Data-Based Analysis of Extreme Events: Inference, Numerics and Applications

Doctoral Dissertation submitted to the
Faculty of Informatics of the Università della Svizzera italiana
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

presented by
Olga Kaiser

under the supervision of
Prof. Illia Horenko

January 2015

Dissertation Committee

Prof. Rupert Klein	Freie University Berlin, Germany
Prof. Rolf Krause	Universitá della Svizzera italiana, Switzerland
Prof. Simone Padoan	Bocconi University of Milan, Italy
Prof. Igor Pivkin	Universitá della Svizzera italiana, Switzerland
Prof. Olivia Romppainen	University of Bern, Switzerland

Dissertation accepted on 12 January 2015

Research Advisor	PhD Program Director
Prof. Illia Horenko	**Prof. Igor Pivkin**

I certify that except where due acknowledgement has been given, the work presented in this thesis is that of the author alone; the work has not been submitted previously, in whole or in part, to qualify for any other academic award; and the content of the thesis is the result of work which has been carried out since the official commencement date of the approved research program.

Olga Kaiser
Lugano, 12 January 2015

To my grandparents

Ernst & Maria Kaiser
and
Wasilij & Alexandra Adreev

Abstract

The concept of extreme events describes the above average behavior of a process, for instance, heat waves in climate or weather research, earthquakes in geology and financial crashes in economics. It is significant to study the behavior of extremes, in order to reduce their negative impacts. Key objectives include the identification of the appropriate mathematical/statistical model, description of the underlying dependence structure in the multivariate or the spatial case, and the investigation of the most relevant external factors. Extreme value analysis (EVA), based on Extreme Value Theory, provides the necessary statistical tools. Assuming that all relevant covariates are known and observed, EVA often deploys statistical regression analysis to study the changes in the model parameters. Modeling of the dependence structure implies a priori assumptions such as Gaussian, locally stationary or isotropic behavior. Based on EVA and advanced time-series analysis methodology, this thesis introduces a semiparametric, nonstationary and non-homogenous framework for statistical regression analysis of spatio-temporal extremes. The involved regression analysis accounts explicitly for systematically missing covariates; their influence was reduced to an additive nonstationary offset. The nonstationarity was resolved by the Finite Element Time Series Analysis Methodology (FEM). FEM approximates the underlying nonstationarity by a set of locally stationary models and a nonstationary hidden switching process with bounded variation (BV). The resulting FEM-BV-EVA approach goes beyond a priori assumptions of standard methods based, for instance, on Bayesian statistics, Hidden Markov Models or Local Kernel Smoothing. The multivariate/spatial extension of FEM-BV-EVA describes the underlying spatial variability by the model parameters, referring to hierarchical modeling. The spatio-temporal behavior of the model parameters was approximated by locally stationary models and a spatial nonstationary switching process. Further, it was shown that the resulting spatial FEM-BV-EVA formulation is consistent with the max-stability postulate and describes the underlying dependence structure in a nonparametric way. The proposed FEM-BV-EVA methodology was integrated into the existent FEM MATLAB toolbox. The FEM-BV-EVA framework is computationally efficient as it deploys gradient free MCMC based optimization methods and numerical solvers for constrained, large, structured quadratic and linear problems. In order to demonstrate its performance, FEM-BV-EVA was applied to various test-cases and real-data and compared to standard methods. It was shown that parametric approaches lead to biased results if significant covariates are unresolved. Comparison to nonparametric methods based on smoothing regression revealed their weakness, the locality property and the inability to resolve discontinuous functions. Spatial FEM-BV-EVA was applied to study the dynamics of extreme precipitation over Switzerland. The analysis identified among others three major spatially dependent regions.

Acknowledgements

The years of my dissertation studies have been the most explorative, challenging and interesting in my life so far. I would like to express my gratitude to all the people who supported and encouraged me during this time. In the following, I would like to mention the key people involved.

First of all, I want to thank my research advisor and mentor Professor Dr. Illia Horenko for his continuous support during my PhD studies, for his encouragement and guidance, and at the same time the trust and patience when letting me exploring the "essence of science" on my own. Illia contributed significantly towards my development as a scientist by teaching me how to address scientific questions from many different perspectives and how to express ideas clearly and consistent.

I would like to thank the members of my dissertation committee for their helpful comments which helped to improve this work. I thank all the present and past members of the working group "Computational Time Series Analysis". A special thank goes to Dimitri Igdalov, Dr. Lars Putzig, Dr. Philipp Metzner and Dr. Jana de Wiljes for the many helpful discussions that lead to a deeper understanding of scientific problems I faced during this years. I would like to acknowledge all my colleagues at the Institute of Computational Science in Lugano for providing a pleasant working environment. Further, a special acknowledgment goes to Nadine Maier, alias Apka, for her permanent supportive and motivating presence. Nadine is a great online colleague and a wonderful friend. I thank the University of Lugano for financial and organizational support during my studies.

I am deeply grateful to all my family members. I am particularly grateful to my parents, Sergej & Valentina Kaiser not only for their continuous love, support, understanding and care through my life thus far, but also for listening, offering me advice, and motivating me when ever I need it.

The most important acknowledgments go to my best friend, biggest supporter, greatest critic, and my husband, Dimitri Igdalov. You enriched my scientific way of thinking in many different ways. In particular, your way of being curious, your desire of a fundamental understanding of a problem, and your requirements on me to communicate simply and clearly continuously improve my personal development. Thank you for your steady support and patience, the numerous discussions and late night walks during this years and lastly, for having humor and being foolish in difficult and frustrating times.

Contents

Figures

Tables

1 Introduction

It is in our human nature to observe and to understand our environment. Observing a process enables us to draw conclusions about its behavior, namely, to find a rule which explains the underlying dynamics and can be used for future evaluations of the process. In some cases the aim is to study the average behavior of a process, in other cases, we are interested in their outliers defined as events that occur unexpected, rare and irregular. Such events are referred to as extreme events or just extremes. In the context of a statistical model, extremes are usually located in the tails of the corresponding distribution, referring to the illustrated "trunk" and the "tail" of an elephant, please see Figure 1.1. The relevance of an extreme event can be measured by its social and/or financial impact. While

Figure 1.1. *"A boa constrictor digesting an elephant."*[36]

we are pleasantly surprised by extremes with a positive impact, we are concerned about extremes that result in social and financial losses. In order to reduce these potential losses, it is important to study extremes of processes and systems surrounding our daily life, including environmental, industrial and economical processes[2,42]. In this thesis, we will focus on statistical description of extremes that describe the outliers of such processes, for instance, heavy precipitation in hydrological systems, material strength in material sciences and financial crashes in economics. Of particular interest are extremes that are defined either as partial maximal/minimal values or as excesses beyond

a predefined threshold, for instance, annual flood levels and large insurance claims, respectively. Data-based analysis of such events is a challenging task. First, we have to solve an inverse problem given only historical observations of extremes with the aim of finding the best descriptive model of the underlying dynamics in a certain class. Secondly, the corresponding model should not only describe the behavior of observed extremes, but also be sophisticated and reliable for drawing conclusions about the behavior of "more extreme" extremes. For example, dikes or damns for flood control are designed to protect from the 100-year event, while the level of such an event is estimated based on observations of 10-year events[2,6,35]. Lastly, extremes are rare, occur irregularly, and usually there is no known deterministic formulation based on obvious physical laws.

The first attempts to study the behavior of extremes were made in hydrological engineering focusing on the level/intensity of floods and droughts, and date back to ancient times, where the agrarian economy and the main system of communication were strongly dependent on the water flow. The importance increased during the industrial age with the invention of hydro-electric plants[52]. Until the mid of the 20th century engineers applied the less successful empirical methods to study the behavior of extremes. For instance, the frequency of a major flood was estimated by dividing the number of such events by the corresponding time span. Then, the statistical nature of sample extremes was recognized and Fréchet obtained in 1927 the first result: an asymptotic parametric distribution of sample extremes, please review Figure 1.2 for an example. Fréchet also introduced

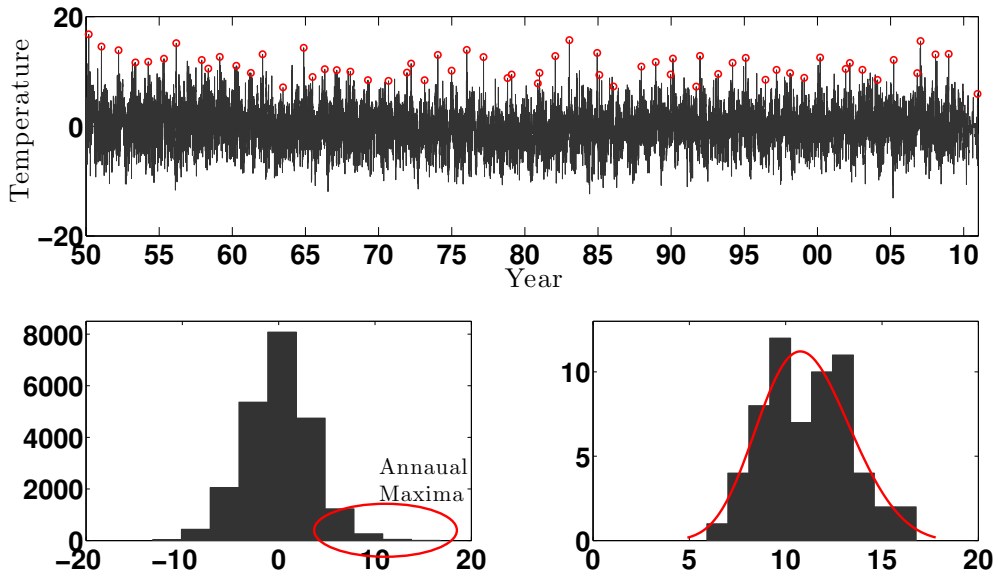


Figure 1.2. *The upper panel shows the deseasonalized daily temperature in Celsius measured at location Lugano, Switzerland (46°N, 8.9667°E). The bottom left panel presents the histogram of the same data. In both figures the annual maxima are circled in red. The histogram of the annual maxima is shown in the bottom right panel. The red line is the estimated extreme value distribution.*

the so called max-stability postulate, under which a distribution is qualified for predictions of "more extreme" events[46,52]. Based on max-stability, Tippet and Fisher extended the result of Fréchet by introducing another two limiting distributions for extremes[44]. The foundation for Extreme Value Theory (EVT) was laid. Other fundamental results were obtained among others by deHaan in 1970 and Pickands in 1975, including asymptotic results for threshold excesses[35,52]. With this, extreme value analysis (EVA) results in fitting the appropriate extreme value distribution to observed extremes.

In order to complete the analysis of extreme events, it is important to analyze their occurrence. For this purpose results from Point Process Theory can be deployed. Point processes such as the Poisson process estimates the probability of the number of events within a predefined period and the duration time between the next event given the occurrence of the previous one[22]. This dissertation thesis will focus on the analysis of the level of extremes in EVT and will not tackle the issues from the Point Process Theory.

The application of EVT assumes that the observations of the process from which the extremes are extracted are independent and identically distributed, implying stationarity of the underlying dynamics. In real applications, this assumptions may not always be true, for example, hydrological processes like precipitation depend on the season. The most general way to account for the nonstationarity of the underlying process is to refer to the time dependent parameters of the corresponding distribution of extremes. Further objective is to understand if the behavior of extremes is governed by external influences. To study this question, the parameters of the distribution are expressed by regression models, referring to statistical regression analysis of extremes. In this context, the aim is to estimate the values of the regression parameters from historical observations. Standard state-of-the-art methods in this field are divided into two groups: parametric and nonparametric regression methods. In parametric approaches the model parameters are expressed as predefined functions such as the sine/cosine functions to model the seasonal trends in meteorology[22]. A disadvantage of parametric approaches is the assumption that either all relevant covariates are known, or that the unknown covariates are independent and identically distributed. As a result, these methods implicitly assume time independence of the involved regression coefficients. But, due to the multiscale nature of most realistic processes, for example, in climate research and economics, one would never be able to guarantee that the set of the information collected about the analyzed process is complete. One would also not be able to guarantee that all of the necessary probabilistic assumptions are fulfilled a priori for the analyzed process. Nonparametric approaches for regression analysis of extreme events are based, for instance, on Local Likelihood Smoothing[32] or Bayesian techniques[24,82]. The limitations of these methods are their locality, that is, the local stationarity assumption in some local temporal window of predefined width, and a priori parametric assumptions about the distributions of the model parameters. Another strategy is to involve mixture models and Hidden Markov Models (HMM)[7,10,74]. Such approaches require a priori knowledge about the probabilistic model for the time-dependent model parameters such as stationarity and the Markov assumption for the hidden switching process.

Another relevant subject in extreme value analysis is the exploration of the relationship among extremes which are extracted from different processes. Here, the main goals are to investigate the

behavior of extremes for each single process and to study if and how the different processes influence each other in terms of the intensity of extremes. The latter is approached by the multivariate EVT and is mainly applied to environmental, climatological and economical problems. In the case when the different processes correspond to observations at different spatial locations, we refer to spatial EVA. Here, the relationship among locations with respect to the occurrence of extremes is denoted as the spatial dependence structure, measured by the joint probability. Unlike univariate EVT, there exists no known closed probability distribution for the multivariate extremes, but rather a wide range of different descriptions of the underlying dependence structure. The state-of-the-art methods approximate the underlying dependence structure assuming, for instance, Gaussian, stationary and isotropic behavior. In nonparametric approaches the dependence structure can be approximated by a combination of predefined kernel functions[25,31,62].

Spatial extreme value analysis is a very active research field. Thereby, it is important to detect the external factors that have the most significant influence on the dynamics of extremes and to describe the underlying dependence structure beyond strong a priori assumptions. The goal of this thesis is to introduce a novel data-based framework for spatio-temporal statistical regression analysis of extremes based on EVT and advanced time series analysis techniques. We will approach explicitly the task of what happens if significant covariates are missing in the deployed regression model. In particular, we will show that unresolved covariates can be reflected by an additive nonstationary offset. In order to resolve this nonstationarity in a nonparametric way, we apply the Finite Element Time Series Analysis Methods with Bounded Variation of model parameters (FEM-BV)[58,78]. The resulting univariate FEM-BV-EVA approach goes beyond probabilistic a priori assumptions of methods based, for instance, on nonstationary Bayesian mixture models, smoothing kernel methods or neural networks. Furthermore, we are interested in a nonparametric description of the underlying dependence structure among different locations. For this, we extend the FEM-BV-EVA towards space-time clustering of extremes remaining consistent with the max-stability postulate. The resulting spatio-temporal FEM-BV-EVA provides a pragmatic, nonparametric and nonstationary description of the spatial dependence structure. Finally, based on FEM-BV-EVA we provide a computationally efficient framework for statistical regression analysis of extremes that can be straightforwardly applied to large real-world problems. Therefore, we consider different optimization techniques including gradient free MCMC based methods and numerical solvers for constrained, large, structured quadratic and linear problems, which can be straightforwardly implemented as highly-scalable applications in HPC context, using existent parallel libraries. We will demonstrate the proposed framework on test-cases and real data. For real applications we will focus on climatological and meteorological data, as these data are easily accessible and can be tagged to real geological locations. Further, in the context of anthropogenic climate change, an increase of extremes in hydrological and climatological systems might be expected[39]. Hence, analyzing extremes of such processes is a present and important problem.

This thesis is organized as follows: in Chapter 2 we review related work of univariate and multivariate EVT, discuss the limitations of state-of-the-art-approaches, and motivate the purpose of this thesis. In Chapter 3 the methodology of the FEM-BV-EVA framework is proposed and derived in details. Additionally, this chapter contains the conceptual comparison of FEM-BV-EVA and the

state-of-the-art-approaches in EVA. Chapter 4 presents the extension of the FEM framework towards spatial Extreme Value Analysis. In Chapter 5 we first demonstrate on test-cases and real applications that parametric approaches provide biased results in the case of unresolved/missing covariates. Further, we compare the proposed framework with nonparametric standard approaches, based on smoothing regression. The performance of spatial FEM-BV-EVA is demonstrated on a real application. Chapter 6 contains the conclusion of the thesis and provides ideas for the extensions of the prosed methodology.

2 Statistical Extreme Value Analysis

In many real applications we intend to solve an inverse problem: given only the measurements of a process the aim is to find the best descriptive model of the underlying dynamics. For some inverse problems, the underlying dynamics can be described by physical laws, for example, in geophysics and biophysics. In such a case, the inverse problem is solved by extracting the corresponding physical parameters from the measurements through the solution of a system of equations[41]. However, in cases where observed measurements do not follow some obvious physical laws, statistical concepts provide an alternative to describe the underlying dynamics by a statistical model. A statistical model is defined by a probability distribution of the particular measurements and summarizes the main descriptive features like the average value and the range of values around it[4,101]. A brief introduction of some important concepts relevant for statistical modeling is given in Appendix A. In the case where the underlying distribution is known, the aim is to estimate the most likely parameters for the given data. Often the underlying distribution is unknown and so we focus on a certain class of distributions which satisfy some desirable properties. Such a class is the class of "stable distributions". A distribution is considered as a "stable distribution", if the sum of $n > 0$ independent random variables from this distribution preserves the shape and skewness while changing the scale and location parameters. Stable distributions provide a wide range of robust models, for instance, the family of Gaussian distributions[83,107,116]. Another approach is to approximate the underlying distribution exploiting statistical limit laws. A famous example is the Central Limit Theorem (CLT), which states that for a set of random identically and independently distributed (i.i.d) variables $X_1, X_2, \ldots$ with finite mean μ and finite variance σ^2, the limit behavior of their normalized partial sums, $\bar{X}_n = \sum_{i=1}^{n} X_i$, is the Normal distribution

$$\sqrt{n}\frac{\bar{X}_n - \mu}{\sigma} \to \mathcal{N}(0,1) \text{ as } n \to \infty \tag{2.1}$$

$$\text{with } \mu = \frac{X_1 + X_2 + \ldots X_n}{n}. \tag{2.2}$$

In particular, for large n this result is often used to approximate the distribution of $\bar{X}_n$ by $N(\mu, \frac{\sigma^2}{n})$[79]. Also the modeling of extreme events is an inverse problem, and since there is no known closed formulation for the underlying dynamics based on physical laws, statistical modeling is widely accepted. Let us first focus on sample extremes (also called block-extremes) defined by

$$Y = \max(X_1, \ldots, X_n) \text{ or } Y = \min(X_1, \ldots, X_n) \tag{2.3}$$

as $n \to \infty$, where $X_1, \ldots, X_n$ are i.i.d with a common distribution function $F(\cdot)$. Then, the objective of statistical modeling of sample extremes is not only to provide the most descriptive distribution but also to predict likely events that lie beyond the observed range. That is, the considered class of distributions should be valid with respect to more extreme events, e.g., a model for annual maxima should stay valid for five-year maxima. This property is called max-stability and was introduce by Fréchet in 1927. By definition, a distribution $G(y)$ is max-stable if for $n = 1, 2 \ldots$ there exist constants $a_n > 0$ and b_n such that

$$(G(a_n y + b_n))^n = G(y), \tag{2.4}$$

where "stability" indicates that the shape and the function class of the distribution is not changing for "more extreme" extremes[22]. Based on max-stability, Extreme Value Analysis (EVA) is a standard tool in statistics for describing the probability distributions of sample extremes[22,35,40]. Inline with CLT, an asymptotical result was obtained by studying the limiting behavior of sample extremes: for a sequence of real constants $a_n > 0$, b_n and for $n = 1, 2, \ldots$ the nondegenerate limit distribution for normalized extremes is given by

$$\mathbb{P}[\frac{Y - b_n}{a_n} \leq y] = \mathbb{P}[\frac{X_1 - b_n}{a_n} \leq y, \ldots, \frac{X_n - b_n}{a_n} \leq y] \tag{2.5}$$

$$\overbrace{=}^{i.i.d} \mathbb{P}[\frac{X_1 - b_n}{a_n} \leq y] \cdots \mathbb{P}[\frac{X_n - b_n}{a_n} \leq y] \tag{2.6}$$

$$= (F(a_n y + b_n))^n . \tag{2.7}$$

The corresponding limit distributions, defined by

$$(F(a_n y + b_n))^n = G(y), \text{ as } n \to \infty, \tag{2.8}$$

are called the extreme value distributions. The key statements of EVA are described next, for a detailed discussion please compare[22,35,52].

2.1 Univariate Extreme Value Analysis

It has been shown by Fisher & Tippet in 1928 that the limit distribution (2.8) of sample maxima or minima defined by (2.3) is the Generalized Extreme Value (GEV) distribution[35,52], defined by its cumulative distribution function (cdf)

$$G(y; \mu, \sigma, \xi) = \begin{cases} \exp\left(-\left[1 + \xi \frac{y-\mu}{\sigma}\right]^{-\frac{1}{\xi}}\right), & \xi \neq 0, \\ \exp\left(-\exp\left[-\frac{y-\mu}{\sigma}\right]\right), & \xi = 0, \end{cases} \tag{2.9}$$

with location, scale and shape parameters, $\mu, \sigma, \xi \in \mathbb{R}$, respectively, subject to $\left[1 + \xi \frac{y-\mu}{\sigma}\right] > 0$ and $\sigma > 0$. The corresponding GEV probability density function (pdf) that is estimated as the derivative of (2.9): $g(y; \mu, \sigma, \xi) = dG(y; \mu, \sigma, \xi)/dy$, is illustrated in Figure 2.1. GEV is a max-

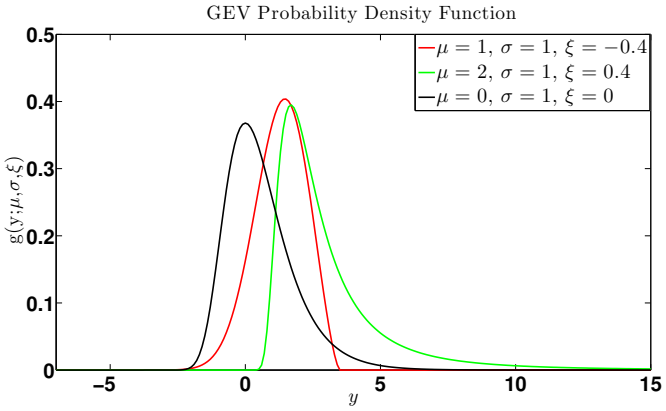


Figure 2.1. *The GEV probability density function $g(y;\mu,\sigma,\xi)$ combines Weibull, Gumbel and Fréchet probability distribution families corresponding to $\xi < 0$, $\xi = 0$ and $\xi > 0$, respectively.*

stable distribution, the reverse is also true: if a distribution is max-stable, then it is a GEV as was shown by Gnedenko in 1943[22,35,52]. Thus, for a series of block-maxima we refer to the GEV as the appropriate distribution. The optimal model parameters can be obtained by maximizing the corresponding log-likelihood function as described in Appendix A.1. Please note that in the following we focus on block-maxima only, analogue results are obtained for block-minima by referring to $Y = \max(-X_1,\ldots,-X_n)$.

