argue this is true only for relatively short distances. However, the uncertainty is rather large, wit a lot of values still falling under the strong dependence case. Therefore, we make the assumption for asymptotical dependence for all four series.

5.E Model diagnostic and posterior predictive checks for reference stations

We obtained Quantile-Quantile plots and Posterior predictive checks to assure that our model adequately represents the observed data. Some of these are shown in Fig.

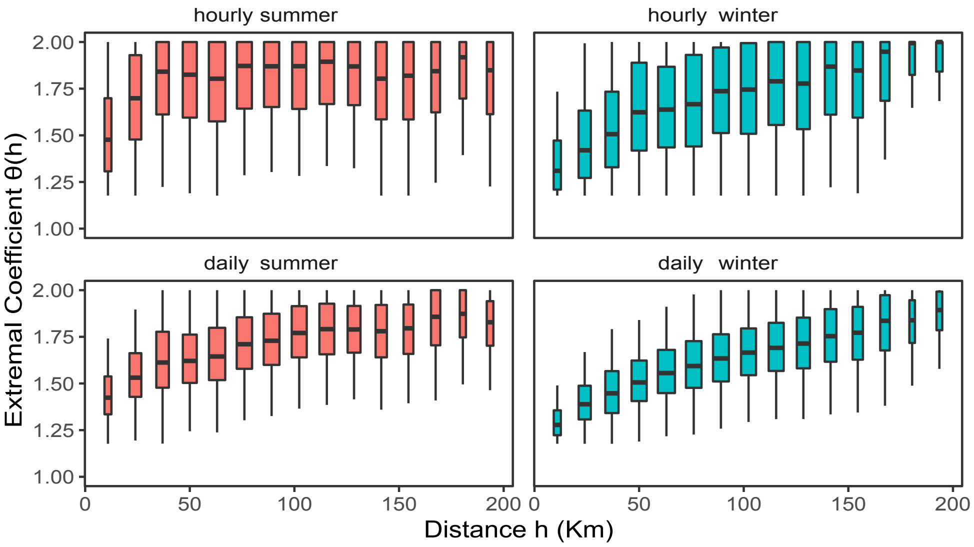

FIGURE 5.11: Boxplots showing the distribution of bootstraped samples (N=500) of the non-parametrical estimate of the extremal coefficient $\hat{\theta}_{\mathrm{NP}}(s_j, s_{j'})$. The width of the boxplots is proportional to the amount of data.

5.12. The included plots come from the hourly summer data.

The QQ plots for the different stations provide evidence that the GEV is mostly appropriate for modeling the marginal distributions, with some minor exceptions. Similarly, the posterior predictive checks allow us to see how well the posterior distributions for the GLM and BR models would be at capturing the original data. In this case, both models' original data seems well-captured.

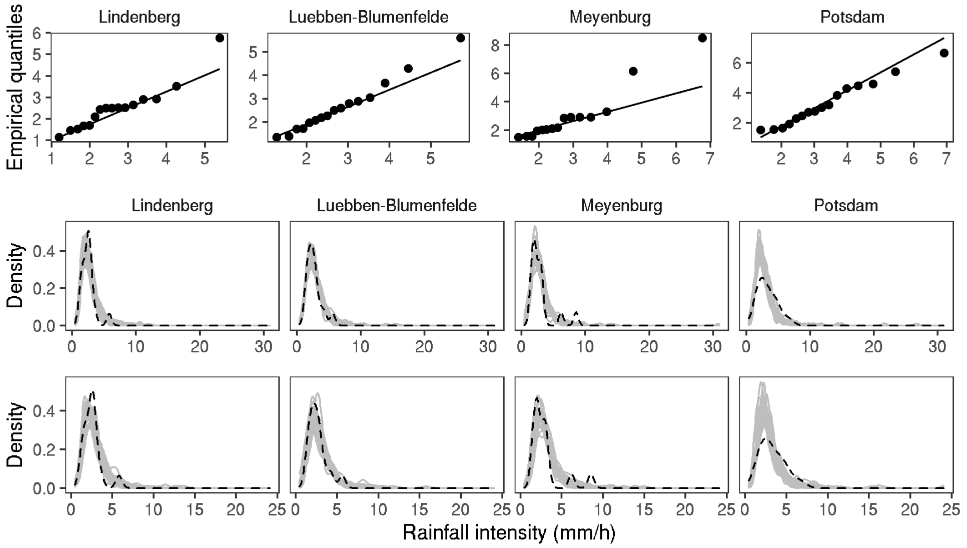

FIGURE 5.12: Top row: Q-Q plots for the four reference stations. Middle row: Posterior predictive check for the GLM model. Bottom row: Posterior predictive check for the BR model. For the middle and bottom row, the dashed line shows the observed density, and the grey lines show 100 samples from 20 GEV distributions with parameters sampled from the respective posterior distributions.

6 Other advances to the modeling of extreme rainfall

In the previous two chapters, two studies were performed to show how the incorporation of different dependence structures can improve statistical models for extreme rainfall. These applications involved a direct incorporation of the dependence, achieved by using a max-stable process. However, these methods are not the only ones that can borrow the strength from spatial/temporal similarity. As mentioned in ch. 3, a VGLM model that assumes independence but incorporates spatial covariates can also be used to model the spatial variability of extreme rainfall. The question is then how good the model is when it (wrongly) assumes that there exists no dependence between the random variables.

In the following sections, we summarize the results of several studies where I used the previously developed techniques from ch. 4 to capture the dependence not in space but across durations for IDF relations. Although the results from this study showed that modeling the dependence does not significantly improve the pointwise estimate of IDF relations, it allows for simulation of maxima series for different durations when including the dependence across duration. These simulations constitute a prerequisite for evaluating strategies for confidence intervals. In section 6.1, we explore this application for a case study in the Wupper region.

Additionally, the relatively short-ranged dependence between durations found for the Wupper stations in ch. 4 is used as a justification for the study presented in section 6.2. This allowed for a more nuanced discussion of the resulting uncertainties when using the independence likelihood.

Finally, the contribution to a study that uses the copula-based modeling approach described in section 3.6.5 to describe the bivariate spatial dependence is given in section 6.3. In this case, the two random variables are ozone and temperature, nevertheless, the methods for modeling in space remain the same.

6.1 Using simulations from a max-stable process to investigate coverage of IDF curves

Publication details

The following section contains a summary of the contribution made to the published study titled **Flexible and consistent quantile estimation for intensity-duration-frequency curves**, published in 2021 by Felix S. Fauer, Jana Ulrich, Oscar E. Jurado, and Henning W. Rust to the HESS journal. The other authors granted permission to reproduce some excerpts of text and figures.

Chapter 3 introduced the concept of IDF curves and the d-GEV distribution. These methods aim to incorporate extreme rainfall events of different durations into a single model, from which inter- and extrapolation to unobserved durations can be performed. An additional advantage is that information from durations with larger datasets can be transferred to durations with less information; for example, the DWD in Germany has much more records for daily precipitation height than for hourly or minutely precipitation height. Choosing an appropriate IDF model can then take advantage of the extra information for the daily precipitation intensity, transferring it to the shorter durations. This occurs by assuming that the function that links intensity and duration is rather smooth, allowing for interpolation.

In past chapters we showed that the d-GEV distribution from eq. (3.12) can handle precipitation intensities of any arbitrary duration. However, in this study, Fauer et al. (2021) showed that the d-GEV is not flexible enough to capture some variations across durations, as seen in annual maxima for various stations. In this study, the main goal is to investigate the effect of adding two parameters to the d-GEV model that allow it to be more flexible: the so-called **intensity offset** and the **multiscaling parameter**. The impact of these parameters was investigated using the QSI verification measure.

6.1.1 The issue of underestimating uncertainty

This study uses rainfall series aggregated from different durations to fit different d-GEV models. The model parameters were fit using MLE, which requires the assumption of independence between rainfall data of different durations. However, as explored by Jurado et al. (2020) (ch. 4), the method for constructing the aggregated intensity series for different durations always induces a statistical dependence, which is known as the dependence between durations. Thus, by assuming independence between durations when performing MLE, the resulting likelihood is misspecified. Similar studies have shown that, under certain conditions, pointwise estimates from misspecified likelihood models are unbiased, but the estimated uncertainty is not always meaningful. Chan et al. (2017), for example, found that a misspecified GEV model that ignored spatial dependence of extreme rainfall resulted in consistently underestimated confidence intervals for the resulting return levels.