However, by considering block-maxima only, we neglect other extreme events. An alternative approach is to define extremes as excesses over/under a predefined higher/lower threshold: considering the above sample of i.i.d variables $X_1,\ldots,X_n$ with $Y \sim G(y;\mu,\sigma,\xi)$ and a large enough threshold u, threshold exceedances are all those variables X_i with $X_i > u$, $i = 1\ldots,m$. Conditioned on $X_i > u$ the threshold excesses are defined as $X_i - u$, $i = 1\ldots,m$. Balkema & de Haan (1974) and Pickands (1975) showed that the distribution of the threshold excesses $\mathbb{P}[X_i - u \leq x | X_i > u]$ can be approximated by the Generalized Pareto Distribution (GPD)[40], defined by its cumulative distribution function

$$H(x;\tilde{\sigma},\xi) = \begin{cases} 1 - \left(1 + \frac{\xi x}{\tilde{\sigma}}\right)^{-\frac{1}{\xi}}, & \xi \neq 0, \\ 1 - \exp\left(-\frac{x}{\tilde{\sigma}}\right), & \xi = 0. \end{cases} \tag{2.10}$$

where $\tilde{\sigma} = \sigma + \xi(u - \mu)$ with ξ, μ, σ defined by (2.9). GPD is parametrized by scale and shape parameters with respect to $\tilde{\sigma} > 0$ and $\left[1 + \xi \frac{x}{\tilde{\sigma}}\right] > 0$. The corresponding GPD pdf $h(x;\tilde{\sigma},\xi) = dH(x;\tilde{\sigma},\xi)/dx$ is illustrated in Figure 2.2. GPD is threshold-stable[22]: threshold excesses for any higher threshold $v > u$ also follow a GPD distribution with the same shape parameter $\xi_v = \xi$ and a shifted scale parameter $\sigma_v = \sigma + \xi(v - u)$. This property is significant for simulating likely threshold excesses that lie beyond the observed range. Please note that the threshold u must be sufficiently high to fulfill the asymptotical behavior. There exist different techniques as how to choose the appropriate threshold[22,96]. In this thesis the threshold will be fixed a priori, for instance, referring to

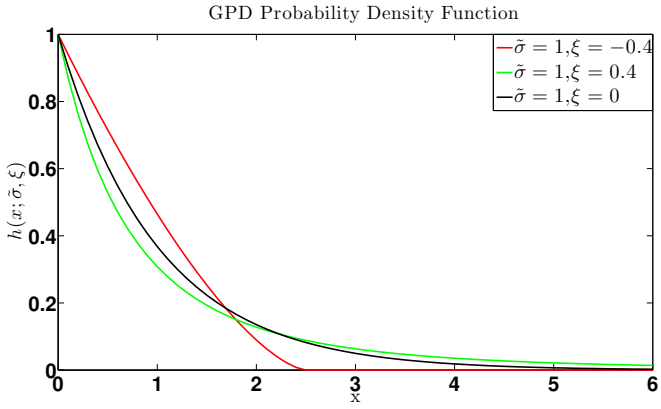


Figure 2.2. *The GPD probability density function. In the case that the shape parameter ξ is zero, the GPD corresponds to the exponential distribution.*

the 0.95 quantile of the observed series.
The optimal model parameters in GEV (2.9) or GPD (2.10) are obtained by maximizing the corresponding log-likelihood function. However, the direct fitting of block-maxima or threshold excesses, respectively, implies independence and identical distribution of the sample $X_1, \ldots, X_n$. This assumption refers to a non-changing behavior of the underlying dynamics (stationarity) and is obviously not always the case, for example, in context of climatology/meteorology the dynamics of monthly temperature or precipitation are affected by the seasonality. The most general way to release this stationarity assumption is to incorporate the time dependent behavior of model parameters, i.e., to express the model parameters in (2.9) and (2.10) by time-dependent functions. Please note that by "model parameters" we refer to both the GEV and the GPD model parameters, unless explicitly specified otherwise.
Further aim is to understand if the behavior of extremes is governed by some external influence. The interest lies in finding the most explanatory variables (denoted as covariates, modes or factors) that significantly influence the parameter dynamics and hence the distribution of extremes. Therefore, in this thesis the model parameters are constructed as functions of covariates, i.e., as regression models. The aim of data-based regression analysis of extreme events will be to infer the values of the regression parameters from observed data. Standard state-of-the-art methods applicable to this task can be roughly divided into two groups: parametric and nonparametric regression approaches. In the following, a review of standard approaches for parametric and nonparametric regression analysis of extremes will be given, methods which are addressed more often in this thesis are explained in more details.

2.1.1 Parametric Regression Analysis

In parametric regression analysis the model parameters are expressed as a priori defined functions depended on covariates. Thereby, as covariates we can involve explicitly known functions

such as sine/cosine functions for modeling the seasonal trends in meteorology, as well as climate phenomena where only measurements are given (in order to study their linkage to the dynamics of extremes). For example, for the location parameter $\mu(t)$ in (2.9) and a set of covariates $U(t) = (u_1(t), \ldots, u_P(t))$, denoted in the following as U_t, we get

$$\mu(t) = f(U_t, \theta), \tag{2.11}$$

where $f(U_t, \theta)$ is the given parametric function, described by the vector of parameters θ. A widely used example is the linear regression, with $f(U_t, \theta)$ defined by

$$f(U_t, \theta) = \mu_0 + \sum_{p=1}^{P} \mu_p u_p(t), \tag{2.12}$$

where the regression coefficients are summarizes in $\theta = (\mu_0, \ldots, \mu_P)$. The main advantages of a linear regression are (a) the simple description of the underlying parameter dynamics and (b) the direct interpretation, since the contribution of each covariate is measured via the corresponding regression coefficient[4,57]. Linear regression might be not appropriate in the case that the influence of the covariates is nonlinear. The linearity can be relaxed by involving nonlinear terms, describing the interacting couplings of the covariates (e.g., $u_1^2(t)$, $u_1(t) \cdot u_2(t)$, $\ldots$) as additional covariates. However, the most general parametric approach is to deploy standard tools from machine learning, e.g., Artificial Neural Networks (ANNs)[9,57] and Support Vector Machines (SVM)[71].
The combination of GEV with a special form of ANN, called Conditional Density Estimation Network (CDN) has led to a creation of the GEV-CDN[14,15], a robust and flexible approach for nonstationary and nonlinear approximation of the GEV model parameters. CDN is an extension of the multilayer perceptron neural network (MLP) for probabilistic models[16,81], where MLP refers to a feed-forward artificial neural network (ANN)[9]. In context of ANN, the covariates are called the input and the model parameters the output. Referring to this notation, the basic idea of ANN applied to regression models is to describe the output by a composition of nonlinear functions dependent on the input. Each nonlinear function is defined by a transformed linear regression

$$h_j(U_t) = m\left(\sum_{p=1}^{P} w_{jp}^{(1)} u_p(t) + w_{j0}^{(1)}\right), \; j = 1, \ldots, M, \tag{2.13}$$

where the superscript (1) refers to the first layer and $w_{j0}^{(1)}$ is the offset. The nonlinear transformation function $m(\cdot)$, e.g., hyperbolic tangent function, and the number M are defined a priori, referring to the activation function and the nodes of a network, respectively. In more complex networks the nonlinear functions can be again constructed by a composition of further functions, referring to a multiple hidden "layer" network[9]. The GEV-CDN approach is based on a simple network, having just three different layers; the input layer, one hidden layer, and the output layer. The output is defined as the composition of $h_j(U_t)$ by

$$o_k(U_t) = \sum_{j=1}^{M} w_{kj}^{(2)} h_j(U_t) + w_{k0}^{(2)}, \tag{2.14}$$

with appropriate weights $w_{kj}^{(2)}$, where the superscript (2) refers to the output layer and k refers to the output dimension, $k = 3$ for the GEV and $k = 2$ for the GPD model. Further, GEV-CDN transforms the output according to

$$\mu(U_t) = o_1(U_t), \tag{2.15}$$

$$\sigma(U_t) = \exp\left(o_2(U_t)\right), \tag{2.16}$$

$$\xi(U_t) = \kappa \tanh\left(o_3(U_t)\right). \tag{2.17}$$

The transformation is chosen such that the scale parameter takes positive values and the shape parameter takes values in $(-\kappa, \kappa)$ only.
The clear advantage of parametric regression analysis is that by inserting the parametric models into (2.9) or (2.10) the optimal model parameters, e.g., linear regression coefficient or the ANN weights, are obtained by maximizing the resulting log-likelihood function. Moreover, the explicit parametrization of the model parameters can be used for simulation and postprocessing, e.g., prediction of the averaged intensity and the frequency of extremes[22]. However, GEV-CDN as well as all other parametric nonstationary extensions of (2.9) and (2.10) rely on the explicit availability of all of the relevant covariates and some strong probabilistic assumptions about the systematically missing/unresolved covariates, e.g., i.i.d. assumption for unresolved covariates. As a result, these methods implicitly assume time independency of involved offset, e.g., in (2.14) the offset $w_{k0}^{(2)}$.

2.1.2 Nonparametric Regression Analysis

Parametric regression analysis makes explicit assumptions about the dynamics of the model parameters. However, in many real applications the underlying dynamics is unknown and the a priori assumptions could imply biased results. Moreover, the aim of regression analysis is not only to identify the significant set of covariates, but also to investigate their kind of influence on parameter dynamics. A parametric approach might be inappropriate, and tools for describing the parameter dynamics beyond a priori assumptions are required. The appropriate techniques are called nonparametric[79].
Well-known examples of nonparametric regression analysis are the Local Likelihood Smoothing[32] and nonparametric Bayesian techniques[24,82]. Local Smoothing based techniques provide at each time step a local estimator by considering only weighted values in the direct neighborhood (window) of a predetermined size. The weights are assigned according to a priori chosen kernel function. There is no closed formulation for the nonstationary parameter dynamics but rather a sequence: a value for each parameter for each time step. In addition, the results depend strongly on the choice of the kernel function and the size of the window. Classical Bayesian techniques rely on a priori assumption about the distribution of the model parameter θ, the so called prior is denoted by $\pi(\theta)$. This a priori information is incorporated into the parameter estimation by applying the Bayes' Theorem resulting in the posterior distribution for the model parameters. Please see Appendix A.2 for more information. Often there is no closed formulation for the posterior and sampling techniques such as Markov Chain Monte Carlo methods (MCMC)[3] are used to obtain the main characteristics of the posterior. The results of Bayesian statistics rely strongly on the choice of the prior.

Another strategy is to involve mixture models and Hidden Markov Models (HMM)[7,10,74]. Such approaches require a priori knowledge about the parametric probabilistic model class for the time-dependent model parameters as well as stationarity and Markov assumptions for the hidden parameter-switching process.
Besides, for nonparametric nonlinear regression Generalized Additive Models (GAM) can be applied[56,57,110]. This class of models provides a general approach to approximate the underlying dynamics and to account for nonstationarity by a linear combination of smooth nonparametric functions of covariates. The GAM approach is widely used to study the dynamics of extremes in context of EVA[18,86,114]. In the following we briefly outline the main idea of GAM. Let Y_t be the output and denote the covariates $U_t = (u_1(t), \ldots, u_P(t))$ as the input, each measured at time steps $t = t_1, \ldots, t_{N_T}$. Then, an additive regression model has the following structure[57]

$$Y_t = f(U_t) + \varepsilon, \quad \text{with} \quad f(U_t) = \sum_{p=1}^{P} f_p\left(u_p(t)\right), \tag{2.18}$$

where ε is the error term with zero expectation and $f_p(.) \in W^2([t_1, t_{N_T}])$, $p = 1, \ldots, P$, is an arbitrary, nonlinear and nonparametric smooth function with

$$W^2([t_1, t_{N_T}]) = \{f_p(.) : f_p(.) \in C^{(2)}([t_1, t_{N_T}]), \int_{t_1}^{t_{N_T}} [f_p''(\tilde{t})]^2 d\tilde{t} < \infty\}. \tag{2.19}$$

For simplicity we will assume that $P = 1$, then the optimal smoothing functions for the additive model in (2.18) is obtained by minimizing[56]

$$\sum_{j=1}^{N_T} \left(Y_{t_j} - f(U_{t_j})\right)^2 + \lambda \int_{t_1}^{t_{N_T}} [f''(\tilde{t})]^2 d\tilde{t}. \tag{2.20}$$

Thereby, the first term minimizes the distance between the observations and the function $f(U_t)$, while the second, the regularization term, penalizes the curvature of $f(U_t)$. The optimal solution is a natural cubic spline[56,110,111] with

$$f(U_t) = \sum_{i=1}^{q_j} \beta_{ji} b_{ji}(U_t), \tag{2.21}$$

where $b_{ji}(.)$ are the basis functions e.g., cubic or thin splines, and β_{ji} are the corresponding coefficients. Parameter q_j should be chosen large enough in order to fit the underlying dynamics. The solution depends on the degree of penalization; for $\lambda > 0$ it refers to a smoothing rather then to an interpolating function, i.e., Y_t and $f(U_t)$ do not necessarily match. Correspondingly, for $\lambda \to \infty$ the solution is a linear function, since we get $f''(.) = 0$, and for $\lambda \to 0$ we are searching for an interpolating function $f(U_t) \in W^2$. Consequently, the smoothing spline is "approximately a kernel smoother"[56]. The optimal choice of λ depends on underlying data and can be chosen with respect to information criteria, as will be discussed in Section 4.1.

Minimization of (2.20) for $P \geq 1$ is carried out by the iterative "backfitting" Algorithm 1 as described in[56,111]: in an alternating order, the smoothing function of each covariate $f_p(u_p(t))$ is fitted to the "partial residual". Thereby, the "partial residual" is the difference between the data and the current GAM model that does not contain the estimate of $f_p(u_p(t))$[111]. The algorithm stops when the changes of the estimates are negligible. The generalized additive regression models can be

Algorithm 1: The Backfitting Algorithm for Generalized Additive Models[56,111]

Initialize: $\hat{f}_p^{(old)}(u_p(t))$ for $p = 1, \ldots, P$;

while $|f_p^{(new)}(u_p(t)) - f_p^{(old)}(u_p(t)))| > eps$ **do**

  for *j=1:P* **do**

1 $partialResidual = Y - \sum_{p=1}^{P} \hat{f}_p^{(old)}(u_p(t))$

2 Fit $\hat{f}_p^{(new)}(u_p(t))$ to the *partialResidual* applying e.g., the gradient based Newton-Raphson, or any other appropriate method[56,111].

directly applied to study the nonlinear and nonstationary impact of covariates on the GEV/GPD model parameters, as exemplified on the scale parameter

$$\sigma(U_t) = \exp\left(\sum_{p=1}^{P} f_p\left(u_p(t)\right)\right). \tag{2.22}$$

The exponential transformation ensures that $\sigma(t) > 0$ for all time steps. For more details of GAM models in context of extreme events we refer to [18,86,114].

2.2 Spatial Extreme Value Analysis

In this section, we review the standard approaches for statistical modeling of spatial extreme events. We focus only on one quantity, for example, extreme precipitation measured at different locations. The objective in spatial modeling of extremes is (a) the detection of spatio-temporal changes in frequency and intensity and (b) the dependency structure of extremes. The dependency structure, which describes the relationship between different locations with respect to the occurrence of extremes measured, for instance, by the joint probability, can depend on direction and/or time, referring to an anisotropic and a nonstationary behavior, respectively. While the spatio-temporal changes within the observed region can be studied exploiting the univariate EVA (referring to marginal distributions), the spatial dependence structure is approached with the application of spatial statistics. The classical spatial statistics (geostatistics) is not appropriate for spatial analysis of extreme events because it handles only the mean behavior and does not capture the dynamics of extremes, referring to the tails of the distribution. The state-of-the-art methods for statistical modeling of spatial extremes comprise of extreme value theory and geostatistics and are classified into max-stable spatial

processes, copula based approaches, and spatial hierarchical models[25,31,62]. In the following, we briefly discuss the latest developments in the statistical modeling of spatial extreme events. For a more comprehensive overview please see[6,25,31,62,92].

2.2.1 Max-stable Processes

Let $Y(s,t)$ be a series of extremes observed at location s and time t, for $s = s_1,\dots,s_{N_S}$ and $t = t_1,\dots,t_{N_T}$, e.g., maximal annual temperature at N_S locations observed for N_T years. Referring to block-maxima, the spatial extension of GEV provides a class of max-stable processes[6,34,68]. The resulting max-stable models are not only verified for simulating the joint behavior of spatial block-maxima, but also of the spatial exceedances over a higher threshold[5,62]. Thereby, the standard approach is to first estimate the marginals $F_s(\cdot)$ by fitting either the GEV or the GPD distribution to each location s and second, to transform the marginals to a common distribution. For the first step often the unit Fréchet distribution is chosen

$$\mathbb{P}[Z \leq z] = \exp\left(-\frac{1}{z}\right). \tag{2.23}$$

The observed sample of extremes has to be standardized with respect to

$$Z(s,t) = -\frac{1}{\log\left(F_s\left(Y(s,t)\right)\right)}, \quad s = s_1\dots,s_{N_S}, \quad t = t_1,\dots,t_{N_T}. \tag{2.24}$$

Then, the dependence structure of the spatial extremes across different locations is modeled by fitting the series of the standardized max-stable random vectors $Z(s,t) = (Z(s_1,t),\dots,Z(s_{N_S},t))$ for $t = t_1,\dots,t_{N_T}$ to the chosen max-stable process. A general formulation of a spatial max-stable process is given by

$$\mathbf{Z}(\mathbf{s},t) = \sup_{i\geq 1} \chi_i W_i(\mathbf{s},t), \text{ with } \mathbf{s} = (s_1,\dots,s_{N_S}), \text{ and fixed } t \tag{2.25}$$

where $0 < \chi_1 < \chi_2 < \dots$ are points from a Poisson process with intensity $\frac{1}{\chi_i^2} d\chi_i$ and $W_i(\mathbf{s},t)$ are independent samples from any stochastic spatial process $W(\mathbf{s},t)$ with $\mathbb{E}[W(\mathbf{s},t)] = 1$[34]. One possible interpretation of (2.25), given by Smith, is the "rainfall-storm" model: Denote $W_i(\mathbf{s},t)$ as the shape of a random storm in space and χ_i as its intensity, then $\mathbf{Z}(\mathbf{s},t)$ defines the point-wise maximal rainfall over all storms for each location. The joint distribution of $\mathbf{Z}(\mathbf{s},t)$ with $z(s_1,t),\dots,z(s_{N_S},t) > 0$ is described by

$$\mathbb{P}[Z(s_1,t) \leq z(s_1,t),\dots,Z(s_{N_S},t) \leq z(s_{N_S},t)] = \exp\left(-\mathbb{E}[\max_{j=1,\dots,N_S} \frac{W(s_j,t)}{z(s_j,t)}]\right). \tag{2.26}$$

There is no finite parametrization of $\mathbf{Z}(\mathbf{s},t)$, since for any appropriate random process $W(\mathbf{s},t)$ we obtain a closed formulation of the max-stable process by evaluating the right hand side of (2.26). The most frequently used parametric models for $W(\mathbf{s},t)$ are based on Gaussian distribution including the resulting Smith, Schlather and Brown-Resnick processes[63,97]. In order to model anisotropic

and/or nonstationary dependence structures of extremes, the process $W(\mathbf{s},t)$ incorporates parametric spatio-temporal covariance functions[31,62]. However, the fitting of a max-stable process to the data is in general not applicable for $N_S > 3$. The two main reasons are: (a) closed formulation of the involved multidimensional integral in (2.26) exists in general for low dimensions ($N_S \leq 2$) only and (b) since the full likelihood is formed by differentiation of (2.26) with respect to $z(s_1,t),\ldots,z(s_{N_S},t)$, the number of terms involved in the log-likelihood explodes combinatorially with N_S[62]. Hence, the standard likelihood-based inference for such models is computationally infeasible for most real applications, e.g., in climate and weather research[25,30,31]. Instead, an approximation of the log-likelihood, the so-called composite log-likelihood, is considered[85,109]. The composite log-likelihood is the sum over the possible pairwise distributions, corresponding to the bivariate density $f(Z(s_i,t),Z(s_k,t),\Theta)$ for any s_i,s_k. Assuming their existence and eligibility for model parameter inference, i.e., the log-likelihood can be estimated, the composite log-likelihood is given by

$$\mathcal{L}_{\mathcal{C}}\left(\mathbf{Z}(\mathbf{s},t),\Theta\right) = \sum_{j\neq k\in\mathcal{K}}\sum_{i=1}^{n}\log f\left(Z(s_i,t),Z(s_k,t),\Theta\right), \tag{2.27}$$

where $\mathcal{K}$ is the set of all pairs[85]. The composite log-likelihood based inference can be successfully applied to spatial extremes[30,62,104].

Another possibility to model the dependence structure of multivariate extremes is the application of extremal copulas[92]: for the random max-stable vector $\mathbf{Z}(\mathbf{s},t)$ with univariate max-stable marginals $F_1,\ldots,F_{N_S}$ and joint distribution $F\left(Z(s_1,t),\ldots,Z(s_{N_S},t)\right)$ an extremal copula $C:[0,1]^{N_S}\rightarrow[0,1]$ characterizes the dependence structure of $\mathbf{Z}(\mathbf{s},t)$ by

$$F\left(Z(s_1,t),\ldots,Z(s_{N_S},t)\right) = C\left(F_1(Z(s_1,t)),\ldots,F_{N_S}(Z(s_{N_S},t))\right), \tag{2.28}$$

subject to $C\left(u_1^m,\ldots,u_{N_S}^m\right) = C\left(u_1,\ldots,u_{N_S}\right)^m$, $0<u_1,\ldots,u_{N_S}$, and with the exponent $m\in\mathbb{N}$[80]. The resulting joint distribution is max-stable. The extremal copula belongs to the class of max-stable processes and sharing the same difficulties when applied to real data[25,30,31].

2.2.2 Bayesian Hierarchical Models

Another often used parametric approach for statistical modeling of spatial extremes is based on Bayesian Hierarchical Models (BHM). Thereby, by exploiting the probability chain rule and the Bayesian Theorem, hierarchical models describe the dynamics of a complex system by decomposing it into different layers of parameter variation[27], compare Appendix A.2. The two main layers are usually the data model, which describes the distribution of the data given a process, and the process model, which describes the true underlying process. The final layer defines the prior distribution for parameters of the process model[27,29]

$$\begin{aligned}\pi\left(\text{process, parameters}|\text{data}\right) \propto\ & \pi\left(\text{data}|\text{process, parameters}\right)\\ & \times\pi\left(\text{process}|\text{parameters}\right)\\ & \times\pi\left(\text{parameters}\right).\end{aligned} \tag{2.29}$$

The application of BHM for modeling spatial extreme events requires the parametric specification of the involved layers. For a given process and the parameter layers it is often assumed that the extremes among different locations are independent, referring to conditional independence. As a consequence, the data model, which defines the joint likelihood, is the product of the likelihood functions for each location[25,31]. The process model refers to either the GEV or to the GPD distribution. The parameter model reflects the spatial variations. For example, we can express the location parameter in GEV as a function of space and time: $\mu(s,t) = f(s,t;\beta) + S(s,\psi)$, where $f(s,t;\beta)$ is a deterministic function dependent on covariates, and $S(x,\psi)$ is a stationary Gaussian process with predefined covariance ψ, describing the linear relationship between locations[79].
The above BHM formulation provides a flexible tool for modeling spatial variation in marginal distributions of extremes and can deal with large problems. However, it can not capture the underlying spatial dependence structure due to the conditional independence assumption in the data model layer. In order to account for the dependence structure as well, more sophisticated formulations of the data model layer should be approached, based, e.g., on Gaussian copula models[95]. Further, classical hierarchical models do not fulfill the max-stability.
A max-stable extension of hierarchical models was recently presented[89]: The main idea is first to fit the marginal GEV/GPD distributions to spatial block-maxima or threshold excesses, respectively, and then to approximate the residual dependence structure by a combination of kernel basis functions. Because this thesis deploys max-stable hierarchical models, this idea will be discussed in more detail: Under the assumption that $Y(s,t)$ is max-stable, the corresponding marginal distributions are GEV for each single location. Equivalently, the residual process estimated by

$$X(s,t) = \left(1 + \xi(s)\frac{Y(s,t) - \mu(s)}{\sigma(s)}\right)^{\frac{1}{\xi(s)}} \tag{2.30}$$

is max-stable and has the univariate Fréchet distributions (2.23) as marginals[90]. The dependence structure of $X(s,t)$ called the residual spatial dependence structure is denoted by $\tilde{\theta}(s,t)$, and is approximated by a linear combination of positive kernel basis functions $w_l(s)$, e.g., Gaussian kernels, such that we obtain

$$\tilde{\theta}(s) = [\sum_{l=1}^{L} A_l(t) w_l(s)^{\frac{1}{\alpha}}]^{\alpha}, \text{ with } \sum_{l=1}^{L} w_l(s) = 1, \text{ and } \alpha \in (0,1). \tag{2.31}$$

The parameter α controls the smoothness of the process; for $\alpha \approx 0$ the resulting process is smooth, while $\alpha \approx 0$ refers to a noisy process[89]. In order to satisfy max-stability of the spatial process and to provide Fréchet marginal distributions, the approximation coefficients $A_l(t)$ must follow a positive stable distribution, which has a Laplace transformation of the particular form

$$\int_0^{\infty} \exp(-At) p(A|\alpha) dA = \exp(-t^{\alpha}). \tag{2.32}$$

In the following, the appropriate distribution for $A_l(t)$ is denoted by $PS(\alpha)$. Given the coefficients $A_1(t),\ldots,A_L(t)$ the observed extremes are conditionally independently distributed, i.e.,

$$\mathbb{P}[Y(s_1,t),\ldots,Y(s_S,t)|A_1(t),\ldots,A_L(t)] = \mathbb{P}[Y(s_1,t)]\ldots\mathbb{P}[Y(s_S,t)]. \tag{2.33}$$

In line with the notation in[89] this can be rewritten as

$$Y(s,t)|A_1(t),\ldots,A_L(t) \overset{independent}{\sim} GEV\left(\mu^*(s,t),\sigma^*(s,t),\xi^*(s)\right), \tag{2.34}$$

$$A_l(t) \overset{i.i.d}{\sim} PS(\alpha) \tag{2.35}$$

where the GEV parameters have the following formulation

$$\mu^*(s,t) = \mu(s) + \frac{\sigma(s)}{\xi(s)}[\theta(s,t)^{\xi(s)} - 1], \quad \sigma^* = \alpha\sigma(s)\theta(s,t)^{\xi}(s), \quad \xi(s) = \alpha\xi(s). \tag{2.36}$$

For the final layer it is assumed that $\mu(s), \sigma(s), \xi(s)$ are Gaussian distributed. Thereby, the mean value can be represented by a regression model in order to describe the parameter dynamics dependent on some covariates. The resulting model parameters are obtained exploiting MCMC sample techniques and the joint distribution of the residual spatial process is given by[89]

$$\mathbb{P}[X(s_1) < c_1,\ldots,X(s_n) < c_n] = \exp\{-\sum_{l=1}^{L}[\sum_{i=1}^{n}\left(\frac{w_l(s_i)}{c_i}\right)^{\frac{1}{\alpha}}]^{\alpha}\}. \tag{2.37}$$

The max-stable hierarchical model provides a flexible approach for spatio-temporal modeling of extremes which enables also a nonstationary description of the dependence structure. However, the resulting description of the underlying dynamics of extremes depends on the choice of the kernels and the a priori probabilistic assumption of the approximation coefficients and the marginal model parameters $\mu(s), \sigma(s), \xi(s)$. Further, models based on hierarchical formulation may suffer from the assumption of conditional independence. It was shown in[51] that such models can not simulate dependence of very rare events at two different locations regardless of the distance between them.