Underestimation of the uncertainty can result in a severe issue when generating confidence intervals: if the resulting intervals are too narrow, then the true value will be missed by the interval in a greater ratio than the expected one. Figure 6.1 shows an example of this phenomenon, where the same estimation method was used to construct two sets of 100 confidence intervals. The confidence intervals were constructed in a way that, if we took many samples and constructed CIs for each one, we expect 95%

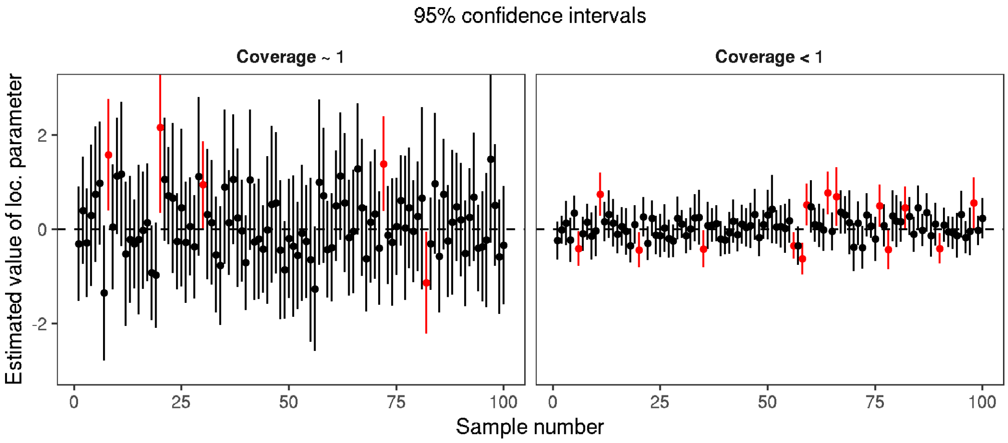

FIGURE 6.1: Example of confidence intervals with two different coverage probability values. For this example, the true value of the estimated parameter is zero. Red intervals denote confidence intervals that did not contain the true value inside. For both examples the nominal coverage is set at .95, which means that we expect 95 out of 100 CIs to contain the true value. The left plot shows 95% CIs with the proper nominal coverage. The right plot shows CIs that underestimate the uncertainty, resulting in more intervals not containing the true value than expected.

of the intervals to contain the true value (in statistical jargon: the nominal coverage was set to 0.95). In the left panel we see that 5 out of 100 intervals did not contain the true value, which is expected. In the right one, the estimates were modified so that the uncertainty was underestimated, resulting in much narrower intervals. From this panel we can see that around 14 out of 100 intervals did not contain the true value, which represents a larger proportion than the expected 5. This example illustrates the main problem with underestimating the uncertainty: the intervals are too narrow, increasing the number of intervals that do not contain the true value. The quantity that summarizes the proportion of time that a confidence interval contains the true value is known as the coverage probability, explained in section 2.4.1.

In any case, obtaining the proper uncertainty values from the d-GEV model is crucial for the correct interpretation of the resulting return levels, which are used by e.g. water management agencies. The question for this study was then: What impact does the dependence between durations have in the MLE-based uncertainty estimates for the flexible d-GEV model? To investigate this question, a coverage study of the confidence intervals was required to make sure that the coverage probability did not deviate too strongly from the nominal coverage.

As far as we know, very little studies investigating the effect of ignoring the dependence between aggregation durations existed before doing the work for this dissertation, as most studies focused exclusively in quantifying the effects of spatial dependence. This knowledge gap was the focus of study 1 (ch. 4), where the impact of including the dependence between durations is investigated.

6.1.2 Using simulated data to calculate coverage probabilities

In that study, we do not delve explicitly in the quantification on uncertainty, but we did propose a method to simulate rainfall series with a specified level of dependence. For the study of Fauer et al. (2021), we had the idea to adapt this simulation in order to perform a coverage analysis of bootstrapped CIs via a simulation study, where the simulated data had different (known) dependence levels. This proceedure is as follows:

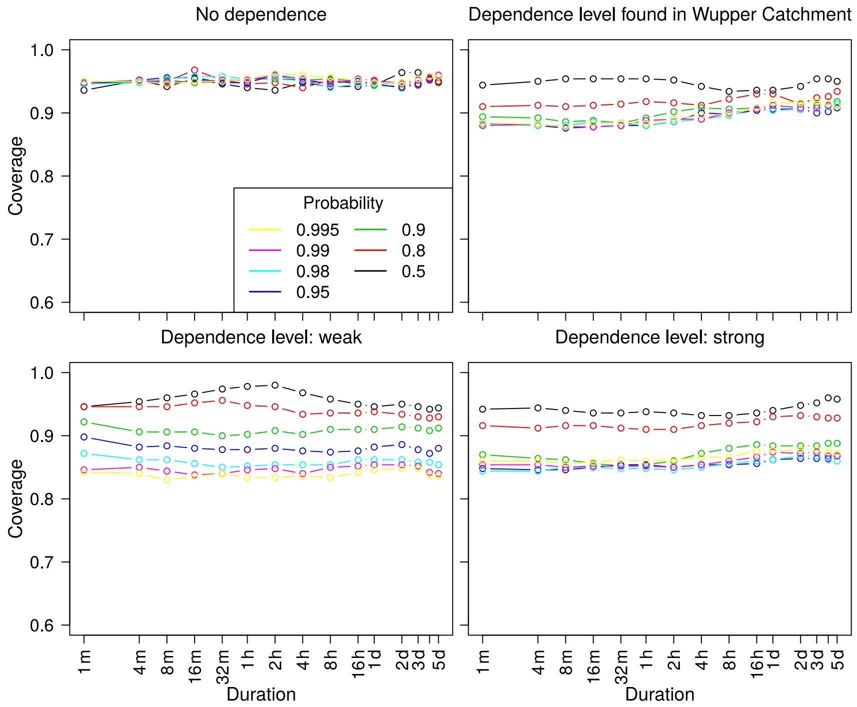

FIGURE 6.2: Coverage probability for different levels of dependence between durations. The nominal coverage was of 0.95 for all panels. The different colors represent the probabilities of non-exceedance for which the return level was computed. Image reproduced from Fauer et al. (2021) with permission of the authors.

1. For the d-GEV model, fix the value of all parameters to known values that are reasonable for the study area.

2. Set the values of the dependence parameters (ρ and σ in eq. (4.8)) used for the simulation. This step allows us to choose between a high or low level of dependence.

3. Generate a dataset of rainfall maxima with different aggregation durations containing the level of dependence between durations set in the last step. Note that here we directly simulate annual maxima.

4. Estimate parameters for the proposed flexible d-GEV model using the simulated data from the last step. Additionally, obtain the 95% confidence interval of the parameters and propagate to the return levels.

5. Repeat the last two steps many times, keeping track of how many times the resulting confidence intervals contain the true value of the parameters set in the first step. Using this information, obtain the coverage probability cov_p, given by dividing the relative number of times that the interval contained the true value by the nominal coverage.

The above procedure allowed us to study how the level of dependence can affect the resulting coverage of the confidence intervals. Our previous hypothesis was that for strong dependence, the coverage probability would differ strongly from the nominal values. The resulting coverage values for the different dependence levels are shown in Fig. 6.2. For both the weak and strong dependence, the coverage for the 95% CI was around 0.9, which was found to be an appropriate trade off for using the d-GEV with the independence assumption. This was a rather surprising finding, but it aligns with the findings from study 1, where the different dependence levels did not seem to have a big impact in the final estimates.

In order to create the simulated rainfall data from different duration with different dependence levels, a max-stable process was constructed using the same procedure as in the study from ch. 4. For this, a 1-dimensional "duration-log-space" was constructed, where the log of each duration $\log_2(d)$ acted as a location, and the distance between durations d and d' was given by the log-ratio $\log_2(d) - \log_2(d') = \log_2(d/d')$. This construction allowed the use of a simple max-stable Brown-Resnick process, for which the dependence could be set using the range and smoothness parameters. Max-stable processes were then used to capture the dependence to learn more about the modeling of this particular dataset.

Performing the simulation study for determining the coverage was crucial for the correct interpretation of the results from this study. This was due to the likelihood misspecification resulting from the assumption that the maxima from different aggregation durations were independent, even though we know this dependence exists. Without the simulation study, we would not know if the misspecification would result in strong deviations of the coverage from the confidence levels below their nominal levels. The specific contribution to this study was the following: design of the simulation study, writing of the R code for the simulations, and co-writing/editing of the paragraph describing the simulation study.

6.2 Modeling extreme rainfall of different durations with a VGLM

Publication details

The following section contains a summary of the contribution made to the published study titled **Estimating IDF curves consistently over durations with spatial covariates**, published in 2020 by Jana Ulrich, Oscar E. Jurado, Madlen Peter, Marc Scheibel, and Henning W. Rust to the Water Journal. The other authors granted permission to reproduce some excerpts of the results.

So far, in the examples and studies presented in this thesis (particularly in ch. 5), the dependence between different spatial locations has been one of the main topics of interest when creating models. We have seen that including this dependence explicitly in the models can lead to reductions in uncertainty when pooling data from different locations. However, this effect can be relatively small in some cases, and when the interest is only in the parameters of the marginal GEV distribution for different locations, the added complexity of a max-stable model could prove to be too much to justify their use. In these cases, a valid alternative is using a VGLM (or its Bayesian alternative, the Distributional Model), as explained in section 2.2.