2.3 Conclusion

The main aim of the spatio-temporal modeling of extreme events is to study (a) the spatio-temporal variability of the model parameters, referring to the marginal distribution and thus to the univariate extreme value analysis, and (b) the spatial dependence structure. On one hand, the univariate extreme value analysis is well investigated and established, often times deploying the statistical regression analysis where the choice of the covariates depends on a priori knowledge about the investigated system. On the other hand, modeling of the dependence structure in extremes is an active research field. Most of the approaches in this area (as explained above) are based on some parametric a priori assumptions. For example, Smith and Schlather max-stable processes assume Gaussian behavior of the underlying spatial process, whereas Bayesian hierarchical models are based on conditional independence and a priori assumptions about the distribution of the model parameters[25,31]. Mixture models, e.g., the max-stable hierarchical model[89] and the Dirichlet-based copula[49], provide nonparametric approaches. Thereby, the dependence structure is approximated by a linear combination of a priori defined kernel functions and a priori parametric assumptions on the distribution of the approximation coefficients are made[89]. Further, the maximization of the corresponding spatio-temporal likelihood function is computationally infeasible in real-world application

and so the inference is based on approximations, e.g., considering only the pair-wise interactions between the different spatial locations, bivariate densities are deployed for the composite likelihood function[91,109].

In order to approach some of the above problems in spatio-temporal modeling of extremes, a non-stationary and semiparametric framework for spatio-temporal statistical regression analysis of extremes will be introduced. Special emphasis will be done on the issues of systematically missing covariates, numerical instability and computational efficiency.

3 FEM-BV-EVA Methodology

In this chapter, we present a methodology for the nonstationary spatio-temporal regression analysis of extremes that accounts for the issues of systematically missing covariates, ill-posedness, and numerical complexity. The methodology investigates the main objectives of data-based extreme value analysis, namely, (a) the changes in marginal distributions and (b) the description of the underlying dependence structure among the observed locations. For the investigation of marginal distributions, we exploit univariate EVA and express the model parameters in GEV and GPD distributions as linear combinations of covariates. We will show that the involved regression model becomes nonstationary if some of the relevant covariates are systematically missing. The resulting nonstationarity and the ill-posedness of the inverse problem are resolved by deploying the recently introduced Finite Element Time Series Analysis Methodology with Bounded Variation of model parameters (FEM-BV)[58–60,78]. It will be demonstrated that the proposed FEM-BV-EVA approach allows a well-posed problem formulation and goes beyond probabilistic a priori assumptions of methods for analysis of extremes based on, e.g., nonstationary Bayesian mixture models, smoothing kernel methods or neural networks. Based on max-stable hierarchical models and advanced time-series techniques, spatial FEM-BV-EVA extension provides a pragmatic, nonparametric, and nonstationary description of the underlying spatial dependence structure.

3.1 Data-based regression analysis of univariate extremes

This section addresses the issue of unresolved covariates in statistical univariate regression analysis of extremes based on our results proposed in[64,65]. Here we will focus on univariate data-based regression analysis of block-maxima first.

3.1.1 Regression analysis of block-maxima

In order to account for the nonstationarity of the underlying dynamics of extremes, we focus on the fully time-dependent GEV distribution defined by its probability density function (pdf)

$$f\left(x;\mu\left(t\right),\sigma\left(t\right),\xi\left(t\right)\right)=\begin{cases} c(t)\exp\left(-[1+\xi\left(t\right)\frac{(x-\mu(t))}{\sigma(t)}]^{-\frac{1}{\xi(t)}}\right) & ,\xi(t)\neq 0 \\ c(t)\exp\left(-\exp\{-\frac{x-\mu(t)}{\sigma(t)}\}\right) & ,\xi(t)=0, \end{cases} \tag{3.1}$$

where t denotes the time variable and $c(t)$ the normalization constant where

$$c(t) = \begin{cases} \frac{1}{\sigma(t)}[1+\xi\,(t)\,\frac{(x-\mu(t))}{\sigma(t)}]^{-\frac{1}{\xi(t)}-1} & ,\xi(t)\neq 0 \\ \frac{1}{\sigma(t)}\exp\left(-[\frac{(x-\mu(t))}{\sigma(t)}]\right) & ,\xi(t)=0. \end{cases} \tag{3.2}$$

The model parameters must fulfill the constraints:

$$[1+\xi(t)\frac{(x-\mu(t))}{\sigma(t)}]>0, \quad \text{and} \quad \sigma(t)>0 \quad \forall t. \tag{3.3}$$

The aim is to investigate if the parameter dynamics is influenced by some external factors. For this purpose we express each GEV parameter as a function dependent on a vector of covariates. Let us denote all covariates with significant influence on the parameter dynamics by the vector $U^{all}(t) = \left(u_1^{all}(t),\ldots,u_{\mathcal{J}}^{all}(t)\right)$. Under the assumption that $U^{all}(t)$ is known, we focus on linear regression as exemplified on the location parameter,

$$\mu\left(U^{all}(t)\right) = \mu_0 + \sum_{j=1}^{\mathcal{J}} \mu_j u_j^{all}(t), \tag{3.4}$$

where μ_j, $j=1\ldots,\mathcal{J}$, are the regression coefficients. However, in real applications, one is usually confronted with the problem that some (or most) of potentially relevant covariates are missing in the measurements. One possible source for the systematically missing covariates is the multi-scale dynamics nature of the underlying process, e.g., processes in climate or molecular dynamics may involve multiple time and length scales[20,75,77]. That is, by only observing covariates on a slow time scale (resolved covariates) we neglect covariates on the faster scale (unresolved covariates). An additional reason for the missing covariates is that, even on just one single time scale we can not resolve all covariates due to the interest in regression models with finite numbers of degrees of freedom. In particular, this is true for regression analysis of extremes because of the relatively small statistics. We have to select a set of resolved covariates, usually based on expert knowledge, and have to account for the effect of the systematically unresolved/missing covariates. Several disciplines cover the issue of missing information, e.g., in statistical regression analysis the issue of unresolved information is often addressed as the "unobserved heterogeneity"[13]. The unobserved covariates are included in the regression model via a stationary probabilistic error term. The posterior distribution of extremes is obtained by exploiting the Bayesian inference as discussed in Appendix A.2 and depends on the a priori assumption about the distribution of this error term. Furthermore, there is often no closed expression of the posterior. In this thesis we reduce the involved linear regression model to resolved covariates only, and express the influence coming from unresolved covariates by a nonstationary additive offset. Considering the involved covariates $U^{all}(t)$, we now split them into resolved $U_t = (u_1(t),\ldots,u_P(t))$ and unresolved $U_t^{un} = \left(u_1^{un}(t),\ldots,u_Q^{un}(t)\right)$ covariates. Without loss of consistency, we normalize the unresolved covariates and rewrite (3.4)

$$\mu\left(U_t, U_t^{un}\right) = \mu_0 + \sum_{p=1}^{P} \mu_p u_p(t) + \frac{1}{Q}\sum_{q=1}^{Q} \nu_q u_q^{un}(t), \tag{3.5}$$

where μ_p, $p = 1\ldots,P$ and ν_q, $q = 1\ldots,Q$ are the regression coefficients. Based on the assumption that $\nu_q u_q^{un}(t)$ are i.i.d for all time steps, we can apply the Central Limit Theorem to reduce their influence approximately to an additive Gaussian noise, i.e., the reduced formulation of (3.5) is given by

$$\mu\left(U_t\right) = \mu_0 + \sum_{p=1}^{P} \mu_p u_p(t) + \varepsilon(t), \text{ with } \varepsilon \sim \mathcal{N}\left(\hat{\mu}, \hat{\sigma}(t)\right). \tag{3.6}$$

In real applications where the i.i.d assumption may be too strong, we adjust by applying the Central Limit Theorem for independent variables in a formulation that requires a much weaker Lindeberg condition[72]. This condition reveals that the contribution of a random standardized variable to a sum is small relative to the total sum[94]. Then, under the assumption that the Lindeberg condition holds we rewrite (3.5)

$$\begin{aligned}\mu\left(U_t, U_t^{un}\right) = \mu_0 + \sum_{p=1}^{P} \mu_p u_p(t) \\ + \underbrace{\frac{1}{Q}\sum_{q=1}^{Q} \nu_q \left(u_q^{un}(t) - \mathbb{E}[u_q^{un}(t)]\right)}_{\to \varepsilon(t)} + \frac{1}{Q}\sum_{q=1}^{Q} \nu_q \mathbb{E}[u_q^{un}(t)].\end{aligned} \tag{3.7}$$

For cases in which the covariates are not independent, the Karhunene-Loève transformation can be used to provide a orthogonal representation of the covariates, i.e., to decorrelate them[69,73]. By inserting $\mu_0(t) = \mu_0 + \frac{1}{Q}\sum_{q=1}^{Q} \nu_q \mathbb{E}[u_q^{un}(t)]$ into (3.7) we obtain the reduced, nonstationary regression model

$$\mu\left(t, U_t\right) = \mu_0(t) + \sum_{p=1}^{P} \mu_p u_p(t) + \varepsilon(t), \text{ with } \varepsilon \sim \mathcal{N}\left(0, \hat{\sigma}\right). \tag{3.8}$$

Please note that in (3.8) the offset $\mu_0(t)$ is a time-dependent function and not a constant number as in the case of parametric statistics (3.6). In presence of systematically missing covariates, the application of parametric approaches could lead to biased results, then requiring nonparametric statistical methods. Further, we generalize (3.8) by releasing the stationarity assumption of the coefficients μ_p for $p = 1,\ldots,P$ and get

$$\mu\left(t, U_t\right) = \mu_0(t) + \sum_{p=1}^{P} \mu_p(t) u_p(t) + \varepsilon(t). \tag{3.9}$$

This generalization is significant, especially in situations when we observe a long time period the influence of covariates can change with time. Analogous to (3.9), we express the scale and shape parameters by focusing on the same set of covariates as for the location parameter, such that we

obtain

$$\sigma(t,U_t) = \sigma_0(t) + \sum_{p=1}^{P} \sigma_p(t) u_p(t) + \tilde{\varepsilon}(t), \quad \tilde{\varepsilon}(t) \sim \mathcal{N}(0, \breve{\sigma}(t)), \tag{3.10}$$

$$\xi(t,U_t) = \xi_0(t) + \sum_{p=1}^{P} \xi_p(t) u_p(t) + \overline{\varepsilon}(t), \quad \overline{\varepsilon}(t) \sim \mathcal{N}(0, \overline{\sigma}(t)). \tag{3.11}$$

The regression models in (3.9-3.10-3.11), reduced to resolved covariates only, become stochastic due to their normally-distributed additive noise terms. The normal additive noise corresponds to a prior in Bayesian inference context (compare Appendix A.2) and there exists no closed formulation for the resulting posterior[22]. However, in the interest of simplicity, the current manuscript we focus on the mean behavior of parameters. Consequently, we neglect the normally-distributed noise terms in (3.9-3.10-3.11). Please note that by considering the mean behavior we obtain deterministic model parameters, which still account for the unresolved information through the nonstationary off-set terms $\mu_0(t), \sigma_0(t), \xi_0(t)$. The consideration for complete stochastic regression model with explicit noise terms remains for future study. Finally, the nonstationary GEV distribution (3.1) is parametrized by

$$\Theta_{GEV}(t) = (\mu_0(t), \ldots, \mu_P(t), \sigma_0(t), \ldots, \sigma_P(t), \xi_0(t), \ldots, \xi_P(t)). \tag{3.12}$$

The reduced regression model (3.8) is a general example of how to reflect the unresolved covariates in context of linear regression. In the following section, we will directly apply this model to the regression analysis of threshold excesses as summarized in the next section.

3.1.2 Regression analysis of threshold excesses

In the following we aim to model the behavior of threshold excesses as defined in Section 2.1 by a fully nonstationary GPD described by its pdf

$$f(x; \tilde{\sigma}(t), \xi(t)) = \begin{cases} \frac{1}{\tilde{\sigma}(t)} \left(1 + \frac{\xi(t)x}{\tilde{\sigma}(t)}\right)^{-\frac{1}{\xi(t)}-1} & , \xi(t) \neq 0, \\ \frac{1}{\tilde{\sigma}(t)} \exp\left(-\frac{x}{\tilde{\sigma}(t)}\right) & , \xi(t) = 0, \end{cases} \tag{3.13}$$

where t denotes the time variable and the following constraints are satisfied

$$\left[1 + \frac{\xi(t)x}{\tilde{\sigma}(t)}\right] > 0 \quad \text{and} \quad \tilde{\sigma}(t) > 0 \quad \forall t, \quad x > 0. \tag{3.14}$$

The extension to statistical regression analysis of threshold excesses based on resolved covariates is straightforward. Under the assumption that the Lindeberg condition holds, we summarize the reduced regression behavior of the GPD model parameters in line with (3.9-3.10-3.11) and obtain

$$\tilde{\sigma}(t,U_t) = \tilde{\sigma}_0(t) + \sum_{p=1}^{P} \tilde{\sigma}_p(t) u_p(t) + \breve{\varepsilon}(t), \quad \breve{\varepsilon}(t) \sim \mathcal{N}(0, \breve{\sigma}(t)), \tag{3.15}$$

$$\xi(t,U_t) = \xi_0(t) + \sum_{p=1}^{P} \xi_p(t) u_p(t) + \breve{\varepsilon}(t), \quad \breve{\varepsilon}(t) \sim \mathcal{N}(0, \breve{\sigma}(t)). \tag{3.16}$$

The deterministic formulation of the resulting model parameter is given by

$$\Theta_{GPD}(t) = (\tilde{\sigma}_0(t), \ldots, \tilde{\sigma}_P(t), \xi_0(t), \ldots, \xi_P(t)), \tag{3.17}$$

and where the offsets $\tilde{\sigma}_0(t)$ and $\xi_0(t)$ reflect the nonstationary impact from unresolved covariates.

3.1.3 Univariate FEM-BV-EVA

Given the series of extreme events, X_t, observed at time steps $t = t_1, \ldots, t_{N_T}$ and referring either to block-maxima or to threshold excesses, we aim to estimate the most descriptive model parameters $\Theta_{GEV}(t)$ or $\Theta_{GPD}(t)$. In the following, $\Theta(t)$ will refer to both the GEV and the GPD model parameters, unless noted otherwise. To get the optimal parameters, we minimize the corresponding negative log-likelihood function (NLL)

$$\mathcal{L}(X_t, \Theta(t)) = \sum_{j=1}^{N_T} g\left(X_{t_j}, \Theta(t_j)\right) \tag{3.18}$$

with respect to $\Theta(t)$, where $g(X_t, \Theta(t))$ refers to the NLL for a fixed t. Before doing so, we must parametrize the model parameter $\Theta(t)$. As discussed in 3.1.1, the application of a parametric approach implies a constant offset and could produce biased results. Nonparametric statistical methods are more appropriate in these situations. Exploiting smoothing regression[56,110,111] the nonstationarity in $\Theta(t)$ can be resolved by a smoothing spline according to (2.21). For example, we can resolve the nonstationary offset term $\tilde{\sigma}_0(t)$ in (3.15) by

$$\tilde{\sigma}_0(t) = \sum_{i=1}^{q} \beta_i b_i(t), \tag{3.19}$$

where $b_i(t)$ represents the basis function e.g., cubic or thin spline, β_i is the corresponding coefficient and q is the dimension of the basis functions. Estimation of the optimal coefficients β_i involves penalizing of the "wiggliness" of $\tilde{\sigma}_0(t)$. However, as already outlined in Section 2.1.2, the smoothness of the spline implies the locality property of a kernel smoother and becomes a drawback in the case that the underlying function is discontinuous.

In this thesis the resulting nonstationarity is handled by applying the Finite Element time series analysis methodology (FEM)[58–60]. FEM formulates the inverse problem for a nonstationary dynamical systems as a regularized variational clustering problem. The nonstationarity is approximated by a set of locally stationary models and a nonstationary convex switching process. To ensure well-posedness of the inverse problem the switching process discretized with Finite Elements is restricted, for instance, to the class of functions with bounded variation. For a detailed introduction into the FEM approach, we refer the interested reader to[58–60,78]. In the following we will formulate the FEM-BV-EVA approach in two steps (i) approximation and (ii) regularization.

The FEM approach assumes that the model parameter $\Theta(t)$ changes slower than the observed series of extremes. Then, the underlying dynamics can be approximated by a set of $K \geq 1$ locally stationary models, each parametrized by θ_k for $k = 1, \ldots, K$ and a nonstationary switching process

$\Gamma(t) = (\gamma_1(t), \ldots, \gamma_K(t))$. FEM interpolates the nonstationary model "distance" function $g(X_t, \Theta(t))$ by a linear convex combination of K locally stationary model distance functions

$$g(X_t, \Theta(t)) = \sum_{k=1}^{K} \gamma_k(t) g_k(X_t, \theta_k) \tag{3.20}$$

with corresponding constraints on Θ and convexity constraints on $\Gamma(t)$

$$\sum_{k=1}^{K} \gamma_k(t) = 1, \qquad t = t_1, \ldots, t_{N_T}, \tag{3.21}$$

$$\gamma_k(t) \geq 0, \qquad t = t_1, \ldots, t_{N_T}, \qquad k = 1, \ldots, K. \tag{3.22}$$

By inserting (3.20) into (3.18) we obtain the average (interpolated) model distance functional

$$\mathcal{L}(\Theta, \Gamma(t)) = \sum_{j=1}^{N_T} \sum_{k=1}^{K} \gamma_k(t_j)\, g\left(X_{t_j}, \theta_k\right) \tag{3.23}$$

In contrast to mixture models and HMMs[7,10,74], FEM avoids a priori assumptions on $\Gamma(t)$ like stationarity, Gaussian or Markovian behavior, but deploys a nonparametric and nonstationary hidden switching process. Elimination of a priori assumptions implies ill-posedness of the optimization problem in sense of Hadamard[54]. FEM regularizes the ill-posed problem by exploiting the observation that many realistic problems demonstrate a persistent (metastable or regime-switching) behavior for their parameters. In the general FEM formulation, the persistence property/condition on $\Gamma(t)$ can be imposed by referring to the space of weakly differentiable functions or to the space of functions with bounded variation (BV)[58,59]. The latter can also account for a binary switching process, i.e., $\Gamma(t) \in \{0, 1\}$. In this case, the underlying parameter dynamics is directly interpolated by K locally stationary models

$$\Theta \approx \sum_{k=1}^{k} \gamma_k(t) \theta_k, \tag{3.24}$$

and the average model distance function (3.23) interpolates the true negative log-likelihood function (3.18)[78]. Because of these considerations, this thesis refers to the space of functions with bounded variations only and incorporates this property by

$$||\gamma_k(t)||_{BV[t_1, t_{N_T}]} = \sum_{j=1}^{N_T - 1} |\gamma_k(t_{j+1}) - \gamma_k(t_j)| \leq C_k(N_T), \quad k = 1, \ldots, K, \tag{3.25}$$

where C_k denotes the maximal number of allowed transitions between the model k and all the other models in the time interval $[t_1, t_{N_T}]$. Later on, we will refer to $C = \max\{C_1(N_T), \ldots, C_K(N_T)\}$[58]. For the observations at finite discrete times, the natural boundary of C is given by N_T (the length of the series of extremes), and thus including constraint (3.25) into the optimization problem will not a priori confine the solution space. We will denote the resulting minimization problem as the "FEM-BV-EVA" approach. The minimization of (3.23) with convexity and persistency constraints (3.21-3.22-3.25) on $\Gamma(t)$ becomes well-posed with respect to $\Gamma(t)$[78]. In the following, we consider the explicit formulation for both the GEV and the GPD approach.

Univariate FEM-BV-GEV

We apply the FEM approach to resolve the nonstationarity in the regression analysis of block-maxima by considering following parametrization for each local GEV model

$$\mu_i(U_t) = \mu_{k0} + \sum_{p=1}^{P} \mu_{kp} u_p(t),\ k = 1,\ldots,K, \tag{3.26}$$

$$\text{and analogue expressions for } \sigma_k(U_t) \text{ and } \xi_k(U_t),$$

and define the local model distance function as the local negative log-likelihood function with $\theta_k = (\mu_{k0},\ldots,\mu_{kP},\sigma_{k0},\ldots,\sigma_{kP},\xi_{k0},\ldots,\xi_{kP})$, $k = 1,\ldots K$, for $\xi_k(U_t) \neq 0$

$$\begin{aligned} g_{GEV}(X_t,\theta_k) = \log(\sigma_k(U_t)) &+ \left(1+\xi_k(U_t)\frac{(X_t-\mu_k(U_t))}{\sigma_k(U_t)}\right)^{-\frac{1}{\xi_k(U_t)}} \\ &+ \left(1+\frac{1}{\xi_k(U_t)}\right)\log\left(1+\xi_k(U_t)\frac{(X_t-\mu_k(U_t))}{\sigma_k(U_t)}\right), \end{aligned} \tag{3.27}$$

and for $\xi_k(U_t) = 0$

$$g_{GEV}(X_t,\theta_k) = \log(\sigma_k(U_t)) + \frac{(X_t-\mu_k(U_t))}{\sigma_k(U_t)} + \exp\left(-\frac{X_t-\mu_k(U_t)}{\sigma_k(U_t)}\right). \tag{3.28}$$

Then for $\Theta_{GEV} = (\theta_1,\ldots,\theta_K)$ the average model distance functional is defined by

$$\mathcal{L}(\Theta_{GEV},\Gamma(t)) = \sum_{j=1}^{N_T}\sum_{k=1}^{K} \gamma_k(t_j)\, g_{GEV}\left(X_{t_j},\theta_k\right), \tag{3.29}$$

with constraints on model parameters

$$[1+\xi_k(U_t)\frac{(X_t-\mu_k(U_t))}{\sigma_k(U_t)}] > 0, \quad \sigma_i(U_t) > 0 \quad \text{for} \quad t = t_1,\ldots,t_{N_T}, \quad k = 1,\ldots,K, \tag{3.30}$$

and with convexity and the persistency constraints on $\Gamma(t) = (\gamma_1(t),\ldots,\gamma_K(t))$.

Univariate FEM-BV-GPD

In the next step we apply FEM to resolve the nonstationarity in regression analysis of threshold excesses. Let us first consider a locally stationary parametrization for the GPD model parameters

$$\tilde{\sigma}_k(U_t) = \tilde{\sigma}_{k0} + \sum_{p=1}^{P} \tilde{\sigma}_{kp} u_p(t), \quad k = 1,\ldots,K, \tag{3.31}$$

and an analogue expression for $\xi_k(U_t)$. Then, the local model distance function is defined as the local negative log-likelihood function with $\theta_k = (\tilde{\sigma}_{k0},\ldots,\tilde{\sigma}_{kS},\xi_{k0},\ldots,\xi_{kS})$, $k = 1,\ldots K$. For

$\xi(U_t) \neq 0$ we get

$$g_{GPD}(X_t, \theta_k) = \log(\tilde{\sigma}(U_t)) + \left(1 + \frac{1}{\xi(U_t)}\right) \log\left(1 + \frac{\xi(U_t) X_t}{\tilde{\sigma}(U_t)}\right), \tag{3.32}$$

$$\text{with } \tilde{\sigma}(U_t) > 0 \text{ and } \left[1 + \frac{\xi(U_t) x}{\tilde{\sigma}(U_t)}\right] > 0, \tag{3.33}$$

and for $\xi(U_t) = 0$ we get

$$g_{GPD}(X_t, \theta_i) = \log(\tilde{\sigma}(U_t)) + \frac{1}{\tilde{\sigma}(U_t)} X_t, \tag{3.34}$$

$$\text{with } \tilde{\sigma}(U_t) > 0. \tag{3.35}$$

For $\Theta_{GPD} = (\theta_1, \ldots, \theta_K)$, the average model distance functional is defined by

$$\mathcal{L}_{GPD}(\Theta_{GPD}, \Gamma(t)) = \sum_{j=1}^{N_T} \sum_{k=1}^{K} \gamma_k(t_j) \, g_{GPD}\left(X_{t_j}, \theta_k\right), \tag{3.36}$$

with constraints (3.33) or (3.35) on model parameters, convexity and persistency constraints on $\Gamma(t) = (\gamma_1(t), \ldots, \gamma_K(t))$.

3.1.4 Conceptual Comparison with the State-of-the-Art Methods

In the following we provide a conceptual comparison between the univariate FEM-BV-EVA and state-of-the-art methods for univariate regression analysis of extremes by adopting the discussions in[64,65,78]. The FEM-BV-EVA provides a tool for nonstationary statistical regression analysis of extremes based on resolved covariates only. FEM-BV-EVA is a semiparametric approach as a combination of the parametric GEV/GPD and the nonparametric FEM description of the hidden switching process. The influence of unresolved factors, expressed as the nonstationary offset term in GEV or GPD model parameters, compare (3.9-3.10-3.11) or (3.15-3.16), respectively is reflected by $\Gamma(t)$. Please note that the key issue that makes the FEM-BV-EVA problem well-posed, is the fact that decreasing the value of C in (3.25) results in shrinking of the parameter space for $\Gamma(t)$, limiting the number of the local minima for $\mathcal{L}(\Theta, \Gamma(t))$.

FEM-BV-EVA approach includes some state-of-the-art approaches as special cases: In the case that the whole information is provided for the regression analysis of extremes, the FEM-BV-EVA with $C = 0$ (no transitions between models, i.e., $K = 1$) corresponds to stationary parametric regression models and results in a well-posed inverse problem. The presence of different models ($K > 1$) indicates the presence of systematically missing covariates in the statistical regression analysis (in contrast to other methods, e.g. based on moving window[32] where the kernel is a priori chosen as some fixed local parametric function like a Gaussian of a certain width). FEM-BV-EVA goes beyond strong a priori probabilistic and deterministic assumptions typical for standard approaches deploying, for instance, Hidden Markov Models (HMM)[7,33,74]. In particular, being a part of the FEM-BV model family it includes methods based on parametric regression, HMM and local kernel smoothing as special cases[64,78].

The deployed linear regression of the model parameters becomes a weakness as soon as the influence of covariates is strongly nonlinear, although, this can be relaxed by involving nonlinear terms, describing the interacting couplings of the covariates (e.g., $u_1^2(t), u_1(t) \cdot u_2(t), \dots$), as additional covariates. The proposed linear and nonstationary FEM-BV-GEV will be compared to the nonlinear and stationary GEV-CDN methodology, which exploits a conditional density network (CDN) for nonlinear regression analysis based on time dependent covariates with stationary neuron weights and offsets [14].
Further, FEM-BV-GPD will be compared to the state-of-the-arts methods based on generalized additive models [18,57,86,114]. These methods are able to resolve the involved nonstationarity in a nonparametric way. We will refer to the gamGPD approach [18], as discussed in detail in Section 2.1.2. The gamGPD approach accounts for the nonstationary offset term in Θ by a smoothing spline regression dependent on time, inheriting the locality property and the inability to describe discontinuous functions. In contrast, the FEM-BV-GPD approach also accounts for discontinuous functions and provides for $K > 1$ a nonlocal extension of the nonparametric smoothing approach, where the nonstationary process $\Gamma(t)$ allows us to consider all observations that belong to the similar dynamics as a single ensemble.

3.2 Data-based spatial extreme value analysis

Let us denote $Y(s,t)$ as an extreme event observed at location s and time t for $t = t_1, \dots, t_{N_T}$ and $s = s_1, \dots, s_{N_S}$. Let $U_{s,t}^{all} = \left(u_1^{all}(s,t), \dots, u_{\mathcal{J}}^{all}(s,t)\right)$ be the set of all covariates that significantly influence the dynamics of $Y(s,t)$. We distinguish between local covariates observed at each location like temperature and humidity and global covariates such as global oscillation patterns like the Arctic Oscillation index (AO), being the same for all locations. In the following, we focus on regression analysis of extremes by exploiting the Bayesian hierarchical approach: in order to model the dynamics of spatial extremes, we incorporate the spatio-temporal variability, namely, the marginal behavior and spatial dependence, into the model parameters. We will refer to a max-stable hierarchical formulation [89] as outlined in Section 2.2.2.
Let us assume that all the significant covariates $U_{s,t}^{all}$ are known and observed. Then, the process $Y(s,t)$ is independent for all time steps t and all locations s

$$Y(s,t)|U_{s,t}^{all} \overset{independent}{\sim} Model\left(\Theta\left(s, U_{s,t}^{all}\right)\right), \tag{3.37}$$

where $Model$ refers either to GEV or to GPD and $\Theta\left(s, U_{s,t}^{all}\right)$ describes the spatio-temporal parameter dynamics. The optimal $Model$ parameters are obtained through constrained minimization of the corresponding negative log-likelihood function (NLL). Since $Y(s,t)$ is conditionally independent for given $U_{s,t}^{all}$, the corresponding likelihood is the product, and NLL the sum, of the corresponding marginal likelihood over all locations

$$NLL\left(Y(s,t); \Theta\left(s, U_{s,t}^{all}\right)\right) = -\sum_{i=1}^{N_S}\sum_{j=1}^{N_T} g_{Model}\left(Y_{s_i,t_j}, \Theta\left(s_i, U_{s_i,t_j}^{all}\right)\right). \tag{3.38}$$

However, as discussed in Section 3.1.1, in real applications we have to deal with systematically missing covariates. For this issue, we will extend the univariate FEM-BV-EVA towards spatio-temporal regression analysis of extremes. In order to resolve the resulting nonstationary spatio-temporal behavior beyond a priori assumptions, we adapt the spatial FEM formulation[38]. In line with the univariate FEM, spatial FEM resolves the involved nonstationarity by interpolating the resulting distance function $g_{Model}(.)$ by locally stationary distance functions and a nonstationary spatial switching process. It will be shown that we obtain a max-stable hierarchical formulation for the spatial FEM-BV-EVA. In the following we propose the spatial FEM-BV-GEV and FEM-BV-GPD formulations.

3.2.1 Spatial FEM-BV-GEV

Let $Y(s,t)$ for $t = t_1,\ldots,t_{N_T}$ and $s = s_1,\ldots,s_{N_S}$ be a series of block-maxima observed over a region. Assuming that the observed process $Y(s,t)$ is max-stable, the marginal distribution for each location is the GEV[89,90]. A further assumption is that the spatio-temporal variability (marginal behavior and spatial dependence) can be described by the GEV model parameters with $Y(s,t) \sim GEV(\mu(s,t),\sigma(s,t),\xi(s,t))$. We now express each GEV parameter as a function of resolved covariates only and account for unresolved covariates by extending the univariate reduced regression formulation (3.9) towards spatial variability. The behavior of each model parameter, for instance, the location parameter, is described by

$$\mu(s,t,U_{s,t}) = \mu_0(s,t) + \sum_{p=1}^{P} \mu_p(s,t)u_p(s,t) + \varepsilon(s,t), \quad \text{with } \varepsilon(s,t) \sim \mathcal{N}(0,\hat{\sigma}(s,t)), \tag{3.39}$$

for $t = t_1,\ldots,t_{N_T}$ and $s = s_1,\ldots,s_{N_S}$. Inline with the univariate FEM-BV-EVA we neglect the noise term $\varepsilon(s,t)$ and obtain a deterministic formulation of the model parameters. With this, the nonstationary spatio-temporal behavior of $Y(s,t)$ is described by

$$\Theta(s,t,U_{s,t}) = (\ \mu_0(s,t),\ldots,\mu_P(s,t),\sigma_0(s,t),\ldots,\sigma_P(s,t),\xi_0(s,t),\ldots,\xi_P(s,t)\ ), \tag{3.40}$$

with appropriate constraints on the scale and shape parameters. In max-stable Bayesian hierarchical models the nonstationary spatio-temporal behavior of the model parameters is approximated by a set of kernel functions with a priori probabilistic assumptions regarding the approximation coefficients[89]. To go beyond such a priori assumptions, i.e., the appropriate choice of the kernel function and the a priori distribution of the approximation coefficients, we adapt the idea proposed in[38]: the underlying parameter dynamics is approximated by $K \geq 1$ locally stationary models and a hidden spatio-temporal switching process $\Gamma(s,t) = (\gamma_1(s,t),\ldots,\gamma_K(s,t))^{\dagger}$ with convexity constraints

$$\gamma_k(s,t) \geq 0, \forall k,t,s \quad \text{and} \quad \sum_{k=1}^{K} \gamma_k(s,t) = 1, \forall t,s. \tag{3.41}$$

That is, the model parameters are approximated as follows:

$$\mu(s,t,U_{s,t}) \approx \sum_{k=1}^{K} \gamma_k(s,t)\mu_k(U_{s,t}) \text{ with } \mu_k(U_{s,t}) = \mu_{k0} + \sum_{p=1}^{P} \mu_{kp} u_p(s,t), \tag{3.42}$$

and analogue expressions for $\sigma(s,t,U_{s,t})$, $\xi(s,t,U_{s,t})$.

The deterministic parametrization of the resulting spatio-temporal behavior of $Y(s,t)$ is $(\Theta,\Gamma(s,t))$ with $\Theta = (\theta_1,\ldots,\theta_K)$ and $\theta_k = (\mu_{k0},\ldots,\mu_{kP},\sigma_{k0},\ldots,\sigma_{kP},\xi_{k0},\ldots,\xi_{kP})$. We avoid probabilistic a priori assumptions on $\Gamma(s,t)$ like stationarity, Gaussian or Markovian behavior but consider $\Gamma(s,t)$ as a spatio-temporal deterministic process. The resulting hierarchical description of the underlying dynamics of $Y(s,t)$ is summarized in Corollary 3.2.1.

Corollary 3.2.1 *Under the assumption that $Y(s,t)$ is a max-stable process and the underlying spatio-temporal dynamics can be described by a set of covariates $U_{s,t}$ together with a convex spatio-temporal process $\Gamma(s,t)$, $Y(s,t)$ is conditionally independent and its distribution is a hierarchical GEV with*

$$Y(s,t)|U_{s,t},\Gamma(s,t) \overset{independent}{\sim} GEV\left(\mu(s,t,U_{s,t}),\sigma(s,t,U_{s,t}),\xi(s,t,U_{s,t})\right), \tag{3.43}$$

where the parameters $\mu(s,t,U_{s,t})$, $\sigma(s,t,U_{s,t})$, $\xi(s,t,U_{s,t})$ are defined according to (3.42). Further, the hierarchical formulation in (3.43) fulfills the properties of a max-stable process.

Proof: The first statement is a direct consequence of considerations made in (3.37) and (3.39 - 3.42). Next, we show that (3.43) fulfills the properties of a max-stable process. For this, according to Resnick[90] (Proposition 5.10), we have to show that the resulting residual process

$$X(s,t) = [1 + \frac{\xi(s,t,U_{s,t})}{\sigma(s,t,U_{s,t})}\{Y(s,t) - \mu(s,t,U_{s,t})\}]^{1/\xi(s,t,U_{s,t})} \tag{3.44}$$

has Fréchet marginals, i.e., $GEV(1,1,1)$, and is max-stable. First, we show that $X(s,t)$ has unit Fréchet marginal distributions by computing for a fixed s^* $\mathbb{P}[X(s^*,t) < c]$ with respect to (3.43)

$$\mathbb{P}[X(s^*,t) < c] = \exp\{-[1 + \frac{\xi(s,t,U_{s,t})}{\sigma(s,t,U_{s,t})}\{c - \mu(s,t,U_{s,t})\}]^{-1/\xi(s,t,U_{s,t})}\}, \tag{3.45}$$

where the model parameters are expressed according to (3.42). Taking into account that for $X(s,t)$ we have $\mu_k(U_{s,t}) = \sigma_k(U_{s,t}) = \xi_k(U_{s,t}) = 1$ for $k = 1,\ldots,K$ and exploiting the convexity constraints on $\Gamma(s,t)$ we get

$$\mathbb{P}[X(s^*,t) < c] = \exp\{-[1 + \frac{1}{1}\{c-1)\}]^{-1/1}\} = \exp\{-\frac{1}{c}\}. \tag{3.46}$$

Since we assume that the dependence structure is completely resolved by the deterministic description of the model parameters, the residual process $X(s,t)$ is independent and the joint distribution is given by

$$\mathbb{P}[X(s_1,t) < c_1,\ldots,X(s_{N_S},t) < c_{N_S}] = \exp\{-\sum_{i=1}^{N_S} \frac{1}{c_i}\}. \tag{3.47}$$

Following, it fulfills the max-stability, since a process is max-stable if for every time step, any set of locations and for any $n > 0$ the following equation

$$\mathbb{P}[X(s_1,t) < nc_1,\ldots,X(s_{N_S},t) < nc_{N_S}]^n = \mathbb{P}[X(s_1,t) < c_1,\ldots,X(s_{N_S},t) < c_{N_S}] \tag{3.48}$$

holds[115]. □

Describing the dynamics of $Y(s,t)$ by the set of parameters $(\Theta,\Gamma(s,t))$ results in an approximation of the model defined in (3.37), where $\Gamma(s,t)$ reflects the unresolved covariates, while also describing the nonstationary spatial dependence structure.

Remark 3.2.2 *The hierarchical model in (3.43) does not provide a formulation of a max-stable process in the classical sense, please see (2.25). The nonparametric description of the spatio-temporal dependence structure has no closed formulation. Consequently, without additional analysis on $\Gamma(s,t)$, (3.43) can not be used for inference like measuring the strength of the spatial dependence or interpolating missing locations. Instead, (3.43) provides a pragmatic description of the underlying spatio-temporal dynamics of extremes, which is consistent with the max-stable postulate.*

In the next step we aim to estimate the optimal parameter set $(\Theta,\Gamma(s,t))$ for (3.43) by minimizing the resulting NLL

$$NLL(Y(s,t);\Theta,\Gamma(s,t)) = -\sum_{i=1}^{N_S}\sum_{j=1}^{N_T} g_{GEV}\left(Y_{s_i,t_j},\Theta,\Gamma(s,t)\right) \tag{3.49}$$

with respect to $(\Theta,\Gamma(s,t))$. There is no analytical solution and the minimization problem becomes more complex with an increase in N_S and N_T. In the following, we restrict $\Gamma(s,t)$ to binary values, i.e., $\Gamma(s,t) \in \{0,1\}$ and obtain

$$NLL(Y(s,t);\Theta,\Gamma(s,t)) = -\sum_{i=1}^{N_S}\sum_{j=1}^{N_T}\sum_{k=1}^{K} \gamma_k(s_i,t_j) g_{GEV}\left(Y_{s_i,t_j},\theta_k\right), \tag{3.50}$$

which provides a locally stationary approximation of the true NLL (3.38). Formulation (3.50) allows a more efficient minimization than (3.49), as will be discussed in detail in Section 4.2.

Remark 3.2.3 *The idea of direct interpolation of NLL, instead of interpolating the model parameters, comes from the original FEM approach[58]. In cases when the distance function g_{Model} is convex, we can also consider $\Gamma(s,t) \in [0,1]$[78]. Then, exploiting the Jensen's inequality we get*

$$g_{Model}\left(X_t,\sum_{k=1}^{K}\gamma_k(t)\theta_k\right) \leq \sum_{k=1}^{K}\gamma_k(t) g_{Model}\left(X_t,\theta_k\right), \tag{3.51}$$

where X_t is the analyzed data, $g_{Model}(.)$ is a convex distance function. In this particular cases the interpolation of the NLL provides an upper bound for the original NLL. For $\Gamma(s,t) \in \{0,1\}$, (3.51) becomes an equality for any g_{Model}. However, the model distance function in the FEM-BV-EVA context is not convex and we have to stick with a binary switching process.

In order to ensure well-posedness of the inverse problem (constrained minimization of (3.50) with respect to $(\Theta,\Gamma(s,t))$, the temporal FEM persistency constraint on $\Gamma(s,t)$ is incorporated

$$||\gamma_k(s,t)||_{BV([t_1,t_{N_T}])} \leq C(N_T), \forall s,k. \tag{3.52}$$

The resulting spatial FEM-BV-GEV approach results in the minimization of

$$\mathcal{L}(\Theta,\Gamma(s,t)) = -\sum_{i=1}^{N_S}\sum_{j=1}^{N_T}\sum_{k=1}^{K}\gamma_k(s_i,t_j)g_{GEV}\left(Y_{s_i,t_j},\theta_k\right), \tag{3.53}$$

with convexity (3.41) and temporal persistency (3.52) constraints on $\Gamma(s,t)$, and with constraints on model parameters

$$[1+\xi_k(U_{s,t})\frac{(X_t-\mu_k(U_{s,t}))}{\sigma_k(U_{s,t})}] > 0, \quad \sigma_i(U_{s,t}) > 0 \quad \text{for} \quad t=t_1,\ldots,t_{N_T}, \quad k=1,\ldots,K. \tag{3.54}$$

3.2.2 Spatial FEM-BV-GPD

In the next step we refer to $Y(s,t)$ as a series of threshold excesses observed at location s for $s=s_1,\ldots,s_{N_S}$ and time t with $t=t_1,\ldots,t_{N_{T_s}}$. Please note that in order to extract the threshold excesses, we fix a quantile and estimate the threshold according to this quantile individually for each location. Consequently, each location has different length of observed threshold excesses. The results obtained in Section 3.2.1 can be applied directly for spatio-temporal regression analysis of threshold excesses by expressing the GPD model parameters analogous to (3.39). In order to avoid a priori probabilistic assumptions, we employ the spatial FEM approach and obtain in line with spatial FEM-BV-GEV a nonstationary parametrization of the GPD model parameters

$$\tilde{\sigma}(s,t,U_{s,t}) \approx \sum_{k=1}^{K}\gamma_k(s,t)\tilde{\sigma}_k(U_{s,t}) \text{ with } \tilde{\sigma}_k(U_{s,t}) = \tilde{\sigma}_{k0} + \sum_{p=1}^{P}\tilde{\sigma}_{kp}u_p(s,t), \tag{3.55}$$

and analogue expression for $\xi(s,t,U_{s,t})$,

with

$$\Theta(s,t,U_{s,t}) = (\ \tilde{\sigma}_0(s,t),\ldots,\tilde{\sigma}_P(s,t),\xi_0(s,t),\ldots,\xi_P(s,t)\ ). \tag{3.56}$$

Conditioned on $U_{s,t}$ and $\Gamma(s,t)$, the process $Y(s,t)$ is independent

$$Y(s,t)|U_{s,t},\Gamma(s,t) \overset{independent}{\sim} GPD(\tilde{\sigma}(s,t,U_{s,t}),\xi(s,t,U_{s,t})). \tag{3.57}$$

Then, analogous to the FEM-BV-GEV formulation, (3.57) provides a max-stable description with Fréchet margins for the residual process obtained by transforming $Y(s,t)$ with respect to

$$X(s,t) = -\frac{1}{\log(F_s(Y(s,t);\Theta(s,t,U_{s,t})))}, \tag{3.58}$$

for $s = s_1 \ldots, s_{N_S}, t = t_1, \ldots, t_{N_{T_s}}$ and $F_s(.)$ are the GPD marginals with $\Theta(s,t,U_{s,t})$ as defined in (3.56). For a binary $\Gamma(s,t)$ the corresponding NLL, which is given by

$$NLL(Y(s,t);\Theta,\Gamma(s,t)) = -\sum_{i=1}^{N_S}\sum_{j=1}^{N_{T_s}}\sum_{k=1}^{K}\gamma_k(s_i,t_j)g_{GPD}\left(Y_{s_i,t_j},\theta_k\right), \tag{3.59}$$

provides a locally stationary approximation of the true NLL (3.38). The resulting spatial FEM-BV-GPD approach results in the minimization of

$$\mathcal{L}(\Theta,\Gamma(s,t) = -\sum_{i=1}^{N_S}\sum_{j=1}^{N_{T_s}}\sum_{k=1}^{K}\gamma_k(s_i,t_j)g_{GPD}\left(Y_{s_i,t_j},\theta_k\right), \tag{3.60}$$

with convexity and temporal persistency constraints on $\Gamma(s,t)$ and with

$$\left[1+\frac{\xi_k(U_{s,t})x}{\tilde{\sigma}_k(U_{s,t})}\right] > 0 \quad \text{and} \quad \tilde{\sigma}_k(U_{s,t}) > 0 \quad \text{for} \quad t = t_1,\ldots,t_{N_{T_s}}, \quad k = 1,\ldots,K. \tag{3.61}$$

In total, the application of spatial FEM-BV-EVA approach to a spatiotemporal series of extremes results in a set of $K \geq 1$ locally stationary model parameters $\Theta = (\theta_1,\ldots,\theta_K)$ and a spatiotemporal switching process. FEM-BV-EVA is a max-stable hierarchical description of the underlying dynamics of extremes. And while the model parameters are accessible for all locations, the nonstationary switching process is assigned to each location separately (each location s^* is associated with a $\Gamma(s^*,t)$, which describes the temporal affiliation to one of the models).

3.2.3 Conceptual Comparison with State-of-the-Art Methods

The spatial FEM-BV-EVA approach can be ranged into the class of max-stable hierarchical models. The spatial dependence in a kernel-based max-stable hierarchical model is approximated by some predetermined kernels and a priori assumptions about the involved coefficients[89]. Spatial FEM-BV-EVA approximates the dependence structure via a nonparametric and nonstationary switching process with bounded variation. Spatial FEM-BV-EVA goes beyond a priori assumptions commonly made in geostatistics of extremes, for instance, Gaussian assumptions about the dependence structure. Standard techniques approximate the underlying true likelihood by the composite likelihood while FEM-BV-EVA interpolates it by the means of convex linear combination of locally stationary likelihoods. A consequence of a nonparametric description of the spatial dependence structure is that there is no model which can be used for further investigation of the spatial dependency. For instance, neither it is possible to measure the strength of the spatial dependence, nor can the resulting description be used for spatial interpolation of missing locations without additional analysis of the resulting switching process. In contrast, these issues can be directly approached by the parametric max-stable models[25,31]. The appropriate analysis of the FEM-BV-EVA switching process remains for future work. In this thesis, in order to understand the underlying spatial dependence structure in more detail, we will analyze $\Gamma(s,t)$ by exploiting the results obtained in the field of complex networks[76], please see Section 4.3.2.

3.3 Conclusion

In this chapter we proposed the nonstationary, nonparametric FEM-BV-EVA approach for univariate and spatial regression analysis of extreme events based on resolved covariates only. Exploiting theoretical aspects of extreme value analysis (EVA) and the FEM time series analysis methodology for approaching the involved nonstationarity, the resulting FEM-BV-EVA avoids some a priori assumptions made in state-of-the-art approaches. Although, the fully nonstationary and nonparametric modeling of univariate extreme events is widely used, compare Section 2.1.2, the formulation for the reduced regression model proposed in this thesis is new. Based on resolved covariates only, the reduced regression model goes beyond i.i.d assumptions of the unobserved covariates by exploiting Lindeberg and Karhunene-Loève Theorems. In particular, this result emphasizes the significance of nonparametric nonstationary regression analysis of extreme events, since in real applications we often have to deal with unresolved covariates. The spatial FEM-BV-EVA extension is a hierarchical max-stable formulation that describes the spatio-temporal dynamics of extremes by expressing the parameters as spatial and nonstationary regression models based on resolved covariates only. The nonstationarity is resolved by incorporating the spatial FEM formulation, i.e., describing the underlying dynamics by a set of locally stationary models and a spatial nonstationary switching process $\Gamma(s,t)$. In addition, $\Gamma(s,t)$ provides a pragmatic nonparametric and nonstationary description of the underlying spatial dependence structure by grouping together all locations that exhibit similar behavior via the model-affiliation function $\Gamma(s,t)$.
In conclusion, FEM-BV-EVA provides a purely data-driven, space-time clustering approach to study the spatio-temporal dynamics of extremes beyond strong a priori assumptions. However, a clear limitation is that there is no closed formulation for the dependence structure and further investigations, for instance, analysis of the strength of the spatial dependence, can not be carried out directly. In the next chapter we present the FEM-BV-EVA framework.