The dataset used for this study consists of minutely, hourly and daily rainfall observations from 92 rain gauges located inside the Wupper catchment (Germany). The observations from each gauge were aggregated into 15 different durations; the resulting series were used to obtain yearly maxima. This process resulted in ca. 15 series of annual maxima for each rain gauge. The base assumption for this study is that, for a fixed duration, the maxima from the different locations was independent and identically distributed. That is, we assume that the observations from one site are independent from the observations from all other sites. Furthermore, a second

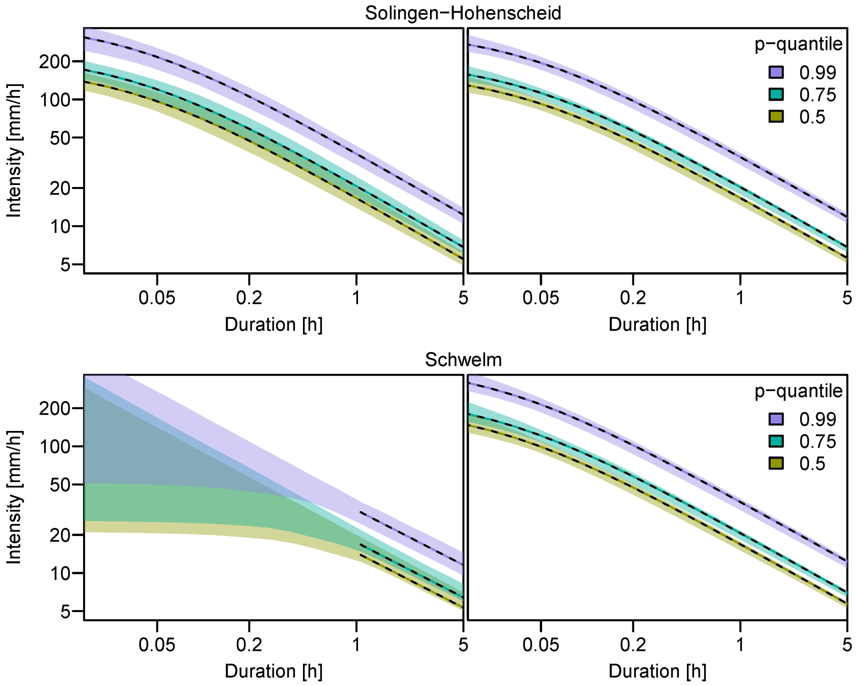

FIGURE 6.3: Bootstraped 95% confidence intervals for IDF-estimates at the example stations Solingen-Hohenscheid and Schwelm, using the station-wise d-GEV (left column) and the spatial d-GEV approach (right column). Image reproduced from Ulrich et al. (2020) with permission of the authors.

assumption of independence was done, for which maxima from different aggregation durations are seen as independent, even if they come from the same location. The validity of this second assumption is the one that was studied in ch. 4.

The goal of the study performed by Ulrich et al. (2020), was to estimate consistent IDF curves using a model that describes the spatial variability of the estimates, but does not explicitly include spatial dependence. This is done via a VGLM (see section 2.2.1 for more information about these models). The VGLM proposed in this study uses the d-GEV (eq. (3.12)), where each parameter includes spatial covariates (transformed longitude and latitude). This model is very similar to that of eqs. (5.4)-(5.6), with the difference that the d-GEV is used instead of the GEV distribution. This method was encoded in the R-package `IDF`, which uses our study as a reference (Ulrich et al., 2019). The benefits of this type of model is that it allows for very quick and efficient estimation of the estimates for each location, with reduced uncertainties thanks to the pooling of information from different rain gauges. Furthermore, the use of spatial covariates allowed for the interpolation at ungauged sites.

In addition to the model described above, two pointwise models were also fitted for comparison. For these models (1) the d-GEV was fitted to every single location, and (2) a GEV is fitted to each individual location and duration. Finally, a comparison using the verification framework detailed in section 5.2.4 of this thesis was performed.

An essential aspect of this study was determining the uncertainty of the return levels resulting from the proposed spatial d-GEV approach, and how it compared to the pointwise approach. Figure 6.3 shows a comparison between the pointwise and spatial approach of the estimated return levels for three non-exceedance probabilities with their respective 95% confidence intervals. The right column for both stations shows a clear reduction in the width of the confidence intervals, suggesting that the uncertainty was reduced when using the spatial d-GEV approach compared to the pointwise d-GEV. The biggest effect is seen for gauges where no previous information about a certain duration existed, as seen in station Schwelm.

The results shown in Fig. 6.3 appear to be a promising way to reduce the uncertainty of the estimates, but they should be interpreted carefully. This is because the assumption of independence between durations and locations for the spatial d-GEV approach results in a misspecified likelihood, as discussed above. Thus, we expected that the uncertainty estimated from the spatial d-GEV approach would be underestimated. This underestimation has two main contributions: ignoring the dependence between different locations, and ignoring the dependence between different aggregation durations. The effect of ignoring the dependence between locations was already known from Zheng et al. (2015). In contrast, the possible effect of ignoring the dependence between durations was unknown at the time. For this, the results from the study presented in ch. 4 were used, as they included an analysis of the dependence between durations for the same precipitation data in the same region. As the results of ch. 4 showed that the dependence between durations had a (mostly) negligible effect on the marginal estimates, the iid assumption for rainfall of different durations was justified.

The specific contribution to this study was the following: discussion and investigation of the effect of dependence between durations, collaboration in software development for the `IDF` package, review and editing of the manuscript, and collaboration with the data curation process.

6.3 Modeling the bivariate spatial dependence of ozone and temperature maxima using copulas

Publication details

The following section contains a summary of the contribution made to the published study titled **The impact of atmospheric blocking on the compounding effect of ozone pollution and temperature: a copula-based approach**, published in 2022 by Noelia Otero, Oscar E. Jurado, Tim Butler, and Henning W. Rust to the ACP journal. The other authors granted permission to reproduce some excerpts of text.

The previous studies shown in this dissertation have dealt only with extreme rainfall. However, the methods presented for spatial models in ch. 5 also apply to model the dependence between other environmental variables. For example, in the study of Otero et al. (2022), we investigated the spatial distribution of the bivariate dependence between the following variables:

- Daily maxima of the 8-hour average ozone concentration ($MDA8O_3$); and
- daily maximum temperature extracted from the measured values at 2 m above the ground in a 6-hourly frequency (T_{max}).

The aim of this study is to investigate how the spatial distribution of the bivariate dependence between $MDA8O_3$ and T_{max} changed when subjected to a meteorological phenomenon known as **blocking**. Blocking is a large-scale atmospheric phenomenon where persistent anticyclones interrupt the westerly flow in the midlatitudes, leading to a standstill of many weather systems. Our study hypothesized that under blocking conditions, the bivariate dependence between ozone and temperature maxima would

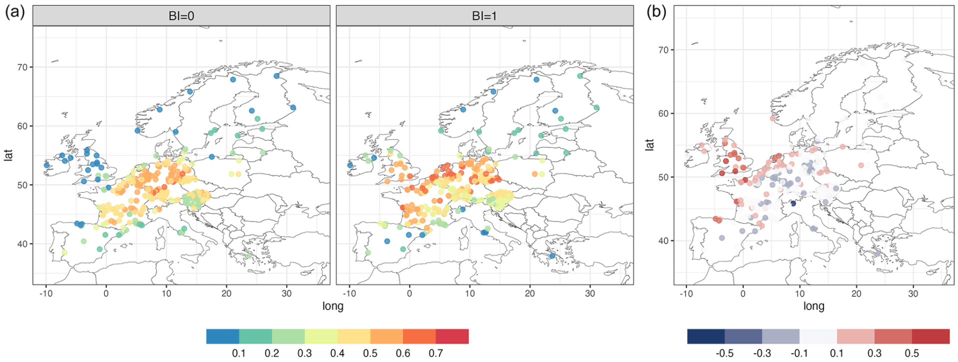

FIGURE 6.4: (a) Spatial distribution of the upper tail dependence parameter derived from the copulas when BI=0 and BI=1. (b) Difference of the upper tail dependence parameter between the two blocking scenarios. Image reproduced from Otero et al. (2022) with permission of the authors.

increase, leading to an increase in the probability of co-occurrence of both events. Knowing how much this probability changes is of great interest to the general public, as significant health issues can follow when high ozone concentrations are combined with high temperatures (Hertig, 2020).

In contrast to the max-stable process models used for the studies presented in chs. 4 and 5, the bivariate dependence for this study was modeled with copulas. As defined in section 3.6.5, copulas are models that describe the dependence structure between uniform marginals. For this study, we focused mainly on a subtype of copulas known as **extreme value copulas**. Under certain conditions, extreme value copulas can be seen as fulfilling the max-stable property, making them a valid alternative to max-stable processes to model the dependence between two extreme-valued random variables [1].