4 Computational/Algorithmic Aspects of FEM-BV-EVA Framework

In this chapter we present the FEM-BV-EVA framework to be integrated into the object-oriented FEM MATLAB toolbox[1]. Applied to a series of extremes, the FEM-BV-EVA framework describes their underlying dynamics by the following set of parameters: the number of models K, the corresponding switching process $\Gamma(s,t)$, the model parameter Θ and the maximal number of switches between the models $C(N_T)$. Thereby, for every fixed set of K and $C(N_T)$ FEM-BV-EVA results in minimization of the following objective functional

$$\mathcal{L}(\Theta,\Gamma(s,t)) = -\sum_{i=1}^{N_S}\sum_{j=1}^{N_T}\sum_{k=1}^{K}\gamma_k(s_i,t_j)\, g_{Model}\left(Y_{s_i,t_j},\theta_k\right), \tag{4.1}$$

where $g_{Model}(.)$ is either the GEV or the GPD log-likelihood and $\theta_k, k = 1,\ldots,K$, fulfills either (3.30) or (3.33/3.35), respectively. Further, $\Gamma(s,t)$ fulfills the convexity constraints:

$$\sum_{k=1}^{K}\gamma_k(s,t) = 1, \quad s = s_1,\ldots,s_{N_S}, \quad t = t_1,\ldots,t_{N_T}, \tag{4.2}$$

$$\gamma_k(s,t) \geq 0, \quad s = s_1,\ldots,s_{N_S}, \quad t = t_1,\ldots,t_{N_T}, \quad k = 1,\ldots,K, \tag{4.3}$$

and the persistency constraint:

$$||\gamma_k(s,t)||_{\mathcal{BV}([t_1,t_{N_T}])} \leq C(N_T), \qquad s = s_1,\ldots,s_{N_S}, \quad k = 1,\ldots,K. \tag{4.4}$$

As the objective functional (4.1) is not convex, there exists no global solution of this constrained minimization problem. A local optimal solution can be found through alternating optimization with respect to $\Gamma(s,t)$ and Θ, as will be discussed in Section 4.2. In order to obtain the global optimal solution, we need to explore the whole solution space as far as possible. Deterministic exploration becomes computationally infeasible with increasing dimension of the problem, for instance, when the number of observed locations is increasing. Thus, random exploration is required[101]. In the

[1]The FEM MATLAB toolbox combines the family of FEM-based methods developed in the working group of Illia Horenko at the Institute of Computational Science (Universitá della Svizerra Italiana) in Lugano. The toolbox was implemented by Dimitri Igdalov.

FEM framework the random exploration is implemented by starting the local optimization several times with random initializations of the switching process (assuming that the chosen number of trials is hight enough to provide the global optimal parameter set).
The selection of optimal K and $C(N_T)$ from a set of all possible combinations is carried out according to the information criteria (IC) of a model [12]. The significant subset of resolved covariates is determined either by considering all possible combinations and deploying IC, or by employing shrinkage techniques [105]. For instance, deploying the Lasso shrinkage approach: by constraining the $L1$ norm of the model parameters Θ, the coefficients of insignificant covariates are set to zero. In the following sections we discuss the implementation of the FEM-BV-EVA framework in detail.

4.1 Model Selection and Lasso Regularization

In many real applications not only one, but a set of candidate models is considered for a given data set. In FEM context, such a set is obtained for different combinations of the parameters K and C. The comparison of the candidate models or rather the selection of the most appropriate one is an important issue. Often, the principle of "parsimony" also called the "Ockam's razor" is applied [12,29]. Indicating that the less complex and at the same time the most informative model should be chosen. Based on the Kullback-Leibler "distance"-measure between probability distributions, Akaike derived in 1973 a model selection criteria for probabilistic models; the Akaike Information Criteria (AIC) [1,61]:

$$AIC = -2L + 2|M|, \tag{4.5}$$

where L is the log-likelihood function for the estimated model M and $|M|$ denotes the number of parameters in this model. AIC is viewed as an extension of the classical maximum likelihood approach and coincides with the "parsimony" principle: maximizing the likelihood L and at the same time penalizing the number of parameters $|M|$.
Since FEM-BV-EVA is minimizing a negative log-likelihood function, we can straightforwardly apply AIC. In FEM-BV-EVA formulation the objective functional (4.1) corresponds to the averaged negative log-likelihood (NLL) and the number of parameters is dependent on $K, C(N_T)$ and the number of involved covariates P. We obtain that $-L = \mathcal{L}(\Gamma(t), \Theta)$ and $|M| = |M(K, C(N_T), P)|$. Please note that the derivation of AIC assumes that the number of the sample, N_T, is sufficiently large. This is not always provided, particularly when studying extreme events. Therefore, FEM-BV-EVA also incorporates the second-order AIC (AIC_c) [61] as a valid estimate for the information content of a sample with a (small) finite length [12]. AIC_c is defined by

$$AIC_c = AIC + \frac{2|M|(|M|+1)}{N_T - |M| - 1}, \tag{4.6}$$

and converges to AIC for large N_T. We compute the appropriate IC for each model M, and choose the best model, denoted by M^*, with respect to $\min(IC)$. Estimating IC values for N different models, we see that some have a relatively small discriminant $\Delta_v = IC(v) - min(IC)$, $v = 1 \ldots, N$ and other

values can have a large discriminant. In order to explain the differences Δ_v, one can compute the corresponding Bayesian posterior model probabilities known as Akaike model weights[12] by

$$\rho\left(M_v\right) = \frac{\exp(-\frac{\Delta_v}{2})}{\sum_{v=1}^{N_{(K,C,S)}} \exp(-\frac{\Delta_v}{2})}, \text{ with } v = 1,\ldots,N, \tag{4.7}$$

where the denominator specifies the normalization constant. The most accurate, i.e., IC-minimal, model gets the maximal weight[12]. Further, we aim to detect the most significant set of covariates out of $U_{s,t} \in \mathbb{R}^P$. One approach is to start FEM-BV-EVA with all possible combinations of $U_{s,t}$, in total:

$$\sum_{p=1}^{P} \frac{P!}{(P-p)!p!}, \tag{4.8}$$

and to choose the optimal one according to model selection criteria, e.g., AIC_c. In the cases when P is big, the number of models M increases quickly and attempting to test all possible combinations of $U_{s,t}$ becomes computationally expensive. Alternatively, the significant subset of resolved covariates can be determined by incorporating shrinkage techniques on model parameter Θ such as the Lasso and Ridge techniques[105]. The Ridge technique shrinks the parameters by constraining their $L2$ norm and the Lasso technique by constraining their $L1$ norm. Additionally, Lasso shrinkage selects the most significant covariates by setting the insignificant coefficients to zero[57,110]. In the FEM-BV-EVA framework, both shrinkage techniques are implemented. In this thesis the Lasso shrinkage is deployed, i.e., it is required that

$$||\Theta||_{L1} \leq C_L. \tag{4.9}$$

By incorporating the $L1$ constraints on Θ via Tikhonov regularization, we obtain the final FEM-BV-EVA minimization problem for $\lambda \geq 0$

$$\mathcal{L}\left(\Theta,\Gamma(s,t)\right) = -\sum_{i=1}^{N_S}\sum_{j=1}^{N_T}\sum_{k=1}^{K} \gamma_k\left(s_i,t_j\right) g_{Model}\left(Y_{s_i,t_j},\theta_k\right) + \lambda||\Theta||_{L1}, \tag{4.10}$$

with constraints on Θ: (3.30) or (3.33/ 3.35), for GEV and GPD respectively
and constraints on $\Gamma(s,t)$: (4.2-4.3-4.4).

Please note that in case we incorporate some shrinkage on Θ, i.e., $\lambda \neq 0$, the number of model parameters, $|M|$, includes only "nonzero" coefficients of Θ (all coefficients with an absolute value above some predefined threshold, ε_Θ, are considered as "nonzero").
Concluding, the "Lasso FEM-BV-EVA" formulation aims to find the optimal set of $K, C(N_T), \lambda$. We start the minimization of (4.10) for different values of $K, C(N_T), \lambda$ and obtain the optimal model by applying either AIC or AIC_c. In the next section we describe the general FEM framework and the deployed algorithms for estimating the optimal EVA model parameters.

4.2 Implementation

The main steps of the general FEM formulation are outlined in Algorithm 2: In the first step a candidate model M is estimated for different values of K, $C(N_T)$ and λ, summarized in K_{list}, $C(N_T)_{list}$, λ_{list} (see Algorithm 2, line 4). The second step is to select the optimal model M^*: choose the optimal K^*, $C^*(N_T)$, λ^* according to the appropriate IC (see Algorithm 2, line 5).
For a fixed set of $\{K, C(N_T), \lambda\}$ model M is obtained by solving (4.10). As already mentioned,

Algorithm 2: The general FEM algorithm

input : Observed series $Y_{s,t}$, covariates $U_{s,t}$, K_{list}, $C(N_T)_{list}$, λ_{list}
output: Optimal model M^* consists of optimal $K^*, C^*(N_T), \lambda^*$ and $(\Theta^*, \Gamma^*(s,t))$

1 **for** λ **do**
2 **for** *K list* **do**
3 **for** *C list* **do**
4 **Step1:** $(\Theta^*, \Gamma^*(s,t)) = getOptimalParameterSet(\lambda, K, C(N_T))$; For fixed λ, K and $C(N_T)$ estimate the global optimal parameter set $(\Theta^*, \Gamma^*(s,t))$ (compare Algorithm 3).
5 **Step2:** $M^* = updateOptimalModel(\Theta^*, \Gamma^*(s,t), \lambda^*)$; Estimate the IC value according to (4.5) or (4.6) for every model M. If the current IC value is smaller then the previous one assign $M^* = M$.

there is no global solution of (4.10). To obtain a local optima, the general FEM framework solves the minimization problem (4.10) in an alternating order[58,78]. It exploits the fact, that for fixed Θ the minimization of (4.10) with respect to $\Gamma(s,t)$ results either in a constrained linear or constrained quadratic problem, referring to BV and H1 regularization respectively. Analogues, for fixed $\Gamma(s,t)$ (4.10) can be minimized with respect to Θ. In some FEM settings, for instance, in the FEM-Markov approach, an analytical solution is available[60,78], in some others, such as in the presented FEM-EVA approach, gradient- or MCMC-based techniques are required.
In the following, this proceeding, also denoted as the subspace iteration, is explained in more details (see Algorithm 3): beginning with a randomly initialized $\Gamma(s,t)$ in an alternating order we estimate Θ for a fixed $\Gamma(s,t)$ and then $\Gamma(s,t)$ for a fixed Θ. In each alternating step the value of $\mathcal{L}(\Gamma(s,t), \Theta)$ in (4.10) is reduced. The subspace iteration converges to a local optimum when the decrease in $\mathcal{L}(\Gamma(s,t), \Theta)$ is less then a predefined minimization threshold, denoted in the following by Tol. In order to obtain the global optimum, FEM framework is started with random initializations several times[78] when $\Gamma(s,t)$ is initialized randomly (see Algorithm 3, line 2).
The two steps of the subspace iteration are carried out as follows. For a fixed parameter Θ, $\Gamma(s,t)$ is discretized by the Finite Element Method and estimation of $\Gamma_{opt}(s,t)$ results in a linear constrained minimization problem that can be solved using standard numerical tools, e.g., simplex method[59,78].

Algorithm 3: getOptimalParameterSet(); Annealing and subspace iteration

input : Observed series $Y_{s,t}$, covariates $U_{s,t}$, fixed $\{K, C(N_T), \lambda\}$, minimization threshold value, Tol, number of random initializations *numberInit*, *annealing*, maximal number of subspace iterations, *maxSubspace*

output: Global optimal parameter set $(\Theta^*, \Gamma^*(s,t))$

$\mathcal{L}(\Theta^*, \Gamma^*(s,t)) = inf$

for *r = 1:numberInit* **do**

  $\Gamma_{old}(s,t)$ generate random wrt constraints (4.2-4.3-4.4)

  $\Theta_{old} = \underset{\Theta}{\operatorname{argmin}}\, \mathcal{L}(\Theta, \Gamma_{old}(s,t))$

  while $|\mathcal{L}(\Theta_{opt}, \Gamma_{opt}(s,t)) - \mathcal{L}(\Theta_{old}, \Gamma_{old}(s,t))| > Tol$ *or* *maxSubspace* **do**

    Step1: $\Gamma_{opt}(s,t) = \operatorname{argmin}\, \mathcal{L}(\Theta_{old}, \Gamma(s,t))$; The constrained minimization wrt. $\Gamma(s,t)$ results for BV-regularization in a linear problem, standard methods, e.g., simplex method, can be applied.

    Step2: $\Theta_{opt} = \operatorname{argmin}\, \mathcal{L}(\Gamma_{opt}(s,t), \Theta)$; The required numerical optimization method wrt Θ depends on the model distance function $g_{Model}(.)$. In FEM-BV-EVA $g_{Model}(.)$ is the GEV or GPD negative log-likelihood and the minimization is carried out by applying MCMC method (compare Algorithm 4).

  if $\mathcal{L}(\Theta^*, \Gamma^*(s,t)) > \mathcal{L}(\Theta_{opt}, \Gamma_{opt}(s,t))$ **then**

    $\Theta^* = \Theta_{opt}$

    $\Gamma^*(s,t) = \Gamma_{opt}(s,t)$

For a fixed $\Gamma(s,t)$, Θ_{opt} is obtained by constrained minimization of $\mathcal{L}(\Gamma(s,t), \Theta)$ with respect to Θ. Exploiting the fact that $\mathcal{L}(\Gamma(s,t), \Theta)$ is uncoupled for different $k = 1, \ldots, K$, Θ_{opt} can be estimated separately for each k by solving

$$\min_{\theta_k} \sum_{i=1}^{N_S} \sum_{j=1}^{N_T} \gamma_k(s_i, t_j)\, g_{Model}\left(Y_{s_i,t_j}, \theta_k\right), \quad k = 1, \ldots, K \tag{4.11}$$

with constraints (4.2-4.3-4.4),

with standard likelihood maximization techniques[22,40]. Note that the corresponding function in (4.11) is strongly nonlinear and non-convex. Additionally, in practical applications it may be non-differentiable (or may exhibit very large values of the first derivative). Therefore, minimization using standard gradient-based methods like Newton's method and gradient descent approaches would be strongly dependent on the initial value and on the boundedness of the first derivatives (e.g., as in the case of the Levenberg-Marquardt optimization algorithm deployed in GEV-CDN[14]). To avoid these difficulties, we will consider a gradient-free optimization technique based on the Metropolis algorithm, which is a Markov Chain Monte Carlo (MCMC) method. In particular, we employ

the adaptive MCMC methodology proposed in[37], where the adaptive MCMC optimization method considers the Boltzmann distribution as the target density

$$\pi(.) = \frac{1}{z}\exp\left(-\beta h(.)\right), \tag{4.12}$$

with normalization constant z, inverse temperature parameter β and some energy function $h(.)$. For $\beta \to \infty$ Boltzmann distributed samples converge towards the minimal energy of $h(.)$. The adaptivity of the MCMC in[37] comes from adjusting the noise, used for proposing the next sample, and from increasing β. This approach can be used as an optimization method to obtain Θ_{opt} for fixed $\Gamma(s,t)$. For that we set $h(\Theta) = \mathcal{L}(\Gamma(s,t),\Theta)$, modify the MCMC by adjusting the "initialization" and the "proposing the next sample" steps (taking into account the constraints on and the dimensionality of Θ)[37]. The main steps of the deployed adaptive MCMC are sketched in Algorithm 4. The implementation of the deployed MCMC algorithm was carried out in C++. We would like to emphasize that in each run of the MCMC algorithm, it is sufficient to sample a parameter Θ_{new} that provides a smaller value of $\mathcal{L}(\Theta_{new},\Gamma(s,t))$ rather than sampling the entire distribution (refer to the algorithm 4 lines 2, 5). The subspace iteration deployed by FEM-BV minimizes in each step $\mathcal{L}(\Theta,\Gamma(s,t))$ and provides the optimal parameter set $(\Theta_{opt},\Gamma_{opt}(s,t))$ for each annealing step. Moreover, the deployed MCMC optimization technique does not depend on the initial start values: since MCMC algorithm also accepts parameters with higher value of $\mathcal{L}(\Theta,\Gamma(s,t))$, there is a chance to obtain the global minima starting from any initial value. Further, as will be demonstrated on the numerical examples in Chapter 5, the deployed MCMC optimization technique is efficient in terms of computational time.

4.2.1 Details on the adaptive MCMC algorithm

In the following, we show the main steps of the deployed MCMC-based optimization proposed in Algorithm 4. The algorithm is based on the work of[37] and differs mainly in two steps; line 1 and line 4 (as explained in the following sections). Note that the convergency conditions for this algorithm are fulfilled if the MCMC proposes a new parameter set that provides a smaller $\mathcal{L}(\Theta_{new},\Gamma(s,t))$ value for a fixed $\Gamma(s,t)$. It is recommended to limit the number of samples, since as soon as we get into to the area of the local optima it becomes difficult to propose an improved parameter set. In FEM-BV-EVA the number of samplings is limited by the parameter *sampleSizeMCMC* (see Algorithm 4, line 3). Should the algorithm fail, meaning it does not provide an improved set of parameters, it returns with $\Theta_{opt} = \Theta_{old}$ (compare Algorithm 4, line 9).

Algorithm 4: MCMC-based optimization algorithm for fixed $\Gamma(s,t)$

input : $Y_{s,t}$ series of extremes, covariates $U_{s,t}$, λ, $\Gamma(s,t)$, $\mathcal{L}(\Theta_{old},\Gamma(s,t))$
output: Θ_{opt}

1 Θ_{new} = generateInitialValue($\Gamma(s,t)$, $Y_{s,t}$, $U_{s,t}$)

if $\mathcal{L}(\Theta_{new},\Gamma(s,t)) < \mathcal{L}(\Theta_{old},\Gamma(s,t))$ **then**
  $\Theta_{opt} = \Theta_{new}$
2 **return** Θ_{opt}

Initialize: δ, β, Σ, counterAccept = 0;

3 **for** *sampleStep=1:sampleSizeMCMC* **do**
4 Θ_{next} = proposeNext(Θ_{new}, $\Gamma(s,t)$, $Y_{s,t}$, $U_{s,t}$, Σ, *noise*, β)
  if $\mathcal{L}(\Theta_{next},\Gamma(s,t)) < \mathcal{L}(\Theta_{old},\Gamma(s,t))$ **then**
    $\Theta_{opt} = \Theta_{next}$
5 **return** Θ_{opt}
6 **else if** *checkAcceptance(β, Θ_{next}, Θ_{new})* **then**
    $\Theta_{new} = \Theta_{next}$
    counterAccept= +1
7 updateCovMatrix(Θ_{new}, Σ)
  if $sampleStep >= 50$ **then**
8 [δ, β] = adaptStep(δ, β, counterAccept, sampleStep)

9 $\Theta_{opt} = \Theta_{old}$

Generating initial value

The first step of the MCMC sampling is the generation of an initial value (refer to Algorithm 4, line 1). In FEM-BV-EVA the scale and shape parameters fulfill the following constraints

$$0 < \sigma_k(U_{s,t}) = \sigma_{k0} + \sum_{p=1}^{P} \sigma_{kp} u_p(s,t) < const, k = 1,\ldots,K, \forall s,t, \tag{4.13}$$

$$-0.5 < \xi_k(U_{s,t}) = \xi_{k0} + \sum_{p=1}^{P} \xi_{kp} u_p(s,t) < 0.5, k = 1,\ldots,K, \forall s,t. \tag{4.14}$$

The constraint (4.14) ensures a regular likelihood estimator[23,98]. That is, for a large sample, the obtained model parameter follows a normal distribution where the mean is the true parameter and the variance corresponds to the observed information matrix, which is the negative Hessian matrix of the log-likelihood with respect to the model parameters[22,29].
An initial value, which is sampled from a uniform distribution, does not necessarily ensure the constraints (4.13-4.14). In order to hold the constraints, we reformulate them: Since $\xi_k(U_{s,t})$ and

$\sigma_k(U_{s,t})$, for $k = 1,\dots,K$, attain their unique maxima/minima values in one of the corners of the convex hull defined by $U_{s,t}$ [60], it is sufficient to fulfill the constraints (4.13-4.14) on all corners of the convex hull of $U_{s,t}$. Using the matrix $A \in \mathbb{R}^{(P+1)\times 2^P}$, which contains all combinations of maximal/minimal values of $U(s,t)$ for $s = s_1,\dots,s_{N_S}$ and $t = t_1,\dots,t_{N_T}$, we reformulate the constraint for $\xi_i = (\xi_{k0},\dots,\xi_{kP})$ by

$$-A\xi_k < -lb_\xi, \quad lb_\xi = -0.5\cdot\mathbf{1} \in \mathbb{R}^{2^S}, \tag{4.15}$$

$$A\xi_k < +ub_\xi, \quad ub_\xi = 0.5\cdot\mathbf{1} \in \mathbb{R}^{2^S}, \tag{4.16}$$

for $k = 1,\dots,K$. The same applies for σ. Lastly, if we strengthen the constraints slightly we get

$$\sigma_k(U_{s,t}) \in [\varepsilon, const], \quad \xi_k(U_{s,t}) \in [-0.5+\varepsilon, 0.5-\varepsilon], \quad k = 1,\dots,K, \tag{4.17}$$

with $\varepsilon > 0$, and $const \in \mathbb{R}$ to be some high value. We can use a convex sampler to get random, uniform distributed values within this convex hull. When studying the behavior of block-maxima, we must also provide a start value for the location parameter, for instance, by applying a similar procedure to sample $\mu_k = (\mu_{k0},\dots,\mu_{kP})$, $k = 1,\dots,K$, in a way that the constraint (3.30) is fulfilled. An alternative is to estimate the initial value for μ_k by applying ordinary least squares [23]. Note that this estimation is not considered as the trend estimate for the GEV distribution, but rather a procedure to generate an initial value that is adjusted within the MCMC and the subspace procedure (compare Algorithm 3). Both possibilities are implemented in the FEM-BV-GEV framework.

Propose Next Sample

The performance of the Metropolis algorithm can be improved with an appropriate proposal distribution [11,70]. However, it might not be obvious which proposal density should be chosen for the current target density. In this work we deploy the Adaptive Metropolis algorithm [11], where the next proposal, denoted here as Y_{n+1}, is sampled according to a mixture distribution with respect to the information of all previous accepted samples, denoted here as $X_0,\dots,X_n$. That is, we obtain the next sample by

$$Y_{n+1} \sim (1-\delta)\mathcal{N}\left(X_n, \frac{2.38^2}{d}\Sigma_n\right) + \delta\mathcal{N}(X_n, \Sigma_0), \tag{4.18}$$

where d is the dimension of X_n and $\Sigma_n \in \mathbb{R}^{d\times d}$ corresponds to the empirical covariance matrix of $X_0,\dots,X_n$. The parameter $0 < \delta < 1$ controls the acceptance rate of the Metropolis algorithm, the acceptance rate is increasing for $\delta \to 1$ and decreasing for $\delta \to 0$. Details on this adaption step can be seen in [37].

4.3 Postprocessing of FEM-BV-EVA Results

The proposed FEM-BV-EVA toolbox can be applied for spatio-temporal statistical regression analysis of extreme events. The resulting optimal model parameters are used for further analysis, such as parameter sensitivity, model validation, and estimation of return levels. The following two sections concentrate on postprocessing of the FEM-BV-EVA results.

4.3.1 Descriptive Statistics

In addition to the optimal parameter set, we are also interested in the sensitivity of the involved estimator with respect to observed measurements. We aim to provide the confidence intervals for the model parameters. For artificial test cases, where a repetition of the experiment is possible, we obtain the confidence intervals via a bootstrapping procedure[21]: we resample the series of interest according to the underlying dynamics N times and apply FEM-BV-EVA each time. Each optimal result $(\Theta^*, \Gamma^*(s,t))$ is then accounted for estimating the averaged parameters as well as the confidence intervals. In many real applications we dispose of only one realization, for instance, in climate research only one observed realization of the process is available. However, we can exploit the fact that the optimal FEM-BV-EVA model parameters are obtained by minimizing the NLL, and that the constraints on the scale parameter insure the regularity condition of the estimator. That is, for large samples, each local model parameter θ_k, $k = 1, \ldots, K$, is normally distributed as discussed in Section 4.2.1. From here, we can evaluate the confidence intervals (as standard errors) for each θ_k as the root of the diagonal of the corresponding covariance, i.e., observed information matrix. Please note that in contrast to univariate FEM-BV-EVA, where the constraint on large sample might be not fulfilled, the size of the sample is increasing with an increasing number of considered locations in spatial FEM-BV-EVA.
Further, using the FEM-BV-EVA model parameters, we can construct the marginal cumulative distribution function (cdf) for each location s, $s = s_1, \ldots, s_{N_S}$, by the convex combination of local functions, as exemplified for the FEM-BV-GEV model

$$f_{FEM-BV-GEV}\left(x;\Theta,\Gamma(s,t)\right) = \sum_{k=1}^{K} \gamma_k(s,t) \exp\left(-[1+\xi_k(U_{s,t})\frac{x-\mu_k(U_{s,t})}{\sigma_k(U_{s,t})}]^{-\frac{1}{\xi_k(U_{s,t})}} \right), \qquad (4.19)$$

where $\mu_k(U_{s,t}), \sigma_k(U_{s,t}), \xi_k(U_{s,t})$ are defined according to (3.42). The cdf can be used for standard inference such as computing the time-dependent probability of an event A by $\mathbb{P}[Y_{s,t} > A, t]$ and/or estimating the return levels and periods. A return level x_p is the expected value, with $f_{FEM-BV-GEV}\left(x_p;\Theta,\Gamma(s,t)\right) = 1-p$, to be exceeded every $1/p$ years[22]. The return level is estimated by inverting the FEM-BV-GEV distribution function (4.19).

4.3.2 Spatial Dependence

The application of spatial FEM describes the underlying model by a set of appropriate local models and a spatio-temporal switching process $\Gamma(s,t)$. The switching process reflects not only the nonstationarity of the underlying dynamics (resolving so the unobserved covariates) but also describes the spatial dependence structure. Since the description of the switching process is nonparametric, we do not obtain a closed formulation of the underlying dependence structure. In order to draw conclusion about the underlying dependence structure, further analysis of $\Gamma(s,t)$ is required. One possibility is to consider all locations that exhibit similar values in $\Gamma(s,t)$ as adjacent. For example, the degree of spatial dependency for any two locations s_i and s_j is strong if $\Gamma(s_i,t) = \Gamma(s_j,t)$, element-wise.
However, this procedure is not directly applicable in the context of extreme events: here we have to account for time delay in the occurrence of extremes among locations. As an alternative, we

exploit results coming from the field of complex networks, referring to the event synchronization (ES) measure[76]: consider the event occurrence for each location s, that is $t(s) = t_1, \ldots, t_{N_{T(s)}}$. Then, for each pair s_i and s_j the number of times an event appears first at s_i and then at s_j is denoted by

$$c(j|i) = \sum_{l=1}^{N_{T(s_j)}} \sum_{m=1}^{N_{T(s_i)}} J_{ji}, \tag{4.20}$$

where

$$J_{ji} = \begin{cases} 1, & \text{if} \quad 0 < t_l(s_j) - t_m(s_i) < \tau_{lm}^{ji} \\ \frac{1}{2}, & \text{if} \quad t_l(s_j) = t_m(s_i) \\ 0, & \text{else,} \end{cases} \tag{4.21}$$

and τ_{lm}^{ji} describes the minimal time lag between two events

$$\tau_{lm}^{ji} = \min\{t_{l+1}(s_j) - t_l(s_j), t_l(s_j) - t_{l-1}(s_j), t_{m+1}(s_i) - t_m(s_i), t_m(s_i) - t_{m-1}(s_i), \}. \tag{4.22}$$

Analogous we obtain $c(i|j)$ and compute

$$Q_{ji} = \frac{c(j|i) + c(i|j)}{\sqrt{N_{T(s_j)} N_{T(s_i)}}}. \tag{4.23}$$

The symmetric matrix Q is a measure for the relative strength of event synchronization among all locations. In the context of FEM-BV-EVA, we estimate the ES matrix for each local model $k = 1, \ldots, K$ according to the temporal affiliation to this particular model. Following, we obtain an extended description of the ES measure: for each local model $k = 1, \ldots, K$ the correlation among locations is described by the matrix Q_k.
Further, a detailed investigation of the switching process $\Gamma(s,t)$ can reveal the spatio-temporal propagation of extremes. For this purpose, the recently proposed FEM-BV-Causality approach enables to study the spatio-temporal interaction of discrete state models in a multiscale context and can be applied to analyze $\Gamma(s,t)$[50]. However, this remains for future work.

4.3.3 Prediction

Another important aspect of extreme value analysis is prediction. This is a very challenging task, containing the prediction of the intensity and the occurrence of extremes. Detailed considerations at this point would go beyond the aims of this thesis. However, a couple of considerations that are relevant in current thesis context will be described in the following.
The classical EVA was designed to study the intensity of extremes and to provide a model that enables one to "predict" more extremal events in terms of return levels. Based on EVA, the proposed FEM-BV-EVA framework approaches the same issue and can be applied straightforwardly to the space-time clustering of extreme events. The resulting FEM-BV-EVA cdf can be used to describe the behavior of return levels and periods in the past in such a manner that possible trends become

visible. However, FEM-BV-EVA is not directly applicable for predictions of the return levels in the future. This task is hampered by two factors. First, the underlying dynamics is described by a set of local model parameters and a nonparametric, nonstationary switching process. Following, there is no closed formulation for the underlying dynamics of $\Gamma(s,t)$ which is required for prediction. To avoid this problem, one can try to find an extended set of covariates in order to resolve the observed dynamics, such that the optimal model is obtained for $K = 1$. An alternative approach is to consider $\Gamma(s,t)$ as a discrete process and to apply time-series analysis methods, e.g., the FEM-BV-Markov or FEM-BV-Causality methods[50,60], for studying the underlying dynamics. The resulting parametric model for $\Gamma(s,t)$ could then be used for making predictions about the affiliation to one of the local models. However, the prediction remains uncoupled from the true continuous timeline resulting in the second challenge.
Nevertheless, regression analysis of extremes allows the identification of the most significant covariates that influence the dynamics of the extremes. We are particularly interested in the identification of covariates that precede the occurrence of extremes. For instance, it has been found that the occurrence of US heat waves is likely preceded by 15-20 days "by a pattern of anomalous atmospheric planetary waves with a wavenumber of 5"[102], linked to the tropical heating. In this context, FEM-BV-EVA can be employed as a robust exploratory regression analysis tool for spatio-temporal extremes as will be demonstrated in Section 5.

4.4 Spatial Regularization

Spatial FEM was first introduced as an approach for spatio-temporal Markov regression analysis of discrete/categorical dynamical processes[38]. Another example of spatial FEM is the FEM-BV-EVA formulation, defined by (4.10). The resulting spatial FEM toolbox can be generalized by appropriately replacing the distance function $g_{Model}(.)$. Then, for an observed series and a set of covariates, the application of spatial FEM results in an optimal descriptive model (parametrized by K, the switching process $\Gamma(s,t)$, $C(N_T)$ and model parameters Θ). As previously discussed, the model parameters Θ are accessible for all locations, but in contrast the switching process is assigned to each location separately. The latter also implies that the estimation of $\Gamma(s,t)$ can be carried out separately for each location, i.e., in parallel, compare Algorithm 3, line 6. However, because we released a priori assumptions on $\Gamma(s,t)$, minimization of $\mathcal{L}(\Theta,\Gamma(s,t))$ with respect to $\Gamma(s,t)$ might become ill-posed. To handle the ill-posedness, the FEM framework in this thesis was extended towards spatial regularization by assuming persistent behavior in space. This assumption is reasonable, in particular in climate or weather research. For instance, the occurrence of spatially-persistent blocking anticyclones might be responsible for heat waves over the affected area[17,43]. The spatial persistency is incorporated into the FEM framework by adding an additional constraint on $\Gamma(s,t)$

$$||\gamma_k(s,t)||_{\mathcal{R}([s_1,s_{N_S}])} \leq C(N_S), \quad t = t_1,\ldots,t_{N_T}, \quad k = 1,\ldots,K, \tag{4.24}$$

where $\Gamma(s,t)$ is either in the space of functions with bounded variation or in the space of weakly differentiable functions, i.e., $\mathcal{R}([s_1,s_{N_S}]) = BV([s_1,s_{N_S}])$ or $\mathcal{R}([s_1,s_{N_S}]) = H^1([s_1,s_{N_S}])$, respectively.

The temporal dimension of $\Gamma(s,t)$ is discretized by applying the Finite Element method exploiting the sequential flow of time[58,59]. In order to discretize the spatial dimension in real applications, we have to account for the region under consideration. Considering geographical distances, e.g., the bee-line between locations, might suppress the topographical properties such as mountains and/or valleys. An alternative approach, for describing the spatial pairwise connectivity among all locations, is to refer to the pairwise correlation between the locations. For example, we can refer to the classical correlation or the cross-correlation matrix of the observed series among all locations in order to account for either the linear or the "time-lagged" relationship, respectively. Please note that the cross-correlation should be applied with caution when dealing with asynchronous measurements. Evaluation of cross-correlation does not account for an asynchronous behavior in the measurements. Consequently, the results refer to a time-lagged relationship where the time-lag has different time-scales and such no direct interpretation is possible. Further, in cases when the underlying relationship between the locations is nonlinear, cross-correlation might be misleading[8]. Instead, one can refer to the "event synchronization measure", which describes the nonlinear relationship between different locations with respect to event occurrence[76], compare Section 4.3.2.
Finally, both the spatio-temporal *BV* and $H1$ regularizations result in a linear or a quadratic constrained minimization problem with respect to $\Gamma(s,t)$. Compare Appendix A.3 for a detailed derivation. Standard methods, e.g., simplex method and constrained quadratic programming, can be applied. Both possibilities, *BV* and $H1$ regularizations, were implemented in the FEM toolbox. In order to solve the corresponding linear or quadratic problems FEM-BV-EVA gets use of the following standard toolboxes: (a) Matlab Optimization Toolbox[103] and (b) Gurobi Optimization Toolbox, which allows to solve the optimization problems in parallel[53].
The choice of an appropriate spatial "distance measure" is not obvious and corresponds in any case to some a priori assumption about the relationship among different locations. Further, incorporation of spatial regularization disables the parallel computation of $\Gamma(s,t)$ for each location s. In cases in which we refer to *BV* regularization, we must account for a further increase of the constrained linear problem dimensionality, as shown in Appendix 4.4. Those aspects lead to an increasing computational complexity, especially in the case of FEM-BV-EVA (where the theoretical aspects are based on a binary switching process). Therefor, for the practical applications examples in this thesis only the time regularization, i.e., without an additional space regularization, was considered. Application and adaptation of a computationally-feasible spatial regularization in the FEM-BV-EVA context remains for future work.

4.5 Conclusion

The FEM-BV-EVA toolbox was presented in this chapter. FEM-BV-EVA implementation deploys a gradient-free MCMC-based optimization technique and numerical solvers for large structured quadratic and linear problems with constraints. FEM-BV-EVA can be easily extended towards highly-scalable applications in HPC context. There are several possible levels of parallelization, for instance, the level of model parameter estimation: for a more wider exploration of the parameter space, the involved MCMC-based optimization can be started simultaneously with random

initializations. Further, in cases in which only time regularization is considered the switching process is uncoupled in space and thus in every iteration of the optimization algorithm $\Gamma(s,t)$ can be estimated for each location separately, that is, embarrassingly parallel, since no communication between different processors is required. In cases of spatial and temporal regularization the resulting linear or quadratic problem is then fully coupled and existent parallel libraries can be deployed. The resulting FEM-BV-EVA toolbox provides a computationally efficient framework for statistical spatio-temporal regression analysis of extremes and can be directly applied to real world problems. In the next chapter we demonstrate the performance of FEM-BV-EVA framework on test cases and real data.

5 Application

In this chapter, we demonstrate the performance of the FEM-BV-EVA framework on various examples and compare it to the state-of-the-art methods in statistical regression analysis of extremes. These include methods based on neuronal networks and smoothing regression. The comparison is performed according to the four criteria: (1) information content of the models (jointly measuring complexity and quality of the models), (2) robustness with respect to the systematically missing information, (3) computational complexity, (4) understandability/interpretability of the models. First, we demonstrate the performance of the univariate FEM-BV-EVA on test-cases and real applications. We will shown that parametric standard approaches provide biased results in cases when significant covariates are missing. Second, by comparing FEM-BV-GPD to methods based on smoothing regression, we emphasize the weakness of smoothing regression as discussed in Section 3.1.4. Finally, we demonstrate the performance of spatial FEM-BV-EVA on a real application analyzing the dynamics of threshold excesses of daily accumulated precipitation over 17 different locations in Switzerland.

5.1 Univariate FEM-BV-GEV

In this section we demonstrate the univariate FEM-BV-GEV methodology on two test cases and on the real data. The two test cases are used to investigate the robustness with respect to the systematically missing covariates, the approximation of nonstationary behavior and the computational performance of the framework (with respect to accuracy and computational time). In the real data example we analyze a series of block-maxima surface temperatures for locations Lugano (Switzerland) and Berlin (Germany). Further, to each application we apply the parametric GEV-CDN approach. GEV-CDN exploits a conditional density network (CDN) for nonlinear regression analysis based on time dependent covariates with constant weights and offsets[14], please see Section 2.1.1 for more details. In order to demonstrate the performance of parametric regression analysis in a presence of systematically missing observations, we apply the GEV-CDN approach to test cases. By applying the GEV-CDN approach to real data, we study either the underlying dynamics of extremes for the considered locations is rather nonlinear than nonstationary. The GEV-CDN analysis is performed using the package GEV-CDN provided in the statistical toolbox R[14,15]. The main tuning parameters of GEV-CDN are: the number of hidden neurons in the network (here denoted by N_H), the hidden layer transfer function (identity or logistic function) and the number of trials (to

avoid the local optima). An optimal configuration of the GEV-CDN with respect to these tuning parameters was determined according to the AIC_c criterion, i.e., deploying the same information criterium that was also used to find the optimal FEM-BV-GEV model.

5.1.1 Stationary Test Case

The first example is aiming to verify the regression analysis of block-maxima based only on resolved covariates. We would like to roughly mimic the true underlying dynamics of block-maxima in real meteorological applications. Therefore, as covariates we consider a linear trend, a periodic function with one year period and daily averaged measurements of the Total Solar Intensity (TSI)[47,48][1]. In general the TSI factor describes the total amount of the solar radiative energy that is hitting the Earth's upper atmosphere[48]. However, for this example we consider only a segment of the TSI measurements (staring from the year 1950) of length $T = 800$ and hence this factor is only responsible for more fluctuation in the generated block-maxima. Now, with covariates $\hat{U}_t = (u_1(t), u_2(t), u_3(t))$ defined by

$$u_1(t) = \frac{1}{400}t, \quad u_2(t) = sin(\frac{\pi}{2} + \frac{2\pi}{365}t), \quad u_3(t) = TSI, \tag{5.1}$$

we generate a series of block-maxima using following parametrization of the GEV model (3.1)

$$\mu(\hat{U}_t) = +1 - 5u_1(t) + 2u_2(t) + 1u_3(t), \tag{5.2}$$

$$\sigma(\hat{U}_t) = +2.1018 - 0.7132u_1(t) - 0.8203u_2(t) + 0.1356u_3(t), \tag{5.3}$$

$$\xi(\hat{U}_t) = -0.0627 - 0.4051u_1(t) + 0.0022u_2(t) - 0.0026u_3(t). \tag{5.4}$$

By assigning a relatively high coefficient to the factor $u_1(t)$ in (5.2) we stress the linear trend behavior in the dynamics of block-maxima. The coefficients in (5.3-5.4) were generated randomly. We use MATLAB function `gevrnd()` for sampling

$$X_t \sim GEV(\mu(\hat{U}_t), \sigma(\hat{U}_t), \xi(\hat{U}_t)) \text{ for } t = 1, \ldots, 800. \tag{5.5}$$

In the next step we split the covariates $\hat{U}_t$ into resolved and unresolved subsets, $U_t = (u_2(t), u_3(t))$ and $U_t^{un} = u_1(t)$, respectively, and apply FEM-BV-GEV and GEV-CDN methods for solving the inverse problem: for given X_t and U_t fit the model parameters to describe the distribution of X_t. We want to emphasize that by purposely missing the most relevant covariate, the linear trend, we would expect both methods to react on this issue by exploiting the nonlinearity in case of the GEV-CDN and the nonstationarity in case of the FEM-BV-GEV.

The FEM-BV-GEV is supplied with $K_{list} = \{1,2,3\}$, $C_{list} = \{2:1:6\}$ and following configurations: Number of annealing steps is fixed to 100, the maximal number of the subspace iterations is set to 150 and the minimization threshold to $Tol = 5.0e-05$. The GEV-CDN approach is configured with $N_H = \{1, 2:2:18\}$, hidden transfer function is the logistic function, number of trials is 100. The results are summarized in Table 5.1, featuring the minimal AIC_c values as been achieved by the respective methods. Resulting optimal models are $K = 3, C = 4$ for FEM-BV-GEV and $N_H = 12$ for CDN-GEV. The regression analysis of X_t based on resolved covariates

[1] The data were retrieved from `http://www.pmodwrc.ch/pmod.php?topic=tsi/composite/SolarConstant`.

	optimal Models for stationary test case			
	Settings	NLL	$\|M\|$	AIC_c
FEM-BV-GEV	$K = 3, C = 4$	1717.3	38	3514.4
GEV-CDN	$N_H = 12$	2111.6	75	4370.3

Table 5.1. *Optimal results for FEM-BV-GEV and GEV-CDN for the stationary test case. By using the original model parameters, we obtain the true negative log-likelihood $NLL_{true} = 1704.2$. As described above in the text, smaller values of NLL indicate the models with a better fit, whereas smaller values of AIC_c indicate more informative models.*

only was performed better by the FEM-BV-GEV than by the GEV-CDN approach (with a smaller NLL and a less total number of model parameters). As we can see from the left upper panel of the Figure 5.1, the optimal switching process $\Gamma^*(t)$, expressed by the affiliation $A(t) \in \mathbb{R}$ (with $A(t) = \{i : i = \text{argmax}\, \gamma_i^*(t) \text{ over } i = 1, \ldots, K\}$), assigns X_t to three different models. FEM-BV-GEV explicitly resolves the implicit linear trend in the systematically missing covariate U^{un} via a switching process that subsequently goes through three local parameter regimes. We can not compare the original and the resulting coefficients for the regression models explicitly. Instead, we evaluate the approximated $\mu^*(U_t)$, $\xi^*(U_t)$, $\sigma^*(U_t)$ according to the FEM-BV-GEV and the GEV-CDN models and compare them with the original evaluations according to (5.2-5.3-5.4). The comparison is shown in Figure 5.1. The top right, bottom left and bottom right panels represent the shape, the scale and the location parameters, respectively. The parameters obtained from the FEM-BV-GEV resolve the underlying trend very reliably. In contrast, due to the assumption that the neuron weights and offsets are constant, the GEV-CDN is not able to recover the impact of this missing covariate.

5.1.2 Nonstationary Test Case

In this section, we consider a nonstationary test case and use it to verify the accuracy and the performance of the FEM-BV-GEV. We generate X_t according to a mixture model with a nonstationary switching process

$$X_t \sim \gamma_1(t) GEV_1 + \gamma_2(t) GEV_2, \tag{5.6}$$

where GEV_1 is parametrized according to (5.2-5.3-5.4) and GEV_2 according to

$$\mu_2(\hat{U}_t) = -0.5 - 3u_1(t) + 0.5u_2(t) + 0.5u_3(t), \tag{5.7}$$

$$\sigma_2(\hat{U}_t) = +0.6729 + 0.0183u_1(t) - 0.4131u_2(t) + 0.1378u_3(t), \tag{5.8}$$

$$\xi_2(\hat{U}_t) = -0.0780 - 0.1398u_1(t) - 0.1608u_2(t) + 0.0266u_3(t). \tag{5.9}$$

We consider the same covariates $\hat{U}_t$ as in the stationary case. The nonstationary switching process $\Gamma(t) = (\gamma_1(t), \gamma_2(t))$ is generated artificially with $C = 6$ switches. Now, for given X_t and $U_t = (u_1(t), u_2(t), u_3(t))$, we apply FEM-BV-GEV and the GEV-CDN approach to capture the nonstationarity of (5.6). The FEM-BV-GEV is supplied with $K_{list} = \{1, 2, 3\}$, $C_{list} = \{2 : 1 : 14\}$, remaining configurations are the same as for the stationary test case. Also the configurations of the

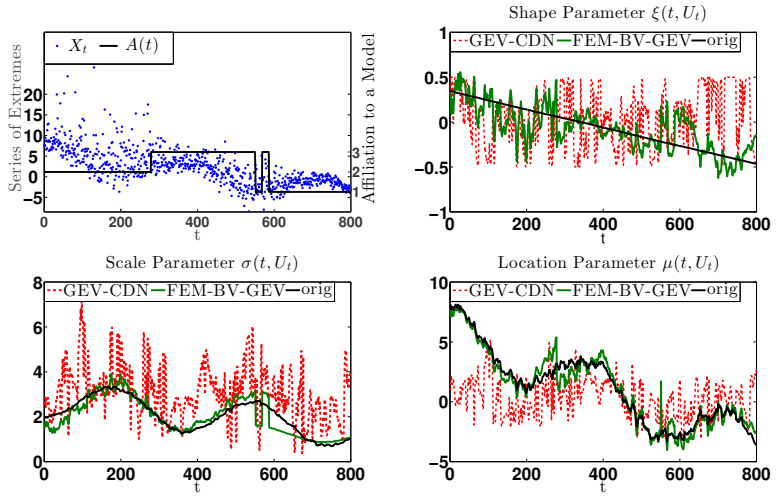


Figure 5.1. *Stationary test case: This figure shows the results for the application of FEM-BV-GEV and GEV-CDN to (5.5). The upper left figure shows the artificially generated series of extremes X_t vs. the optimal switching process $\Gamma^*(t)$, expressed by the affiliation vector $A(t)$. The remaining panels represent the evaluation of the shape, scale and location parameters according to the original (black solid line), optimal FEM-BV-GEV (dashed dotted line) and GEV-CDN (grey solid line) parameters.*

GEV-CDN approach do not change. Because we provide the full information, $U_t = \hat{U}_t$, to both

	optimal Models for nonstationary test case			
	Settings	NLL	$\|M\|$	AIC_c
FEM-BV-GEV	$K = 2, C = 12$	1204.1	37	2485.9
GEV-CDN	$N_H = 7$	1254.5	52	2620.3

Table 5.2. *Optimal results for FEM-BV-GEV ($K = 2, C = 12$) and GEV-CDN ($N_H = 7$) for the nonstationary test case. By using the original model parameters, we obtain the true negative log-likelihood $NLL_{true} = 1228.9$.*

methods, they both perform well, compare Table 5.2 and Figure 5.2. FEM-BV-GEV approximates the dynamics of X_t with less parameters and a smaller NLL. The inconsistency of the number of switches in $\Gamma^*(t)$ with $C = 12$ (compare Figure 5.2, upper left panel) and the original $\Gamma(t)$ with $C = 6$ can be neglected due to the relatively large confidence intervals for $\Gamma^*(t)$ and Θ^* (note displayed here, we refer the interested reader to[64]). Also the GEV-CDN captures the regime switches and the underlying trend in parameters, compare Figure 5.2. The computational performance of FEM-BV-GEV and GEV-CDN is compared by considering the CPU time for one annealing step

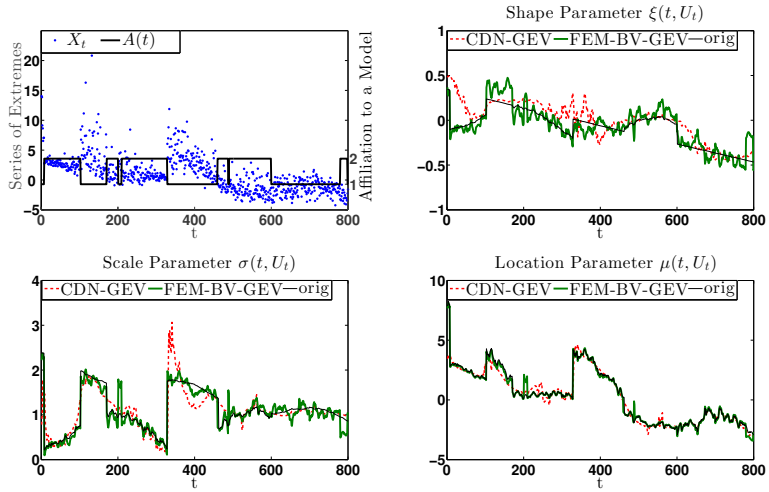


Figure 5.2. *Non-stationary test case: This figure shows the results for the application of FEM-BV-GEV and GEV-CDN to (5.6). The upper left figure shows the artificial generated series of extremes X_t vs. the optimal switching process $\Gamma^*(t)$, expressed by the affiliation vector $A(t)$. The remaining panels represent the evaluation of the shape, scale and location parameters according to original (black solid line), optimal FEM-BV-GEV (dashed dotted line) and GEV-CDN (grey solid line) parameters.*

dependent on the increasing number of parameters (configurations do not change). The results are shown in Figure 5.3. The plots contain the average CPU time over 100 runs. FEM-BV-GEV obviously outperforms the GEV-CDN approach with respect to the computational performance for the growing number of parameters (e.g., corresponding to the larger number of involved covariates, hidden neurons, etc.)