In this study the bivariate spatial dependence between MDA8O$_3$ and T_{max} for different locations s was studied. Let the random variable X represent T_{max} and the random variable Y represent MDA8O$_3$, and $F(\cdot)$ represent their respective distribution function transformed to a standard uniform distribution. The multivariate distribution was then modeled as

$$F_{(X,Y)}(x,y) = C(F_X(x), F_Y(y)), \tag{6.1}$$

where $C(\cdot)$ denotes the copula function. The advantage of this model is that the marginal distributions can be arbitrary, as they are ultimately transformed into uniform distributions. Our study focused on families of $C(\cdot)$ capable of capturing tail dependence: Student-t, Clayton, Gumbel, and Joe.

Inference for the copula parameters was done using MLE, and the copula choice was performed with the AIC [2] In a similar setting to the framework developed in the last two chapters, the copula function was fitted for two scenarios: one when blocking was present and one when blocking was not present. The presence of blocking was derived from the two-dimensional blocking index (BI), where: BI $=1$ indicated blocking, and BI $=0$ indicated no blocking.

[1] Note that for spatial extremes the choice of copulas can seem non intuitive, as copulas are basically (finite dimensional) multivariate distributions, while max-stable processes are (infinite dimensional) stochastic processes. However, in practice, we always have a finite number of measurements, so that inference is intrinsically multivariate (Ribatet et al., 2013)

[2] Once again, this shows how the likelihood (essential for computing MLE and AIC) plays crucial role in all the methods presented in this dissertation.

The use of extreme value copulas allowed for the estimation of the upper tail coefficient, a summary measure of extreme dependence analog to the extremal coefficient θ described in section 3.6.1. This coefficient indicates high levels of dependence when its value is close to 1. Figure 6.4 shows the resulting bivariate upper tail coefficient for all locations under two scenarios: BI $=$ 1, and BI $=$ 0. Observing the difference between the two scenarios, we concluded that blocking increased the upper tail dependence over northwest and central Europe.

The specific contribution to this study was the following: support in the implementation of the copulas using the `copulas` package, help with the discussion and interpretation of results, and review and editing of the manuscript. In summary, this study represents for me a transfer of knowledge from the methods of spatial extremes to a different setting.

Part III

Conclusions

7 Summary and Conclusions

This dissertation set out to improve the understanding of the statistical modeling of extreme rainfall. In particular, the main goal was to incorporate the information encoded by dependence into statistical models for extreme rainfall, expecting that this would result in more efficient use of the data. From this goal, two principal aims were set: the first was to improve the understanding of the mechanisms that lead to statistical dependence in extreme rainfall datasets. The second aim was to improve the extreme rainfall estimates made from models by incorporating the dependence in a meaningful way. To fulfill these aims, two primary studies were presented. Each study tackled the same aims, but with a different perspective. In addition to these two aims, this dissertation also set out to investigate how the methods presented could improve other types of rainfall models. For this, the contribution made to other three studies was summarized in the last chapter.

In this chapter, a summary of the two main studies conducted during this dissertation is presented. For this, the main findings of each study will be highlighted, as well as their implications in the field. Afterwards, we present some of the limitations of this dissertation, followed by possible research lines that could follow the work presented here. We finalize with a short conclusion.

7.1 Summary

Two novel studies were presented in the course of this dissertation. Both studies dealt with incorporating dependence into models for extreme rainfall but differed in the type of dependence included. The first one investigated the impact of dependence between rainfall extremes from different aggregation durations; the second one investigated the impact of spatial dependence of rain gauges in the same geographical catchment.

These studies found that, under certain conditions, this kind of statistical dependence can be modeled using Brown-Resnick max-stable processes. However, several considerations, like isotropy or the choice of distance functions, are essential to get proper results from the model. Overall, the results of the studies show that introducing the dependence to the model is a valuable way to reduce uncertainties without overestimating them; a middle point between the large uncertainty of pointwise estimation and the underestimating from iid GLM models. Furthermore, good verification results from all considered studies showed that including this dependence can lead to better pointwise estimates for specific situations.

A summary of the aims and main results from each study is given in the following subsections.

7.1.1 Summary of study 1: Impact of including dependence for IDF models

In this study, we investigated the impact of including the dependence between rainfall intensities aggregated over different durations in the d-GEV model for estimating IDF curves. This aim was achieved by building two models:

1. A model based in the d-GEV that assumed independence between rainfall maxima of different aggregation durations, which was named the **rd-GEV approach**; and
2. a model derived from a max-stable process that includes dependence between rainfall maxima of different aggregation durations, refereed to as the **MS-GEV approach**.

The development of the MS-GEV approach required the construction of a distance measure for different aggregation durations. This measure needed to reflect that heavy rainfall events are very heterogeneous for short durations (i.e., $d \leq 5$ hours) but homogeneous for long durations (i.e., $d > 12$ hours). Thus, instead of using the euclidean distance proposed by Tyralis et al. (2019), we opted to use a logarithmic distance. By comparing Fig. 4.2 with Fig. 4.3, it becomes apparent that the log-distance makes for a much more appropriate (for our purpose) dependence structure in the MS-GEV approach. Thus, one of the main findings of this study was the following:

Study 1 - Main finding 1

The distance between aggregation durations for rainfall maxima is better represented by a logarithmic distance instead of the Euclidean distance.

After estimating the parameters for the two models, we calculated the out-of-sample performance using an accuracy measure for each model. This measure was the (10-fold) cross-validated Quantile Score (QS). Having a QS value for each model allowed us to compare the performance between the rd-GEV and MS-GEV models. This comparison was summarized via the Quantile Skill Index (QSI), which incorporates the QS of both models. The QSI showed positive values when the MS-GEV model performed better and negative values when the rd-GEV model performed better. Because the MS-GEV approach has the same underlying model as the rd-GEV approach, we interpreted the QSI as a proxy of the impact of the dependence on the estimated return levels from the model.

The resulting QSI for different durations and return periods is seen in Fig. 4.6. This showed that the MS-GEV approach had an advantage for large return periods and short-to-medium durations. Therefore, we can state that:

Study 1 - Main finding 2

Including the dependence between durations in the model had a moderate-to-low positive impact on the QSI values of the final return levels, concentrated mainly on the short duration/large non-exceedance probability events.

Following this finding, for future studies we recommended including the dependence only when the focus of the study is the short duration/large non-exceedance probability region from the distribution.

Finally, a simulation study was also presented, where the impact of including the dependence between durations was shown for a low, medium, and high level of dependence (see Fig. 4.7). The procedure was to fit the rd-GEV and MS-GEV models for each dependence level, from which the QSI was calculated. This procedure allowed us to see the impact of the dependence level in the QSI. The results showed that, as expected, an increasing level of dependence led to better QSI results for the MS-GEV model. By combining these results with those from the case study, we concluded the following:

> Study 1 - Main finding 3
>
> The dependence between rainfall of different aggregation durations appears to decay quickly with increasing distance, particularly for short durations.

This last result was expected, as rainfall events can be grossly divided into two types according to their time-scales. The first one, local convective events, tend to be short and very intense. The second one are frontal/stratiform events, which tend to be much longer. Short durations tend to be associated with convective events. Thus, it makes sense that for short durations the dependence is strong but decays very quickly with increasing duration.

7.1.2 Summary of study 2: Seasonal spatial dependence

This study had two distinct aims. The first one was to describe how the spatial dependence for rainfall maxima changed when looking at two types of rainfall-generating processes: convective and frontal/stratiform. The second aim was to determine the impact of including these dependence structures when estimating return levels with a spatial model. Both aims were investigated with a case study in the region of Berlin-Brandenburg.

For the first aim, a method was proposed to divide the events into convective or frontal/stratiform. For the study region, it has been observed that extreme convective events predominate in the summer months, while frontal/stratiform extreme events are more frequent in the winter months. Therefore, we used season (summer/winter) as a proxy for the type of rainfall-generating process that resulted in the event considered to be the maximum. This method required the use of semi-annual instead of annual maxima. Further justification for this method is given by the results from Ulrich et al. (2021), who found that annual block maxima tend to come from convective events in summer and frontal/stratiform in winter for another similar region in the mid-latitudes.

After classifying the events into winter and summer maxima, we investigated the characteristics of the bivariate extremal coefficient for the pairwise case. The main results for the estimated empirical extremal coefficient are shown in Figs. 5.4 and 5.11. These figures show that a complex dynamic exists for the extremal dependence, as several regimes appeared to coexist. We compared the empirical estimates with those resulting from the model as a way to check that the model adequately captured the dependence structure. This comparison resulted in the following main finding:

Study 2 - Main finding 1

The dependence structure for summer maxima (i.e., presumably convective events) was adequately represented with an isotropic model, while the dependence for winter maxima (i.e., presumably stratiform/frontal events) was not well represented, suggesting that an anisotropic model would perform better. Furthermore, the summer maxima showed a dependence structure with a fast decay with increasing distance, while the dependence structure for the winter maxima showed a slower decay.