5.1.3 Real Data Application

In this section we apply FEM-BV-GEV and GEV-CDN to real data, where we do not have a priori the knowledge about the underlying dynamics. Moreover, we have to account for unresolved covariates because it is not clear a priori which weather/climate covariates are potentially relevant for the analyzed data and which are not. In the following, we consider historical daily records of temperature from 1950-01-01 till 2011-01-01 for locations Lugano, Switzerland ($46°N$, $8.9667°E$) and Berlin, Germany ($52.4649°N$, $13.3017°E$)[100]. Data were retrieved from NOAAs National Climatic Data Center (NCDC) web-page[2]. We restrict the data to this period because observations for some of the involved covariates are available starting from 1950 only. We consider the following set of covariates:

[2]Data were retrieved from `http://gis.ncdc.noaa.gov/map/cdo/?thm=themeDaily`.

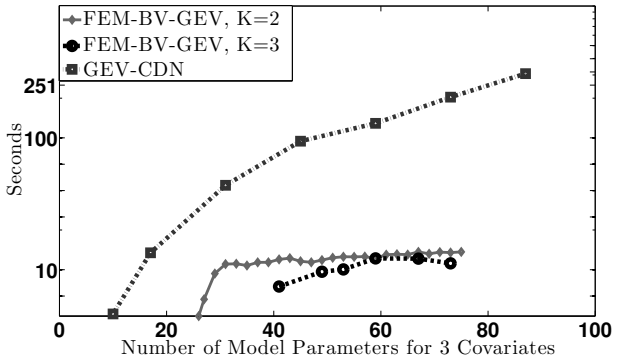


Figure 5.3. *Non-stationary test case: This figure compares the computational time performance of FEM-BV-GEV (diamonds marker for $K = 2$ and circles for $K = 3$) and GEV-CDN (squared markers) using logarithmic time scale (seconds). The number of covariates is fixed, thus the increase of number of model parameters is due to increasing of C for FEM-BV-GEV and number of hidden neurons for GEV-CDN.*

1. Arctic Oscillation (AO)[2],
2. North Atlantic Oscillation (NAO)[3],
3. Total Solar Irradiance (TSI), averaged over one day[47,484],
4. ENSO, represented trough mean sea surface temperature anomalies in the Nino3.4 region[106],
5. $\log(CO2)$, with logarithmic dependence according to[87],
6. Seasonal Periodical Phase: $\mathcal{P}er_I = \sin(\frac{2\Pi}{365}t)$,
7. Seasonal Periodical Phase: $Per_{II} = \sin(\frac{3}{2.1}\Pi + \frac{2\Pi}{365}t)$,
8. Madden-Julian Oscillation (MJO), contains the first two empirical orthogonal functions[5].

In order to interpret and compare the relative influences of covariates on trends in model parameters, $U_t \in \mathbb{R}^8$ is scaled to $[-1,1]$ according to (5.10)

$$u_p(t) = 2\frac{u_p(t) - max(u_p)}{max(u_p) - min(u_p)} + 1, \quad \text{for } p = 1,\ldots,8, \tag{5.10}$$

where $max(u_p)$ and $min(u_p)$ are the maximal and the minimal values of the covariate p over the time from 1950-01-01 till 2011-01-01. Further, before extracting 30 days block-maxima we remove the

[3]Data were retrieved from `ftp://ftp.cpc.ncep.noaa.gov/cwlinks/`.
[4]Data were retrieved from `http://www.pmodwrc.ch/pmod.php?topic=tsi/composite/SolarConstant`.
[5]Data were retrieved from `http://cawcr.gov.au/staff/mwheeler/maproom/RMM/`.

seasonal trend in the data. For this, we estimate the yearly trend by averaging over all values corresponding to the same day and month and then subtract the yearly trend from the data. The dedicated series of block-maxima for each location contains 742 maxima in the observed period. For the regression analysis we refer to covariates measured at the same time steps when the maxima in each block is observed.
In the following, we want to extract the most significant combination of covariates out of all possible, in total 255. For this task, we use the FEM-BV-GEV framework with following configurations: $K_{list} = \{1,2,3\}$, $C_{list} = \{5:5:100\}$, number of annealing steps is fixed to 100, the number of the subspace iterations is set to 250 and the minimization threshold to $Tol = 5.0e-05$. Then, according to the minimal AIC_c, we obtain for each location the optimal model including the most significant combination, denoted by u^*_{comb}. For location Lugano u^*_{comb} is $[NAO, log(CO_2), Per_I, Per_{II}]$, and for location Berlin $u^*_{comb} = [AO, NAO, Per_I]$. In the second step, we compare the FEM-BV-GEV and GEV-CDN applied to two different settings: (a) we provide the complete set of optimal covariates for the regression analysis u^*_{comb} and (b) we provide an incomplete set $\hat{u}^*_{comb} = [NAO, Per_I, Per_{II}]$ and keep back $log(CO_2)$ for location Lugano, and $\hat{u}^*_{comb} = [NAO, Per_I]$ and keep back AO for location Berlin.

	Location Lugano with $u^*_{comb} = [NAO, log(CO_2), Per_I, Per_{II}]$			**Location Lugano with** $\hat{u}^*_{comb} = [NAO, Per_I, Per_{II}]$		
	NLL	$\|M\|$	AIC_c	NLL	$\|M\|$	AIC_c
FEM-BV-GEV	1573.9	70	3302.6	1608.9	64	**3358.0**
GEV-CDN	1494.0	115	**3260.6**	1672.9	45	3441.6

Table 5.3. *Comparison of FEM-BV-GEV and GEV-CDN according to AIC_c model selection criteria for locations Lugano according to the resolved and unresolved covariates. The optimal models for resolved covariates are: FEM-BV-GEV $K = 2, C = 40$ and GEV-CDN $N_H = 14$. The optimal models for unresolved covariates are: FEM-BV-GEV $K = 2, C = 40$ and GEV-CDN $N_H = 6$.*

	Location Berlin with $u^*_{comb} = [AO, NAO, Per_I]$			**Location Berlin with** $\hat{u}^*_{comb} = [NAO, Per_I]$		
	NLL	$\|M\|$	AIC_c	NLL	$\|M\|$	AIC_c
FEM-BV-GEV	1642.8	109	**3541.5**	1675.6	89	**3553.8**
GEV-CDN	1781.8	45	3659.5	1792.7	39	3667.8

Table 5.4. *Comparison of FEM-BV-GEV and GEV-CDN according to AIC_c model selection criteria for locations Berlin according to the resolved and unresolved covariates. The optimal models for resolved covariates are: FEM-BV-GEV $K = 2, C = 85$ and GEV-CDN $N_H = 6$. The optimal models for unresolved covariates are: FEM-BV-GEV $K = 2, C = 70$ and GEV-CDN $N_H = 6$.*

Note that u^*_{comb} is significant according to the FEM-BV-GEV approach and one could argue that for the GEV-CDN approach an other set of covariates could be more important[6]. In return, in real applications we will never know a priori which covariates may be important and in any case the complete set of potentially-relevant covariates will never be available a priori. The results for settings (a) and (b) are shown in Table 5.3 for location Lugano and in Table 5.4 for location Berlin. The optimal GEV-CDN model is chosen out from $N_H = \{2:2:16\}$. Additionally, we compute the expectation value of block-maxima with the corresponding quantiles for both locations and discuss its behavior. Comparing the optimal FEM-BV-GEV and GEV-CDN models, we can conclude that in the case when the set of covariates is "complete" the nonlinear GEV-CDN provides a better description of the block maxima for location Lugano in terms of information theory (as measured by AIC_c). Consequently, the underlying dynamics is rather nonlinear than nonstationary. In contrast, FEM-BV-GEV provides a better description of block-maxima for location Berlin. Moreover, in the particular case when some information is "missing", the nonstationary FEM-BV-GEV approach approximates the underlying dynamics better by reflecting the unresolved modes trough the switching process for both of the considered cases (Berlin and Lugano).

Postprocessing

In the following, we discuss the postprocessing for location Lugano and Berlin according to the optimal FEM-BV-GEV and GEV-CDN models. Local linear FEM-BV-GEV model allows direct interpretation of the influence of covariates on the dynamics of GEV parameters, compare Table 5.5 and Table 5.6. For the GEV-CDN approach the fitted parameter dynamics is not easy to interpret and understand; we obtain a matrix of weights, and have to evaluate the parameters according to the nonlinear transfer functions (being a logistic functions in our particular case). The identification of these factors is physically-meaningful. Positive phase of *AO* causes dry and hot conditions in Mediterranean regions. *AO* has a direct impact on atmospheric circulation blocking events: it induces a ridge of high pressure in the mid latitude jet streams that can cause persistently high temperatures (as well as cold conditions)[55]. Positive phases of *NAO* cause warm, wet winters in Northern and dry winters in Southern Europe. The anthropogenic influence of CO_2 concentration, and so $\log(CO_2)$ holds a positive trend with an oscillating dynamics (with maximum value in May and minimum in October)[66]. The relevance of Per_I and Per_{II} points to a strong seasonal dependence of block-maxima in both locations (this is obvious since we consider monthly maxima). In order to study the long-term trend in distribution of block-maxima, we evaluate the nonstationary expectation value

$$\mathbb{E}_{K=2}[X_t,t] = \sum_{i=1}^{2} \gamma_i(t) \left(\mu_i(U_t) + \sigma_i(U_t) \frac{\tilde{\Gamma}(1-\xi_i(U_t))-1}{\xi_i(U_t)} \right), \tag{5.11}$$

$$\mathbb{E}_{CDN}[X_t,t] = \mu_{CDN}(U_t) + \sigma_{CDN}(U_t) \frac{\tilde{\Gamma}(1-\xi_{CDN}(U_t))-1}{\xi_{CDN}(U_t)}, \tag{5.12}$$

[6]Application of GEV-CDN to identify the most significant combination of covariates is not feasible because of prohibitively high computational cost to get through all 255 covariates combinations (please see Figure 5.3 for computational cost comparisons of the two methods).

with $t = 1,\ldots,742$. Here $K = 2$ corresponds to FEM-BV-GEV (with parametrization according to (3.26)), CDN to GEV-CDN and $\tilde{\Gamma}$ denotes the gamma function. Figures 5.4 and 5.5 show the results according to FEM-BV-GEV and GEV-CDN. The 0.99- and 0.10-Quantiles are the confidence intervals, containing 89% of the distribution. In particular, the 0.99-Quantile corresponds to the 100-year return level. According to the FEM-BV-GEV results, the mean for location Lugano

	Model Parameters for location Lugano with $u^*_{comb} = [NAO, log(CO_2), Per_I, Per_{II}]$														
	μ_0	μ_1	μ_2	μ_3	μ_4	σ_0	σ_1	σ_2	σ_3	σ_4	ξ_0	ξ_1	ξ_2	ξ_3	ξ_4
θ_1^*	4.29	-0.19	1.97	0.74	-1.39	1.99	-0.10	0.05	0.21	-0.61	-0.37	0.39	0.40	-0.15	-0.09
θ_2^*	3.92	0.78	-2.12	1.70	-0.34	1.71	0.60	-0.17	0.39	-0.42	-0.05	0.16	0.03	-0.19	-0.09

Table 5.5. *The table contains optimal parameters θ_1^* and θ_2^* for location Lugano (the values are rounded to two places behind the decimal point).*

	Model Parameters for location Berlin with $u^*_{comb} = [AO, NAO, Per_I]$											
	μ_0	μ_1	μ_2	μ_3	σ_0	σ_1	σ_2	σ_3	ξ_0	ξ_1	ξ_2	ξ_3
θ_1^*	4.73	1.89	0.1	0.38	2.59	-0.52	-0.03	0.6873	-0.22	-0.27	-0.40	0.17
θ_2^*	8.43	2.00	-1.13	0.59	2.15	-0.21	0.69	-0.12	-0.32	0.01	-0.10	0.00

Table 5.6. *The table contains optimal parameters θ_1^* and θ_2^* for location Berlin (the values are rounded to two places behind the decimal point).*

shows a slightly negative trend in the second model. After the 1980-ies the first model dominates, here $log(CO_2)$ has a positive influence, inducing an increasing trend in block-maxima. In contrast, according to GEV-CDN model, there is no obvious trend: however, the confidence intervals for the GEV distribution increase in the last ten years. For location Berlin the trend of the expectation value is separated according to two FEM-BV-GEV models, one model corresponds to higher block-maxima. The GEV-CDN model averages this dynamics, and provides an unchanging behavior with some (few) outliers.

5.2 Univariate FEM-BV-GPD

In this section we demonstrate the robustness of FEM-BV-GPD approach with respect to systematically missing covariates on a test case and on real data. Performance of the introduced FEM-BV-GPD is compared to the gamGPD approach, implemented in the statistics toolbox R [18,88,99]. The gamGPD framework is used to resolve the involved nonstationarity of the bias/off-set according to (3.19). The framework offers following possibilities: (a) to choose between cubic or thin plate spline basis, denoted by $bs = cr$ or $bs = tp$, respectively, (b) to run over different dimensions of the bases for the smooth spline as defined in (3.19), denoted by q and (c) to compute a spline with and without regularization, denoted as $fx = fs$ or $fx = tr$, respectively [88,111–113]. The optimal model

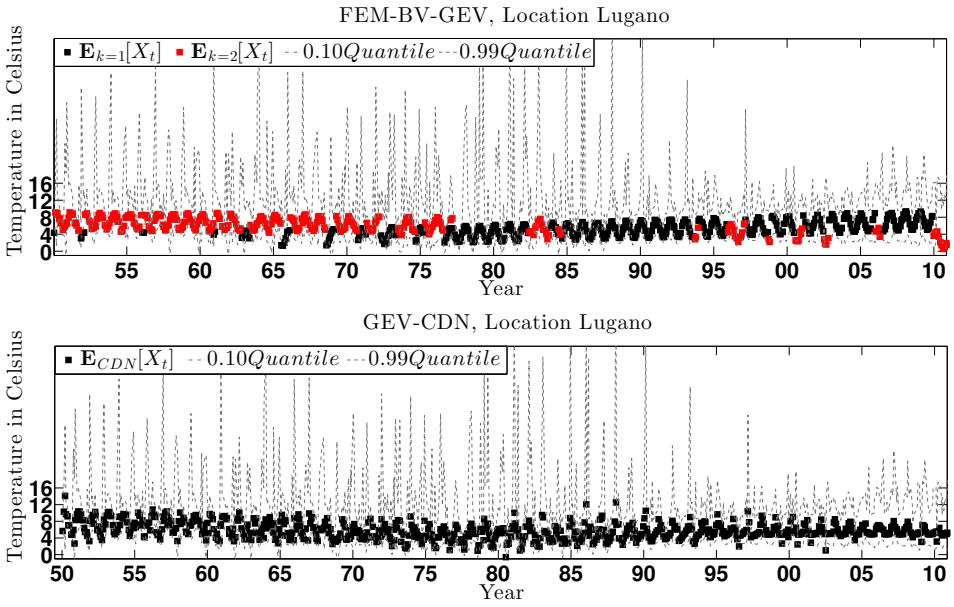


Figure 5.4. *Location Lugano: The figure contains the plot of the expectation value for the optimal FEM-BV-GEV model, $K = 2, C = 40$.*

is chosen with respect to the $AIC/AICc$ criterion (4.5-4.6). In order to interpret the differences in obtained $AIC/AICc$ values, we refer to the Akaike model weights as defined by (4.7).

5.2.1 Nonstationary Test Case

In the following, we construct a test case where the underlying dynamics of extremes is governed by a discrete switching process. We proceed in two steps. First, we generate an artificial series of threshold excesses X_t according to a mixture model with a nonstationary switching process

$$X_t \sim \gamma_1(t) GPD\left(\sigma_1(\hat{U}_t), \xi_1(\hat{U}_t)\right) + \gamma_2(t) GPD\left(\sigma_2(\hat{U}_t), \xi_2(\hat{U}_t)\right), \text{ for } t = 1, \ldots, 1000. \tag{5.13}$$

The model parameters depend on a fixed set of covariates: $\hat{U}_t = (u_1(t), u_2(t), u_3(t))$ with

$$u_1(t) = \frac{1}{400}t, \quad u_2(t) = sin(\frac{2\pi}{500}t), \quad u_3(t) = TSI, \tag{5.14}$$

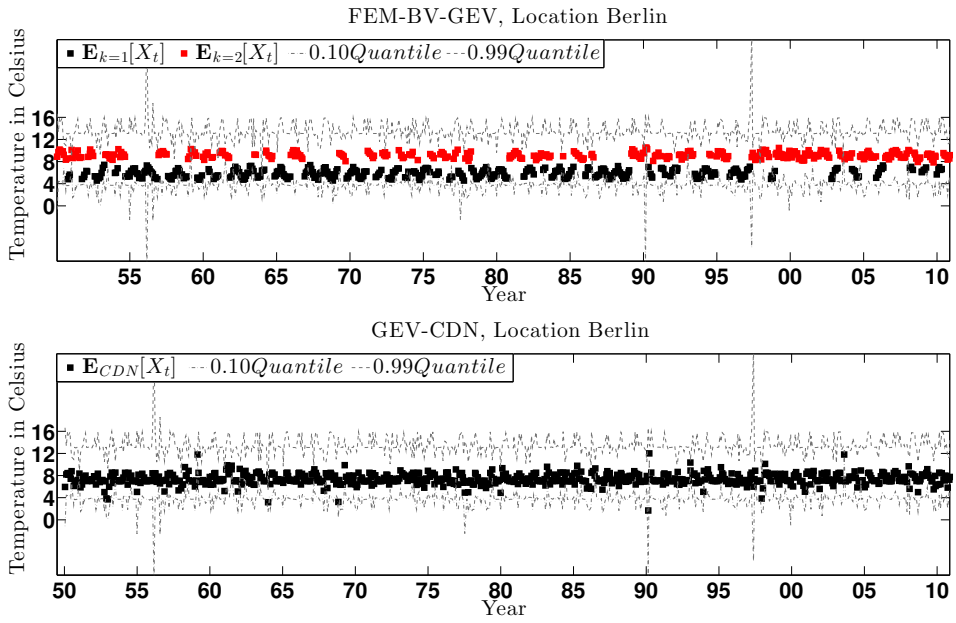


Figure 5.5. *Location Berlin: The figure contains the plot of the expectation value for the optimal FEM-BV-GEV model, $K = 2, C = 85$.*

where TSI (Total Solar Intensity)[47,487] is responsible for more fluctuation in the data. The series of threshold excesses is generated using the following parametrization

$$\sigma_1(\hat{U}_t) = 3.0 + 3.0u_1(t) - 0.5u_2(t) + 0.1u_3(t), \tag{5.15}$$
$$\xi_1(\hat{U}_t) = 0.2 + 0.001u_1(t) - 0.1u_2(t) - 0.01u_3(t), \tag{5.16}$$
$$\sigma_2(\hat{U}_t) = 1.0 + 0.001u_1(t) - 0.5u_2(t) + 0.1u_3(t), \tag{5.17}$$
$$\xi_2(\hat{U}_t) = 0.1 + 0.1u_1(t) - 0.001u_2(t) + 0.01u_3(t). \tag{5.18}$$

We use MATLAB function `gprnd()` for sampling. Second, we ignore the knowledge about the hidden switching process and apply FEM-BV-GPD and gamGPD for solving the inverse problem: for given X_t and $U_t = \hat{U}_t$ find the optimal model parameters.
The FEM-BV-GPD is supplied with $K_{list} = \{1,2,3\}$, $C_{list} = \{2:1:20\}$, $\lambda = 0$ (i.e., no Lasso regularization) and following configurations: number of annealing steps is fixed to 150, the maximal number of the subspace iterations is set to 500 and the convergency criterion is set to 1.0×10^{-3}. We deploy the gamGPD framework for two purposes: (a) for linear regression with a nonstationary offset term and (b) nonlinear regression by expressing the model parameters as additive regression models. We denote the first purpose as $gamGPD_I$, and the second one as $gamGPD_{II}$. The involved additive regression models are resolved by default spline functions as described in [111–113]. For both,

[7] The data were retrieved from `http://www.pmodwrc.ch/pmod.php?topic=tsi/composite/SolarConstant`.

the gamGPD framework is configured with $bs_{list} = \{cr, tp\}$, $q_{list} = 4:25$ and $fx_{list} = \{tr, fs\}$. The maximal number of iterations for the involved backfitting Algorithm 1 is set to 1000 and the convergency criterion is set to 1.0×10^{-3}. The results are summarized in Table 5.7. FEM-BV-GPD

	optimal Models			
Model M	Settings	NLL	AIC	$\rho(M)$
FEM-BV-GEV	$K = 2, C = 6$	2096.8	**4239.7**	0.99
$gamGPD_I$	bs =cr, $q = 15$, $fx = fs$	2099.4	4270.8	1.76×10^{-6}
$gamGPD_{II}$	bs =cr, $q = 7$, $fx = fs$	2092.8	4285.6	1.08×10^{-9}

Table 5.7. *Optimal results for FEM-BV-GPD and gamGPD for the test case in Section 5.2.1. For the original model parameters the true negative log-likelihood is $NLL_{true} = 2064.5$. Smaller values of NLL indicate the models with a better fit, whereas smaller values of AIC indicate more informative models. The values of the model weights $\rho(M)$, estimated according to (4.7), are rounded to two places behind the point.*

outperforms $gamGPD_I$ and $gamGPD_{II}$ in terms of the posterior model weight. As can be seen from the Table 5.7, $gamGPD_{II}$ describes the underlying dynamics by a bigger set of model parameters and thus overfits the data in a sense of AIC. The unregularized gamGPD, i.e., $fx = tr$, is an ill-posed inverse problem. The optimal result for this setting is obtained for $q = 22$, $bs = cr$ with $NLL = 2050.5$ and $AIC = 4270.8$. However, the regression coefficients are not bounded: the values of the scale parameter are in the interval $[-219.733, 1941.244]$ and the values of the shape parameter in $[-1545.644, 1368.124]$. In the following, we refer to the regularized formulation only.
We can not compare the original and the resulting coefficients for gamGPD regression model explicitly. Instead, we evaluate the approximated model parameters according to the optimal FEM-BV-GPD and $gamGPD_I$ models and compare them with the original evaluations according to (5.15)-(5.18). The comparison is shown in Figure 5.6, lower left and right panels represent the shape and the scale parameters, respectively. In the upper right panel we see $\Gamma^*(t)$ expressed by the affiliation $A(t) \in \mathbb{R}$ (with $A(t) = \{i : i = \text{argmax}\, \gamma_i^*(t) \text{ over } i = 1, \ldots, K\}$). Concluding, we can state that the switching process and the model parameters obtained from the FEM-BV-GPD resolve the underlying dynamics of both model parameters very reliably. The $gamGPD_I$ approach approximates the underlying nonstationarity well for the scale parameter, but performs worse in approximating the dynamics of the shape parameter.

5.2.2 Real Data Application

In this section, FEM-BV-GPD and gamGPD are applied to real data. We consider daily accumulated precipitation from 1981-01-01 till 2013-01-01 for Lugano, Switzerland ($46^\circ N$, $8.9667^\circ E$) [8]. As threshold excesses we define the peaks over the 0.95 quantile of the total rainfall. The dedicated series of threshold excesses contains 539 events. We refer to the following set of covariates for the

[8] Data were retrieved from MeteoSwiss (`www.meteoswiss.admin.ch`).

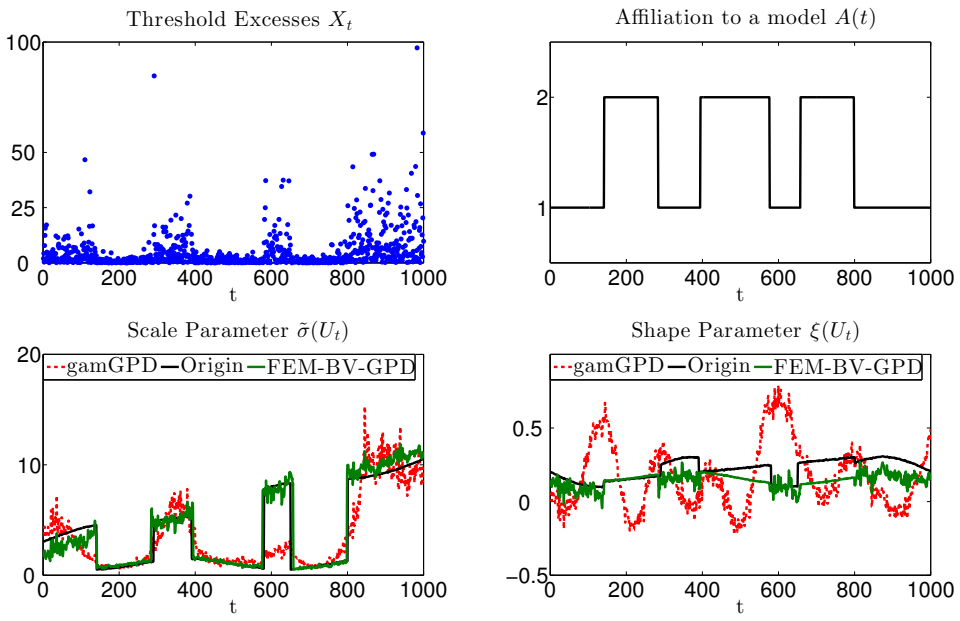


Figure 5.6. *This figure shows the results for the application of FEM-BV-GPD and gamGPD to X_t and U_t described by (5.13). The upper left figure shows the artificially generated threshold excesses X_t, the upper right the optimal switching process $\Gamma^*(t)$, expressed by the affiliation vector $A(t)$. The remaining panels represent the evaluation of the shape and scale parameters according to the optimal gamGPD (solid gray line), the original (dark grey solid line) and the optimal FEM-BV-GPD (dash-dotted black line) models.*

regression analysis:

1. Wind at location Lugano[8],
2. Temperature at location Lugano[8], in the following abbreviated by Temp,
3. Total Solar Irradiance (TSI), averaged over one day[47,489],
4. Seasonal Periodical Phase: $\mathcal{P}er_I = \sin(\frac{2\Pi}{365}t)$,
5. Seasonal Periodical Phase: $Per_{II} = \sin(\frac{3}{2.1}\Pi + \frac{2\Pi}{365}t)$,
6. North Atlantic Oscillation (NAO)[10],
7. Arctic Oscillation (AO)[10],

[9]Data were retrieved from `http://www.pmodwrc.ch/pmod.php?topic=tsi/composite/SolarConstant`.
[10]Data were retrieved from `ftp://ftp.cpc.ncep.noaa.gov/cwlinks/`.