The results from this main finding were expected, given the general geometric considerations for convective and frontal/stratiform events: Convective events tend to be localized, while frontal events tend to be much larger with an elongated shape. As both events have very different time scales, it was necessary to investigate the extremal dependence for a sub-daily (12 hours) and daily aggregation duration.

Once the spatial dependence structure was described for both seasons and durations, we investigated the impact of including the dependence in the VGLM model proposed by Fischer et al. (2017) and used by Ulrich et al. (2020) in a similar setting. This aim was achieved using a Brown-Resnick max-stable process where the response surfaces were equivalent to those from the VGLM approach. We ended up with two types of models:

1. A model assuming spatial independence between the maxima of different locations based on the VGLM model (abbreviated as the **DM approach**); and
2. a Brown-Resnick model that includes spatial dependence (abbreviated as the **BR approach**).

Quantifying the impact of dependence was achieved by following the same verification strategy as in study 1: the QSI quantifies how much skill was gained/lost when estimating return levels of different non-exceedance probabilities by including the dependence in the model. The main results for the QSI are seen in Fig. 5.7. From this figure, we found the following:

Study 2 - Main finding 2

The isotropic Brown-Resnick model showed a good performance for summer maxima (i.e., presumably convective events). On the other hand, it showed an unsatisfactory performance for the winter maxima.

In conclusion to this study, we propose that the design of a model, especially when including dependence, must consider the type of event that will be modeled, or be flexible enough to incorporate various types of rainfall processes. Convective events appear to work well with the isotropic assumption, but other type of events, like frontal/stratiform ones require a more general model for the dependence.

7.1.3 Summary of other studies included in this dissertation

In chapter 6 the contributions made to three further studies dealing with the modeling of extreme events was described. The first study, that of Fauer et al. (2021), involved adapting the simulation developed for study 1 in order to compute the coverage of confidence intervals. This allowed to judge whether the confidence intervals were meaningful under the assumption of independence. Furthermore, in the study of

Ulrich et al. (2020), the main findings from study 1 were used to justify the choice of independence for the design of the VGLM model. Finally, in the study of Otero et al. (2022), a model that includes extreme dependence was developed following the ideas from the other studies.

The contributions done to these studies show the flexibility of the models described in this dissertations. Furthermore, they show the different contexts where methods from EVT and spatial extremes can add valuable information to the study of meteorological extreme events.

7.2 Limitations

It is important to acknowledge limitations to the work done for this dissertation. This dissertation does not engage with the theoretical development of stochastic methods for modeling extreme events. Instead, this work adopts state-of-the-art stochastical methods in order to investigate previously unknown aspects of extreme rainfall modeling. Furthermore, it is beyond the scope of this dissertation to examine the impact of climate change, including non-stationarity, in the estimates of the models. Particular limitations to each study can be found in each chapter.

7.3 Outlook

This work focused mainly on adapting several state-of-the-art statistical models from a meteorological point of view to improve the knowledge gained by modeling extreme rainfall. This research has given way to several questions in need of further investigation. Here I explore some of the main topics of interest that could lead to better models.

7.3.1 Estimating uncertainty for the MS-GEV approach used to estimate IDF curves

The goal of study 1, described in ch. 4, was to measure the impact of including the dependence between durations in the d-GEV model. This study gave valuable insight into the performance of the point estimates of different IDF models; however, it did not investigate how much the uncertainty of the estimates was affected by including the dependence. This uncertainty is essential in any EVT study where the results are used for designing infrastructure. Therefore, a valuable addition to this model would be to devise a method to obtain the uncertainty of the MS-GEV model.

As seen in section 2.4, two different ways of quantifying the uncertainty of a model exist: frequentist or Bayesian. In the published paper, we performed the inference of the parameters using MLE, which is a frequentist method. Thus, to obtain confidence intervals of the predicted return levels, we would have to use theoretical approximations (i.e., the delta method) or resampling (i.e., the bootstrap). However, the complexity of both the pairwise likelihood of the Brown-Resnick model and of the rd-GEV distribution made the necessary computations for the propagation of uncertainty intractable in their current form. Therefore, an alternative way of propagating the uncertainty for this model is required.

The issues described above with the frequentist approach to propagate the uncertainty of the MS-GEV and rd-GEV models suggest that a more straightforward path is to use Bayesian inference. This advantage arises because the Bayesian approach automatically gives the desired uncertainties as long as the assumptions are met. In this case, getting the uncertainty from the MS-GEV or rd-GEV models would require using a Bayesian method like MCMC instead of MLE. From there, it is enough to get the posterior predictive distribution (eq. (2.16)) to get credibility intervals from the return levels.

The problem of obtaining uncertainty reduces to finding an appropriate way of doing Bayesian inference with the MS-GEV model. Fortunately, most of this implementation was already done for study 2, where a Bayesian method was developed to estimate parameters for the Brown-Resnick isotropic model. The difference between the models is in the response surfaces, but this is a straightforward implementation. It should be noted that implementing the OFS correction from Shaby (2014) is necessary to get credibility intervals with the correct coverage probability when using the pairwise likelihood for Bayesian inference.

7.3.2 Incorporating anisotropy for spatial dependence

The Brown-Resnick models introduced in section 4.2 of ch. 5 assumed isotropy, meaning that the dependence was only a function of distance and not of direction. The results showed that this proposed model was appropriate when modeling the semi-annual summer maxima. However, they also showed that the model was not so adequate for the semi-annual winter maxima. An idea that could improve the model for winter events involves extending the model to allow for anisotropy in the dependence structure.

Why would anisotropy (potentially) improve the Brown-Resnick model for winter events? We know that frontal/stratiform events are predominant in the semi-annual winter maxima – As explained before, these events tend to be long and narrow, with a shape along-front of ca. 1000 km and across-front of ca. 100 km. Thus, it is logical to think that an anisotropic dependence structure for the Brown-Resnick model would better capture the properties of these kinds of events.

To include the anisotropy in the model, the variogram of the Brown-Resnick model given in eq. (3.31) needs to be modified to include anisotropy within the distance function. This type of distance that includes anisotropy is commonly known as the reduced distance. The reduced distance requires the computation of the rotation matrix, which requires a parameter ϕ to be given.

Including anisotropy in this fashion ultimately means that at least one extra parameter needs to be estimated with the data. For the case of the pairwise likelihood, the extra parameter is only one angle ϕ of the rotation matrix. This estimation requires a straightforward modification to the likelihood term, from which likelihood-based inference methods like MLE or MCMC can be used. The complexity comes when estimating the uncertainty of this new parameter. However, this information can be extracted directly from the posterior when using Bayesian inference. The problem, in that case, is the proposal of a prior distribution for ϕ. This prior could depend on what we know about the predominant types of events for the specified season. For example, it could be a wide re-centered Beta distribution with a bell shape centered around zero for summer and, for winter, centered around 45. Alternatively, the analyst could propose a more non-informative prior to see how the data shifts the posterior. These analyses would then help the analyst judge whether anisotropy is warranted or not for the data.

7.3.3 Classifying events properly

A significant limitation of study 2 (ch. 5) was using seasons as a proxy for the type of rainfall-generating process predominant in the region. This design choice was justified for the domain used in the case study of this study, as this seasonal behavior has been observed in other studies performed for the midlatitudes ((Berg et al., 2013; Ulrich et al., 2021, see, for example,)). However, for other regions (especially outside of the midlatitudes), using winter and summer as a proxy for convective or stratiform events will not be straightforward. Therefore, using an objective classification of events instead of relying on a proxy would provide a sturdier foundation for investigating the dependence structures of extreme events.

Some work in classifying extreme events already exists. For example, Lengfeld et al. (2021) developed the Catalogue of Radar-based heavy Rainfall Events (CatRaRE). This catalog provides a list of all the heavy rainfall events that affected Germany for a 20-year period; additionally, it provides many parameters describing their characteristics, like objective weather type classification. This information could be combined with annual or monthly block maxima to determine what kind of weather type the maxima came from. With this information, one could do the EVT analysis only for maxima from a specified weather type, such as frontal systems. The main problem is that 20 years could not be long enough for proper extreme value analysis.

7.4 Conclusion

The statistical modeling of extreme rainfall provides crucial information for designing and implementing of protection measures against heavy flooding events. The design of such models depends on many factors, like data availability, which typically is not sufficient to characterize an entire catchment based only on pointwise estimates. Thus, models that “borrow” strength from the similarity in space or temporal scales are needed to extract information for unobserved locations.

This dissertation begins by exploring many of the existing spatial modeling methods developed from a statistical viewpoint. Then, this dissertation aimed to provide different strategies to adapt the methods incorporating knowledge from a meteorological viewpoint. This was mainly provided by investigating how statistical dependence arose from physical process considerations. For example, how the time-scale of the different rainfall-generating mechanisms resulted in dependence between durations, and from there, how to measure its impact using a verification framework.

To summarize this work, we have found that integrating the different types of dependence impacts the estimated return levels, which leads to improved estimates under certain conditions. Taken together, these findings support strong recommendations to incorporate the dependence for extreme rainfall models where the physical characteristics of the predominant extreme events suggest a strong dependence on the spatial and temporal scale of interest.

Bibliography

Andrew, John T. and Eric Sauquet (2017). “Climate change impacts and water management adaptation in two mediterranean-climate watersheds: learning from the Durance and Sacramento Rivers”. In: *Water* 9.2, p. 126.

Aparicio, Francisco (1997). *Fundamentos de hidrología de superficie [Fundamentals of surface hydrology]*. Balderas, Mexico: Limusa, pp. 168–176.

Asadi, Peiman, Anthony C. Davison, and Sebastian Engelke (2015). “Extremes on river networks”. In: *Ann. Appl. Stat.* 9.4, pp. 2023–2050. DOI: 10.1214/15-AOAS863.

Bachmaier, Martin and Matthias Backes (2011). “Variogram or semivariogram? Variance or semivariance? Allan variance or introducing a new term?” In: *Mathematical Geosciences* 43.6, pp. 735–740.

Bentzien, Sabrina and Petra Friederichs (2014). “Decomposition and graphical portrayal of the quantile score”. In: *Q. J. R. Meteorol. Soc.* 140.683, pp. 1924–1934. DOI: 10.1002/qj.2284.

Berg, P. and J. O. Haerter (2013). “Unexpected increase in precipitation intensity with temperature - A result of mixing of precipitation types?” In: *Atmos. Res.* 119, pp. 56–61. DOI: 10.1016/j.atmosres.2011.05.012.

Berghäuser, Lisa et al. (2021). *Starkregen in Berlin: Meteorologische Ereignisrekonstruktion und Betroffenenbefragung [Heavy rain in Berlin: Meteorological reconstruction of events and survey of affected people]*. Tech. rep. Universität Potsdam, p. 44. DOI: 10.25932/publishup-50056.

Betancourt, Michael (2020). *Robust Gaussian Process Modeling*. URL: https://betanalpha.github.io/assets/case_studies/gaussian_processes.html.

Boessenkool, Berry (2021). *rdwd: Select and Download Climate Data from 'DWD' (German Weather Service)*. R package version 1.5.0. URL: https://CRAN.R-project.org/package=rdwd.

Bohnenstengel, Sylvia I., K. H. Schlünzen, and F. Beyrich (2011). “Representativity of in situ precipitation measurements - A case study for the LITFASS area in North-Eastern Germany”. In: *J. Hydrol.* 400.3-4, pp. 387–395. DOI: 10.1016/j.jhydrol.2011.01.052.

Boris, Beranger, Padoan Simone, and Marcon Giulia (2021). *ExtremalDep: Extremal Dependence Models*. R package version 0.0.3-4. URL: https://CRAN.R-project.org/package=ExtremalDep.

Brown, Bruce M and Sidney I Resnick (1977). “Extreme values of independent stochastic processes”. In: *Journal of Applied Probability* 14.4, pp. 732–739.

Buhl, Sven and Claudia Klüppelberg (2016). “Anisotropic Brown-Resnick space-time processes: estimation and model assessment”. In: *Extremes*. DOI: 10.1007/s10687-016-0257-1.

Caretta, M.A. et al. (2022). “Water”. In: *Climate Change 2022: Impacts, Adaptation, and Vulnerability. Contribution of Working Group II to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Ed. by H.-O. Pörtner et al. Cambridge, UK and New York, NY, USA: Cambridge University Press.

Chan, Raymond K.S. and Mike K.P. So (2017). “On the performance of the Bayesian composite likelihood estimation of max-stable processes”. In: *J. Stat. Comput. Simul.* 87.15, pp. 2869–2881. DOI: 10.1080/00949655.2017.1342824.

Chow, Ven Te (1953). “Frequency analysis of hydrologic data with special application to rainfall intensities”. In: *University of Illinois Bulletin* 50.81, p. 86.

CLICOM-SMN (2018). *Datos climáticos diarios del CLICOM del SMN a través de su plataforma web del CICESE*. URL: http://clicom-mex.cicese.mx.

Coles, Stuart. (2001). *An Introduction to Statistical Modeling of Extreme Values*. Springer. DOI: 10.1198/tech.2002.s73.

Cooley, Daniel, Philippe Naveau, and Paul Poncet (2006). “Variograms for spatial max-stable random fields”. In: *Dependence in probability and statistics*. Springer, pp. 373–390.

Cooley, Daniel et al. (2012). “A survey of spatial extremes: Measuring spatial dependence and modeling spatial effects”. In: *Revstat Stat. J.* 10.1, pp. 135–165.

Davison, Anthony C. and M. M. Gholamrezaee (2012a). “Geostatistics of extremes”. In: *Proc. R. Soc. A Math. Phys. Eng. Sci.* 468.2138, pp. 581–608. DOI: 10.1098/rspa.2011.0412.

Davison, Anthony C. and David V. Hinkley (1997). *Bootstrap methods and their application*. 1. Cambridge university press.

Davison, Anthony C. and Raphaël Huser (2015). “Statistics of Extremes”. In: *Annu. Rev. Stat. Its Appl.* 2, pp. 203–235. DOI: 10.1146/annurev-statistics-010814-020133.

Davison, Anthony C., Raphaël Huser, and Emeric Thibaud (2013). “Geostatistics of Dependent and Asymptotically Independent Extremes”. In: *Math. Geosci.* 45.5, SI, 511–529. DOI: 10.1007/s11004-013-9469-y.

Davison, Anthony C., Simone A. Padoan, and Mathieu Ribatet (2012b). “Statistical Modeling of Spatial Extremes”. In: *Stat. Sci.* 27.2, pp. 161–186. DOI: 10.1214/11-STS376.

de Haan, L and SI Resnick (1977). “Limit Theory for multivariate sample extremes”. In: *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* 40.4, pp. 317–337.

de Haan, Laurens (1984). “A Spectral Representation for Max-Stable Processes”. In: *Ann. Probab.* 12.4, pp. 1994–1204.

Dey, Dipak K. and Jun Yan (2016). *Extreme value modeling and risk analysis: Methods and applications*.

Dipak, Dey, Dooti Roy, and Jun Yan (2016). “Univariate Extreme Value Analysis”. In: *Extreme value modeling and risk analysis: methods and applications*. Ed. by Dipak K Dey and Jun Yan. CRC Press.

Dombry, Clément, Sebastian Engelke, and Marco Oesting (2017). “Bayesian inference for multivariate extreme value distributions”. In: *Electron. J. Stat.* 11.2, pp. 4813–4844. DOI: 10.1214/17-EJS1367.

Dombry, Clément, Sebastian Engelke, and Marco Oesting (2016). “Exact simulation of max-stable processes”. In: *Biometrika* 103.2, pp. 303–317. DOI: 10.1093/biomet/asw008.

Durrans, S. Rocky (2010). “Intensity-Duration-Frequency Curves”. In: *Rainfall: State of the Science*. American Geophysical Union (AGU), pp. 159–169. DOI: 10.1029/2009GM000919.

Dyrrdal, Anita Verpe, Alex Lenkoski, Thordis L. Thorarinsdottir, and Frode Stordal (2015). “Bayesian hierarchical modeling of extreme hourly precipitation in Norway”. In: *Environmetrics* 26.2, pp. 89–106.

Engelke, Sebastian, Alexander Malinowski, Zakhar Kabluchko, and Martin Schlather (2015). “Estimation of Hüsler-Reiss distributions and Brown-Resnick processes”. In: *J. R. Stat. Soc. Ser. B Stat. Methodol.* 77.1, 239–265. DOI: 10.1111/rssb.12074.

Fauer, Felix S., Jana Ulrich, Oscar E. Jurado, and Henning W. Rust (2021). “Flexible and consistent quantile estimation for intensity-duration-frequency curves”. In: *Hydrol. Earth Syst. Sci.* 25.12, pp. 6479–6494. DOI: 10.5194/hess-25-6479-2021.

Fischer, Madlen, Henning W. Rust, and Uwe Ulbrich (2017). “A spatial and seasonal climatology of extreme precipitation return-levels: A case study”. In: *Spat. Stat.* DOI: 10.1016/j.spasta.2017.11.007.

Fondeville, R. de and Anthony C. Davison (2018). “High-dimensional peaks-over-threshold inference”. In: *Biometrika* 105.3, pp. 575–592. DOI: 10.1093/biomet/asy026.

Fondeville, Raphaël de and Anthony C. Davison (2020). *Functional Peaks-over-threshold Analysis*. DOI: 10.48550/ARXIV.2002.02711. URL: https://arxiv.org/abs/2002.02711.

Fukutome, S, A Schindler, and A Capobianco (2018). *MeteoSwiss extreme value analyses: User manual and documentation. 3rd Edition.* Tech. rep. MeteoSwiss.

Ganguli, Poulomi and Paulin Coulibaly (2017). “Does Nonstationarity in Rainfall Requires Nonstationary Intensity-Duration-Frequency Curves?” In: *Hydrol. Earth Syst. Sci. Discuss.*, pp. 1–31. DOI: 10.5194/hess-2017-325.

— (2019). “Assessment of future changes in intensity-duration-frequency curves for Southern Ontario using North American (NA)-CORDEX models with nonstationary methods”. In: *J. Hydrol.-Reg. Stud.* 22. DOI: {10.1016/j.ejrh.2018.12.007}.

García-Bartual, R and M Schneider (2001). “Estimating Maximum Expected Short-Duration Convective Storms”. In: *Phys. Chem. Earth, Part B* 26.9, pp. 675–681.

Gebauer, P., G. Myrcik, and F. Schenk (2017). *Beiträge zur Berliner Wetterkarte - Berlin unter Wasser*. URL: https://berliner-wetterkarte.de/Beilagen/2017/BWK_Beitraege_20170714_Berlin_unter_Wasser.pdf.

Gelfand, Alan E., Peter Diggle, Peter Guttorp, and Montserrat Fuentes (2010). *Handbook of spatial statistics*. CRC press.

Gelman, Andrew and Donald B. Rubin (1992). “Inference from Iterative Simulation Using Multiple Sequences”. In: *Statistical Science* 7.4, pp. 457 –472. DOI: 10.1214/ss/1177011136. URL: https://doi.org/10.1214/ss/1177011136.

Glickman, T.S. (2000). *Glossary of meteorology*. American Meteorological Society.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman (2009). *Elements of Statistical Learning 2nd ed.* DOI: 10.1007/978-0-387-84858-7.

Hastings, W. K. (1970). “Monte Carlo sampling methods using Markov chains and their applications”. In: *Biometrika* 57.1, pp. 97–109. DOI: 10.1093/biomet/57.1.97.

Hertig, Elke (2020). “Health-relevant ground-level ozone and temperature events under future climate change using the example of Bavaria, Southern Germany”. In: *Air Quality, Atmosphere & Health* 13.4, pp. 435–446.

Hosking, J. R. M. and James R. Wallis (1997). *Regional Frequency Analysis: An Approach Based on L-Moments*. Cambridge University Press. DOI: 10.1017/CBO9780511529443.

Huser, Raphaël and Jennifer L. Wadsworth (2022). "Advances in statistical modeling of spatial extremes". In: *WIREs Computational Statistics* 14.1, e1537. DOI: https://doi.org/10.1002/wics.1537.

IPCC (2022). "Summary for Policymakers". In: *Climate Change 2022: Impacts, Adaptation, and Vulnerability. Contribution of Working Group II to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Ed. by H.-O. Pörtner et al. Cambridge, UK and New York, NY, USA: Cambridge University Press.

Jurado, Oscar E., Jana Ulrich, Marc Scheibel, and Henning W. Rust (2020). "Evaluating the Performance of a Max-Stable Process for Estimating Intensity-Duration-Frequency Curves". In: *Water* 12.12, p. 3314. DOI: 10.3390/w12123314.

Kabluchko, Zakhar, Martin Schlather, and Laurens de Haan (2009). "Stationary max-stable fields associated to negative definite functions". In: *Ann. Probab.* 37.5, pp. 2042–2065. DOI: 10.1214/09-AOP455.

Koutsoyiannis, Demetris, Demosthenes Kozonis, and Alexandros Manetas (1998). "A mathematical framework for studying rainfall intensity-duration-frequency relationships". In: *J. Hydrol.* 206, pp. 118–135.

Kruschke, John (2014). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. Academic Press.

Lackmann, Gary (2011). *Midlatitude synoptic meteorology*. American Meteorological Society.

Le, Phuong Dong, Michael Leonard, and Seth Westra (2018). "Modeling Spatial Dependence of Rainfall Extremes Across Multiple Durations". In: *Water Resour. Res.* 54.3, pp. 2233–2248. DOI: 10.1002/2017WR022231.

Lehmann, Eric A., Aloke Phatak, Alec G. Stephenson, and Rex Lau (2016). "Spatial modelling framework for the characterisation of rainfall extremes at different durations and under climate change". In: *Environmetrics* 27.4, 239–251. DOI: 10.1002/env.2389.

Lengfeld, Katharina, Ewelina Walawender, Tanja Winterrath, and Andreas Becker (2021). "CatRaRE: A Catalogue of radar-based heavy rainfall events in Germany derived from 20 years of data". In: *Meteorol. Zeitschrift* 30.6, pp. 469–487. DOI: 10.1127/metz/2021/1088.

Lengfeld, Katharina, Tanja Winterrath, Thomas Junghänel, Mario Hafer, and Andreas Becker (2019). "Characteristic spatial extent of hourly and daily precipitation events in Germany derived from 16 years of radar data". In: *Meteorol. Zeitschrift* 28.5, pp. 363–378. DOI: 10.1127/metz/2019/0964.

Marcon, G., S. A. Padoan, P. Naveau, P. Muliere, and J. Segers (2017). "Multivariate nonparametric estimation of the Pickands dependence function using Bernstein polynomials". In: *J. Stat. Plan. Inference*. DOI: 10.1016/j.jspi.2016.10.004.

Mélèse, Victor, Juliette Blanchet, and Gilles Molinié (2018). "Uncertainty estimation of Intensity–Duration–Frequency relationships: A regional analysis". In: *J. Hydrol.* 558, pp. 579–591. DOI: {10.1016/j.jhydrol.2017.07.054}.

Metropolis, Nicholas, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller (1953). "Equation of State Calculations by Fast Computing Machines". In: *The Journal of Chemical Physics* 21.6, pp. 1087–1092. DOI: 10.1063/1.1699114.

Mikkola, Petrus et al. (2021). *Prior knowledge elicitation: The past, present, and future.* DOI: 10.48550/ARXIV.2112.01380. URL: https://arxiv.org/abs/2112.01380.

Monjo, Robert (2016). "Measure of rainfall time structure using the dimensionless n-index". In: *Clim. Res.* 67.1, pp. 71–86. DOI: 10.3354/cr01359.

Muller, Aurélie, Jean Noël Bacro, and Michel Lang (2008). "Bayesian comparison of different rainfall depth-duration-frequency relationships". In: *Stoch. Environ. Res. Risk Assess.* 22.1, pp. 33–46. DOI: 10.1007/s00477-006-0095-9.

Nadarajah, S., E. Afuecheta, and S. Chan (2019). "Ordered random variables". In: *Opsearch* 56.1, 344–366. DOI: 10.1007/s12597-019-00355-6.

Nadarajah, S., C. W. Anderson, and J. A. Tawn (1998). "Ordered multivariate extremes". In: *J. R. Stat. Soc. Ser. B Stat. Methodol.* 60.2, 473–496. DOI: 10.1111/1467-9868.00136.

Nelsen, Roger B. (2007). *An introduction to copulas.* Springer Science & Business Media.

Oesting, Marco, Mathieu Ribatet, and Clément Dombry (2016). "Simulation of Max-Stable Processes". In: *Extreme value modeling and risk analysis: methods and applications.* Ed. by Dipak K Dey and Jun Yan. CRC Press.

Opitz, Thomas (2013). "Extremal t processes: Elliptical domain of attraction and a spectral representation". In: *Journal of Multivariate Analysis* 122, pp. 409–413.

Orlanski, Isidoro (1975). "A rational subdivision of scales for atmospheric processes". In: *Bulletin of the American Meteorological Society* 56, pp. 527–530.

Otero, Noelia, Oscar E. Jurado, Tim Butler, and Henning W. Rust (2022). "The impact of atmospheric blocking on the compounding effect of ozone pollution and temperature: A copula-based approach". In: *Atmos. Chem. Phys.* 22.3, pp. 1905–1919. DOI: 10.5194/acp-22-1905-2022.

Padoan, Simone A., Mathieu Ribatet, and S. A. Sisson (2010). "Likelihood-based inference for max-stable processes". In: *J. Am. Stat. Assoc.* 105.489, pp. 263–277. DOI: 10.1198/jasa.2009.tm08577.

Padulano, Roberta, Alfredo Reder, and Guido Rianna (2019). "An ensemble approach for the analysis of extreme rainfall under climate change in Naples (Italy)". In: *Hydrol. Process.* 33.14, 2020–2036. DOI: {10.1002/hyp.13449}.

R Core Team (2020). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing. Vienna, Austria.

Ribatet, Mathieu (2013). "Spatial extremes: Max-stable processes at work". In: *J. la Société Française Stat. Rev. Stat. appliquée* 154.2, pp. 156–177.

— (2020). *SpatialExtremes: Modelling Spatial Extremes.* R package version 2.0-8. URL: https://CRAN.R-project.org/package=SpatialExtremes.

Ribatet, Mathieu, Daniel Cooley, and Anthony C. Davison (2012). "Bayesian Inference from composite likelihoods, with an application to spatial extremes". In: *Stat. Sinica* 22.2, pp. 813–845. DOI: 10.5705/ss.2009.248.

Ribatet, Mathieu, Clément Dombry, and Marco Oesting (2016). "Spatial Extremes and Max-Stable Processes". In: *Extreme value modeling and risk analysis: methods and applications.* Ed. by Dipak K Dey and Jun Yan. CRC Press.

Ribatet, Mathieu and Mohammed Sedki (2013). "Extreme value copulas and max-stable processes". In: *Journal de la société française de statistique* 154.1, pp. 138–150.

Richards, Jordan, Jonathan A. Tawn, and Simon Brown (2021). *Modelling Extremes of Spatial Aggregates of Precipitation using Conditional Methods.* DOI: 10.48550/ARXIV.2102.10906.

Ritschel, Christoph, Uwe Ulbrich, Peter Nevir, and Henning W. Rust (2017). "Precipitation extremes on multiple timescales - Bartlett-Lewis rectangular pulse model and intensity-duration-frequency curves". In: *Hydrol. Earth Syst. Sci.* 21.12, 6501–6517. DOI: {10.5194/hess-21-6501-2017}.

Rootzén, Holger and Richard W. Katz (2013). "Design Life Level: Quantifying risk in a changing climate". In: *Water Resour. Res.* 49.9, pp. 5964–5972. DOI: 10.1002/wrcr.20425.

Sang, Huiyan (2016). "Composite Likelihood for Extreme Values". In: *Extreme value modeling and risk analysis: methods and applications*. Ed. by Dipak K Dey and Jun Yan. CRC Press.

Schall, Robert (2012). "The empirical coverage of confidence intervals: Point estimates and confidence intervals for confidence levels". In: *Biometrical J.* 54.4, pp. 537–551. DOI: 10.1002/bimj.201100134.

Schlather, Martin (2002). "Models for stationary max-stable random fields". In: *Extremes* 5.1, pp. 33–44.

Serinaldi, Francesco (2015). "Dismissing return periods!" In: *Stoch. Environ. Res. Risk Assess.* 29.4, pp. 1179–1189. DOI: 10.1007/s00477-014-0916-1.

Shaby, Benjamin A. (2014). "The Open-Faced Sandwich Adjustment for MCMC Using Estimating Functions". In: *J. Comput. Graph. Stat.* 23.3, pp. 853–876. DOI: 10.1080/10618600.2013.842174.

Singh, Vijay P. and Lan Zhang (2007). "IDF curves using the Frank Archimedean copula". In: *J. Hydrol. Eng.* 12.6, 651–662. DOI: {10.1061/(ASCE)1084-0699(2007)12:6(651)}.

Smith, Richard L. (1990a). "Max-stable processes and spatial extremes". In: *Unpublished manuscript* 205, pp. 1–32.

Smith, Rl L. (1990b). "Max-stable processes and spatial extremes". In: *Unpubl. Manuscr.*

Stan Development Team (n.d.). *RStan: the R interface to Stan.* R package version 2.26.11. URL: https://mc-stan.org/.

— (2022). *Stan Modeling Language Users Guide and Reference Manual, Version 2.29.0.*

Stephenson, A. G. (2002). "evd: Extreme Value Distributions". In: *R News* 2.2, p. 0. URL: https://CRAN.R-project.org/doc/Rnews/.

Stephenson, Alec (2016). "Bayesian Inference for Extreme Value Modelling". In: *Extreme value modeling and risk analysis: methods and applications*. Ed. by Dipak K Dey and Jun Yan. CRC Press.

Stephenson, Alec G., Eric A. Lehmann, and Aloke Phatak (2016). "A max-stable process model for rainfall extremes at different accumulation durations". In: *Weather Clim. Extrem.* 13, pp. 44–53. DOI: 10.1016/j.wace.2016.07.002.

Thibaud, Emeric, Juha Aalto, Daniel Cooley, Anthony C. Davison, and Juha Heikkinen (2016). "Bayesian inference for the brown-resnick process, with an application to extreme low temperatures". In: *Ann. Appl. Stat.* 10.4, pp. 2303–2324. DOI: 10.1214/16-AOAS980.

Tobler, Waldo R. (1970). "A computer movie simulating urban growth in the Detroit region". In: *Economic geography* 46.sup1, pp. 234–240.

Tyralis, Hristos and Andreas Langousis (2019). "Estimation of intensity–duration–frequency curves using max-stable processes". In: *Stoch. Environ. Res. Risk Assess.* 33.1, 239–252. DOI: 10.1007/s00477-018-1577-2.

Ulrich, Jana, Felix S. Fauer, and Henning W. Rust (2021). "Modeling seasonal variations of extreme rainfall on different timescales in Germany". In: *Hydrol. Earth Syst. Sci.* 25.12, pp. 6133–6149. DOI: 10.5194/hess-25-6133-2021.

Ulrich, Jana, Oscar E. Jurado, and Henning W. Rust (2020). "Estimating IDF curves consistently over durations with spatial covariates". In: *Water* 12.11. DOI: 10.3390/w12113119.

Ulrich, Jana and Christoph Ritschel (2019). *IDF: Estimation and Plotting of IDF Curves.* R package version 0.0.2.

Umlauf, Nikolaus and Thomas Kneib (2018). "A primer on Bayesian distributional regression". In: *Stat. Modelling* 18.3-4, pp. 219–247. DOI: 10.1177/1471082X18759140.

Van de Vyver, H. and J. Van den Bergh (2018). "The Gaussian copula model for the joint deficit index for droughts". In: *J. Hydrol.* 561, 987–999. DOI: {10.1016/j.jhydrol.2018.03.064}.

Van de Vyver, Hans (2015). "Bayesian estimation of rainfall intensity-duration-frequency relationships". In: *J. Hydrol.* 529, pp. 1451–1463. DOI: 10.1016/j.jhydrol.2015.08.036.

— (2018). "A multiscaling-based intensity–duration–frequency model for extreme precipitation". In: *Hydrol. Process.* 32.11, pp. 1635–1647. DOI: 10.1002/hyp.11516.

Vehtari, Aki, Andrew Gelman, and Jonah Gabry (2017a). "Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC". In: *Stat. Comput.* 27.5, pp. 1413–1432. DOI: 10.1007/s11222-016-9696-4.

— (2017b). "Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC". In: *Statistics and computing* 27.5, pp. 1413–1432.

Vehtari, Aki, Andrew Gelman, Daniel Simpson, Bob Carpenter, and Paul-Christian Bürkner (2021). "Rank-normalization, folding, and localization: An improved Rhat for assessing convergence of MCMC (with Discussion)". In: *Bayesian analysis* 16.2, pp. 667–718.

Vettori, Sabrina, Raphaël Huser, and Marc G. Genton (2018). "A comparison of dependence function estimators in multivariate extremes". In: *Stat. Comput.* 28.3, pp. 525–538. DOI: 10.1007/s11222-017-9745-7.

Vogel, Richard M. and Attilio Castellarin (2017). "Risk, Reliability, and Return Periods and Hydrologic Design". In: *Handbook of applied hydrology.* Ed. by Vijay P Singh, D Eng, et al. McGraw-Hill Education.

Wadsworth, Jennifer L. and Jonathan Tawn (2019). *Higher-dimensional spatial extremes via single-site conditioning.* DOI: 10.48550/ARXIV.1912.06560.

Walther, Andi and Ralf Bennartz (2006). "Radar-based precipitation type analysis in the Baltic area". In: *Tellus, Ser. A Dyn. Meteorol. Oceanogr.* 58.3, pp. 331–343. DOI: 10.1111/j.1600-0870.2006.00183.x.

Watanabe, Sumio and Manfred Opper (2010). "Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory." In: *Journal of machine learning research* 11.12.

Wilks, Daniel S. (2011). *Statistical methods in the atmospheric sciences.* Vol. 100. Academic press.

Wood, Simon N. (2015). *Core statistics.* 6. Cambridge University Press.

World Meteorological Organization (2021). *WMO Atlas of Mortality and Economic Losses from Weather, Climate and Water Extremes (1970-2019).* eng. Geneva, Switzerland.

Yee, Thomas W. (2015). *Vector generalized linear and additive models: with an implementation in R.* Vol. 10. Springer.

Zheng, Feifei, Emeric Thibaud, Michael Leonard, and Seth Westra (2015). "Assessing the performance of the independence method in modeling spatial extreme rainfall". In: *Water Resour. Res.* 51.9, pp. 7744–7758. DOI: 10.1002/2015WR016893.