8. ENSO, represented trough mean sea surface temperature anomalies in the Nino3.4 region [106]. We want also to analyze a time delayed influence of ENSO, with a time lag of 3, 12 and 24 month [67,84,93], each denoted in the following as $ENSO_3$, $ENSO_{12}$ and $ENSO_{24}$,

9. $\log(CO2)$, with logarithmic dependence according to [87],

The covariates $U_t \in \mathbb{R}^{11}$ are scaled with $u_p(t) \in [-1,1]$ for $p = 1,\ldots,11$, so we can interpret their relative influences on trends in model parameters. For regression analysis, the covariates are taken at the same time steps as the threshold excesses are observed, minus the corresponding time lags, respectively. Further, we aim to incorporate only uncorrelated covariates. By deploying the Pearson's linear correlation coefficients and considering all covariates with $|corr(u_v,u_p)_{v,p=1,\ldots,13}| > 0.33$ as correlated, we can reduce the set of involved covariates to

$$\hat{U}_1(t) = \{Wind, Temp, TSI, NAO, ENSO_3, ENSO_{24}\}. \tag{5.19}$$

The FEM-BV-GPD models are inferred for all combinations of $K_{list} = \{1,2,3\}$, $C_{list} = \{10:10:100\}$ and $\lambda_{list} = \{0, 1.0\times10^{-5}, 1.0\times10^{-4}, 1.0\times10^{-2}, 1.0, 10, 100\}$. The number of annealing steps is fixed to 150, the maximal number of the subspace iterations is set to 1000 and the convergency criterion is set to 1.0×10^{-3}. Here, we refer to regularized gamGPD only. The dimensions of the bases for the smooth spline are set to $k = 4:10$ and the convergency criterion for the shape parameter is set to 1.0×10^{-2}. The remaining configurations are the same as for the test case. Note that the application of $gamGPD_{II}$ failed with an error output "This most likely comes from non-finite weights in the call to adjustD2()". The function "`adjustD2()`" is responsible for the estimation of the second derivative. This case emphasizes the robustness of the gradient-free MCMC-based optimization method deployed for a numerical minimization of the constrained FEM-BV-GPD problem as discussed in Section 4.2. The results are summarized in Table 5.8.

	optimal Models			
Model M	Settings	NLL	$AICc$	$\rho(M)$
FEM-BV-GPD	$\hat{U}_1(t)$, $K=2, C=40, \lambda=0.1$	2017.71	**4196.67**	0.99
$gamGPD_I$	$\hat{U}_1(t)$, $k=4$, $fx=fs$, $bs=tp$	2089.81	4221.24	4.61×10^{-6}

Table 5.8. *Results for the statistical regression analysis of threshold excesses. Smaller values of NLL indicate the models with a better fit, whereas smaller values of AICc indicate more informative models. The values of the model weights, estimated with respect to (4.7), are rounded to two places behind the point.*

FEM-BV-GPD provides a more descriptive and informative model for threshold excesses as measured by NLL and $AICc$, respectively, and has the maximal posterior model weight, see Table 5.8. Where $K = 2$ points out the intrinsic nonstationarity of the model induced by the presence of missing covariates in the involved regression model. The evaluation of the corresponding model parameters is shown in Figure 5.7. The relative influence of the covariates on FEM-BV-GPD model parameters is summarized in Table 5.9. The appropriate clusters clearly distinguish between two different

dynamic behaviors of the threshold excesses.
The dynamics of extremes spends more time in the first cluster with large values for the scale parameter and dominating negative values for the shape parameter, implying a bounded upper tail of the corresponding GPD distribution. The first cluster is clearly associated with a strong seasonal behavior, because of the strong influence coming from the temperature, compare Table 5.10. The strong correlation with the wind refers probably to heavy rain during thunderstorms. In contrast, the dynamics of threshold excesses in the second cluster exhibits small values for the scale and positive values for the shape parameter, implying a heavy tail of the corresponding GPD distribution and a higher probability of extremes. The behavior of the shape parameter reveals a periodicity associated with the TSI index, rather then with seasonal effect, see Table 5.10. In particular, negative scaled TSI index, i.e., low solar activity, increases the value of the shape parameter and so the probability for larger threshold excesses.

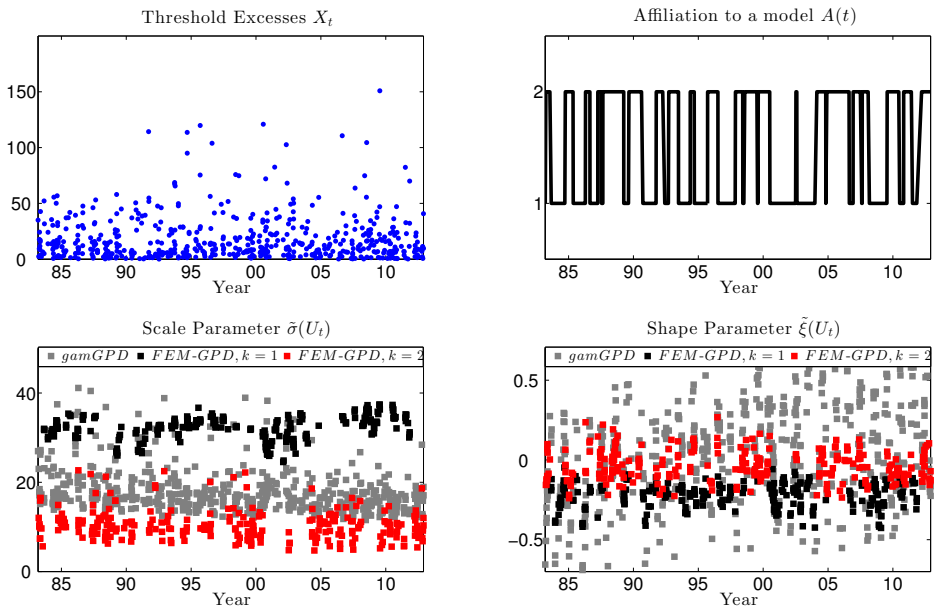


Figure 5.7. *This figure shows the optimal results for the statistical regression analysis of extreme precipitation over Lugano for the period 1981 to 2013 performed by FEM-BV-GPD (in the figure abbreviated as FEM-GPD) and gamGPD. In this figure we projected the results to the real time scale and thus, chose a discrete representation of the model parameters for a better visualization. The top left figure shows the threshold excesses X_t, the top right demonstrates the optimal affiliations $A(t)$ as computed from the optimal switching process $\Gamma^*(t)$ of FEM-BV-GPD. The remaining panels represent the evaluation of the shape and scale parameters according to the optimal gamGPD (gray markers) and the optimal FEM-BV-GPD (black markers) models.*

	Relative influence of the covariates on GPD parameters						
	off-set	Wind	Temp.	TSI	NAO	$ENSO_3$	$ENSO_{24}$
ξ_1^*	-0.0675	0.2238	0.2596	0.0131	0.0154	-0.0391	0.0206
ξ_2^*	0.3330	0.5144	0.2032	-0.0453	0.0013	-0.0146	-0.0053
σ_1^*	30.1295	-0.9715	6.6261	-4.5324	0.4571	0.0076	-0.4093
σ_2^*	27.1608	23.1403	-8.8449	0.0707	-0.2810	0.2425	0.2004

Table 5.9. *The table contains optimal FEM-BV-GPD model parameters for threshold regression analysis of extreme rainfall for location Lugano.*

	Pearson's linear correlation coefficient					
	Wind	Temp.	TSI	NAO	$ENSO_3$	$ENSO_{24}$
ξ_1^*	**0.6048**	**0.9183**	0.0275	-0.0176	0.0391	0.0715
ξ_2^*	**0.8299**	**0.7192**	-0.0569	-0.1061	-0.0417	-0.0601
σ_1^*	0.0476	0.6506	**-0.7948**	-0.1021	0.0808	**0.3088**
σ_2^*	**0.7503**	**-0.4576**	0.0711	0.0253	0.0795	-0.0170

Table 5.10. *The table contains the Pearson's linear correlation coefficients between the optimal FEM-BV-GPD model parameters and the covariates.*

5.3 Spatial FEM-BV-GPD

In this section, we demonstrate the performance of the spatial FEM-BV-GPD on real data. We consider daily accumulated precipitation over 17 different locations in Switzerland from 1981-01-01 till 2013-01-01[11]. For each single location we estimate the threshold as the 0.95 quantile of the accumulated rainfall and define accordingly the threshold excesses. As a result, each location contains different number of involved threshold excesses. For spatial-temporal regression analysis we refer to the a set of local covariates, measured at each location[11] and a set of global covariates, being the same for each locations. The set of local covariates contains

- Wind; hourly maxima,
- Temperature at 2 meters above the ground; hourly average,
- Humidity at 2 meters above the ground; hourly average.

The set of global covariates is the same as considered in Section 5.2.2. In line with Section 5.2, we reduce the full set of covariates to the set of uncorrelated only:

$$\hat{U}(s,t) = \{Wind, Temp, TSI, NAO, ENSO_3, ENSO_{24}\}. \tag{5.20}$$

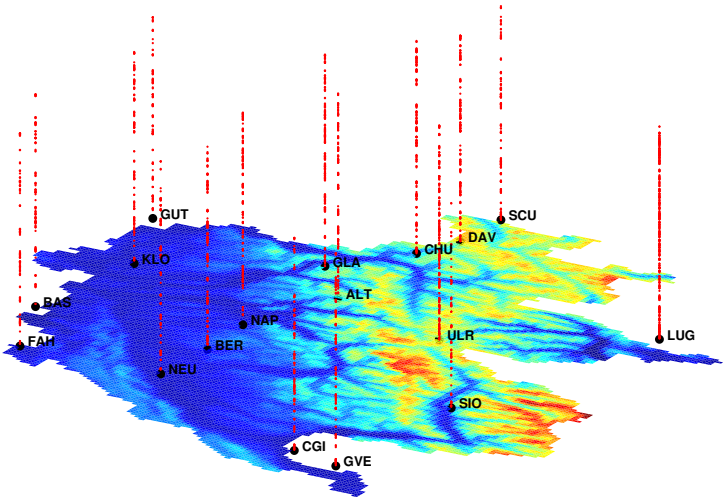
[11] Data were retrieved from "MeteoSwiss" (`www.meteoswiss.admin.ch`).

We apply the spatial FEM-BV-GPD approach with following configurations: $K_{list} = \{1,2,3\}$, $C_{list} = \{10:10:100\}$ and $\lambda_{list} = 0$. The number of annealing steps is fixed to 400, the maximal number of the subspace iterations is set to 2000 and the convergency criterion is set to 1.0×10^{-3}. The optimal result was obtained for $K = 2$ and $C = 50$. That is, FEM-BV-GPD describes the underlying dynamics of threshold excesses over the 17 different location by two different locally stationary GPD-regression models and a nonstationary switching process for each single location. The maximal number of switches for each of the locations is $C = 50$.
The optimal FEM-BV-GPD parameters and their confidence intervals are presented in Table 5.11. We estimated the confidence intervals as the values of the corresponding empirical Fisher Information matrix, for this purpose the MATLAB function `mlecov()` was deployed. The realization of

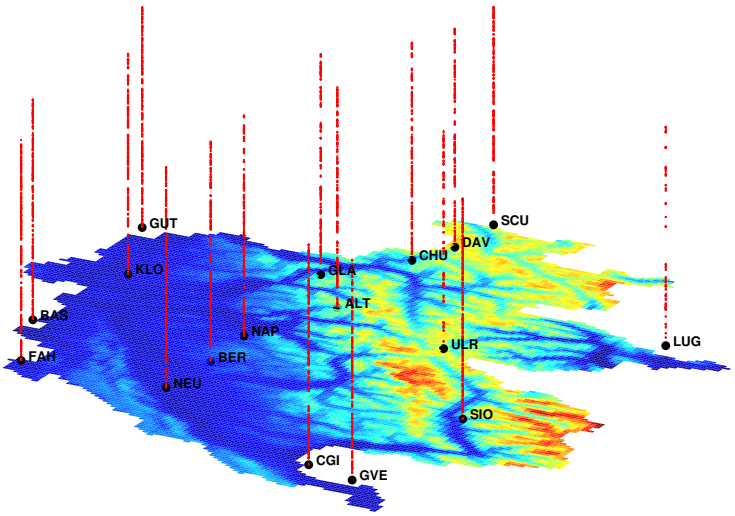
	Optimal FEM-BV-GPD model parameters						
	offset	Wind	Temp.	TSI	NAO	$ENSO_3$	$ENSO_{24}$
ξ_1	-0.0619	-0.0680	-0.4974	-0.0399	0.0161	-0.0168	-0.0094
$\pm$	(0.11)	(0.14)	(0.15)	(0.04)	(0.07)	(0.06)	(0.05)
ξ_2	0.2876	0.2763	-0.1053	-0.0038	0.0003	-0.0010	0.0018
$\pm$	(0.08)	(0.11)	(0.09)	(0.05)	(0.07)	(0.06)	(0.06)
σ_1	28.9111	12.9651	-0.0967	-1.9201	0.1589	0.3906	-0.0106
$\pm$	(2.82)	(3.71)	(2.49)	(1.11)	(1.67)	(1.49)	(1.37)
σ_2	3.0935	-2.3666	2.7117	-0.0444	0.0297	-0.0989	-0.0369
$\pm$	(0.29)	(0.42)	(0.45)	(0.25)	(0.34)	(0.27)	(0.26)

Table 5.11. *The table contains optimal FEM-BV-GPD model parameters and their standard errors (corresponding to the rows indicated by $\pm$) for threshold regression analysis of extreme accumulated rainfall for 17 locations in Switzerland with respect to $\hat{U}(s,t)$ as defined in (5.20).*

the optimal switching process, as the affiliation to one of the models, for each location is shown in Figure 5.8a for the first model and in Figure 5.8b for the second model. Both figures indicate that the switching process at location Lugano, and thus also the dynamics of threshold excesses, behaves very differently to the remaining locations. Given the overall optimal parameters (ξ_1, σ_1) and (ξ_2, σ_2) we can evaluate for each location s_i, for $i = 1,\ldots,17$, the temporal behavior by incorporating the corresponding switching process $\Gamma(s_i,t)$. For instance, for each location we can evaluate the return levels. In order to illustrate the main differences in dynamics of extremes north and south of the Alps, the time dependent values of the FEM-BV-GPD model parameters for locations Lugano and Basel are shown in Figure 5.9. Further, we estimate the correlation between the evaluated model parameters and the covariates deploying the Pearson's linear correlation, compare Table 5.13 and Table 5.12. According to this analysis, the dynamics of threshold excesses for daily accumulated precipitation measured at location Lugano spends more time in the first cluster. Here, the scale parameter exhibits large values, the dominating negative values for the shape parameter imply a bounded upper tail of the corresponding GPD distribution and lower probabilities of ex-



(a) *The temporal affiliation to the first local model*



(b) *The temporal affiliation to the second local model*

Figure 5.8. *The figures display the switching process for each single location.*

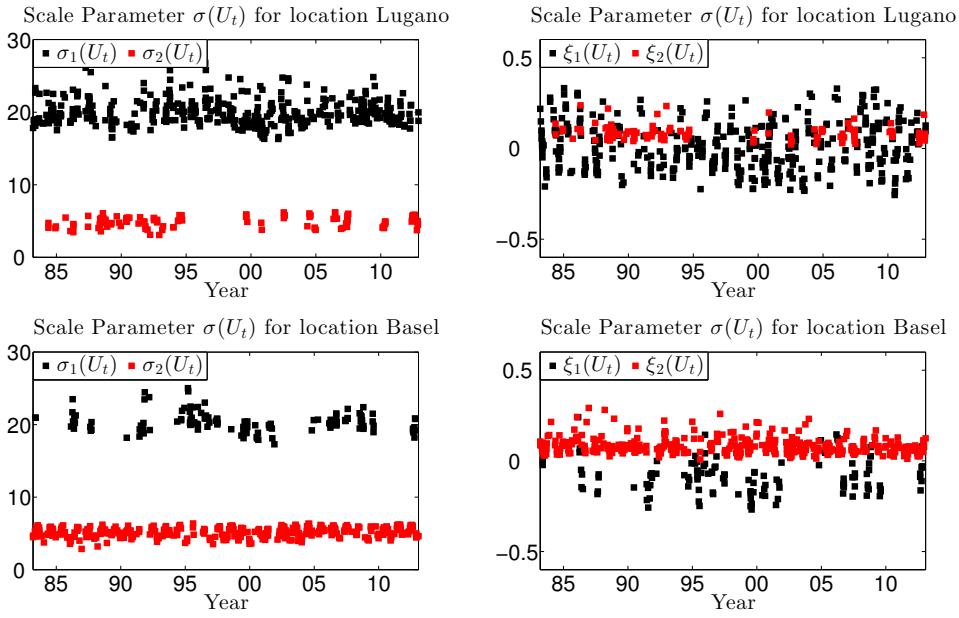


Figure 5.9. *This figure shows the evaluation of the optimal FEM-BV-GPD model parameters for two different locations: Lugano and Basel. The top right and left panels represent the scale and the shape parameters for location Lugano, respectively. The black markers correspond to the first and the red to the second model. Analogues, the model parameters for location basel are shown in the bottom right and left panels.*

tremes. While the scale parameter is positive correlated with the wind, the shape parameter has a negative correlation with wind and temperature. The dynamics in the second cluster is marked by small scale parameters and a mainly positive shape parameter. The scale parameter shows a positive correlation with the temperature. The shape parameter is positively correlated with the wind and negatively with the temperature.
The dynamics of threshold excesses for location Basel exhibits a different behavior. The dominating behavior is represented by the second model, with smaller values for the scale parameter and mainly positive values for the shape parameter.
Further, in the first model, which is responsible for the more extreme events, of interest could be the correlation between the scale parameter and the TSI index. This correlation points to an 11-year periodical behavior of threshold excesses for both locations; in years of low TSI indices the probability for extreme precipitation increases.
To study the underlying spatial dependence we refer to the nonlinear correlation among locations

	Pearson's linear correlation coefficient for Location Lugano					
	Wind	Temp.	TSI	NAO	$ENSO_3$	$ENSO_{24}$
ξ_1^*	**-0.3750**	**-0.9885**	-0.1036	0.1160	-0.1404	-0.0053
ξ_2^*	**0.7584**	**-0.5059**	0.0541	0.0730	0.0925	-0.0192
σ_1^*	**0.8999**	0.2833	**-0.3229**	-0.0560	0.1618	0.0572
σ_2^*	-0.2524	**0.9069**	-0.0657	-0.1070	-0.2592	-0.0448

Table 5.12. *The table contains the Pearson's linear correlation coefficients between the optimal FEM-BV-GPD model parameters and the covariates for location Lugano.*

	Pearson's linear correlation coefficient for Location Basel					
	Wind	Temp.	TSI	NAO	$ENSO_3$	$ENSO_{24}$
ξ_1^*	0.0104	**-0.9805**	-0.2967	0.0775	0.0657	-0.0103
ξ_2^*	**0.9084**	**-0.5607**	0.0503	0.1071	-0.0495	0.0065
σ_1^*	**0.8485**	-0.2064	**-0.4049**	-0.0320	0.2273	0.1738
σ_2^*	**-0.6339**	**0.8651**	-0.1545	-0.0568	-0.0147	0.0601

Table 5.13. *The table contains the Pearson's linear correlation coefficients between the optimal FEM-BV-GPD model parameters and the covariates for location Basel.*

measured by the event synchronization measure (ES)[76], compare Chapter 4.3.2. First, we estimate the stationary ES measure referring to the raw occurrence of extremes. The corresponding matrix is presented in Figure 5.10. Second, we estimate the nonstationary ES measure, i.e., for each local model we estimate a ES measure with respect to the corresponding switching process. The matrices are presented in Figure 5.11.
The stationary ES matrix determines mainly only weak correlation among locations. In contrast, the

nonstationary ES identifies strong correlated regions for each local model pointing to three major climatological regions of Switzerland: Tessin, Rhone Valley, pre-Apls. These results correspond to the scientific report of MeteoSwiss, where the seasonal variability of extreme precipitation were analyzed[108]. In the first model that is responsible for more extreme precipitation, strong correlation is observed between locations in adjacent geographic areas, for instance, with values higher then 0.5 (a) Chur, Davos, Scoul as well as (b) Genéve/Cointrin and Nyon/Changins. The ES for the second model exhibits larger correlated regions. Here location Lugano, representing the region south of the Alps, has only a weak correlation with the rest of Switzerland. Concluding, the stationary ES matrix does not capture the underlying spatial correlation, while the nonstationary extension represents the underlying correlation pointing to three major correlated regions. The above analysis of extreme precipitation over Switzerland revealed different dependent regions on a coarse grain. For future work remains the analysis of spatial extremes on a finer grid for the study of local dependence structure of extremes.

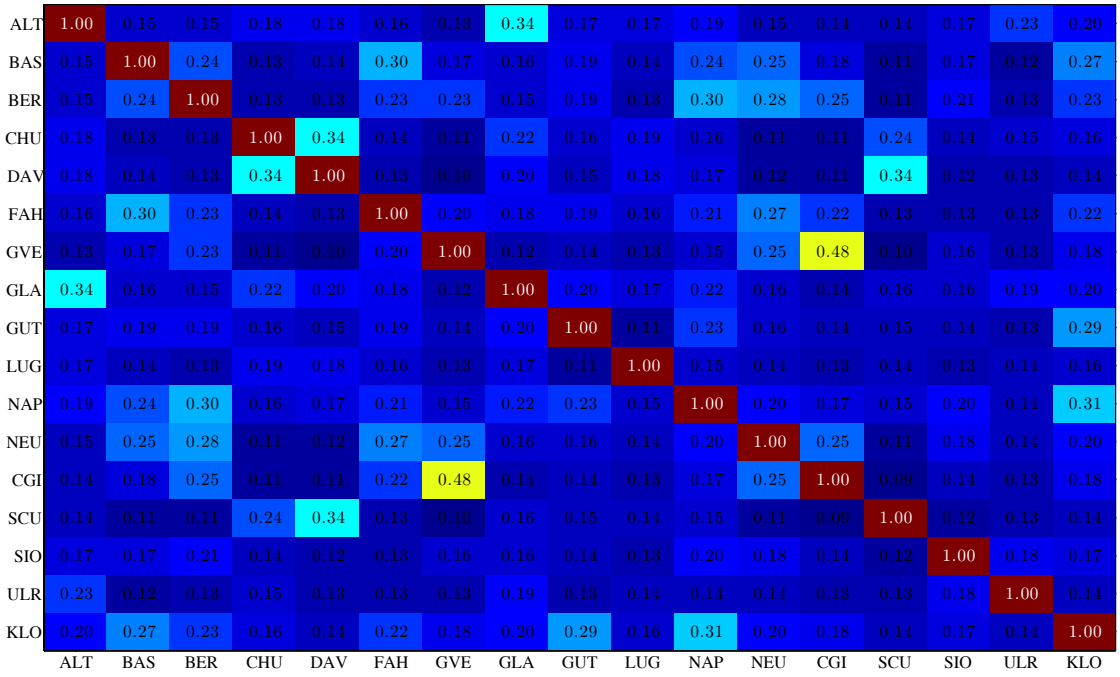
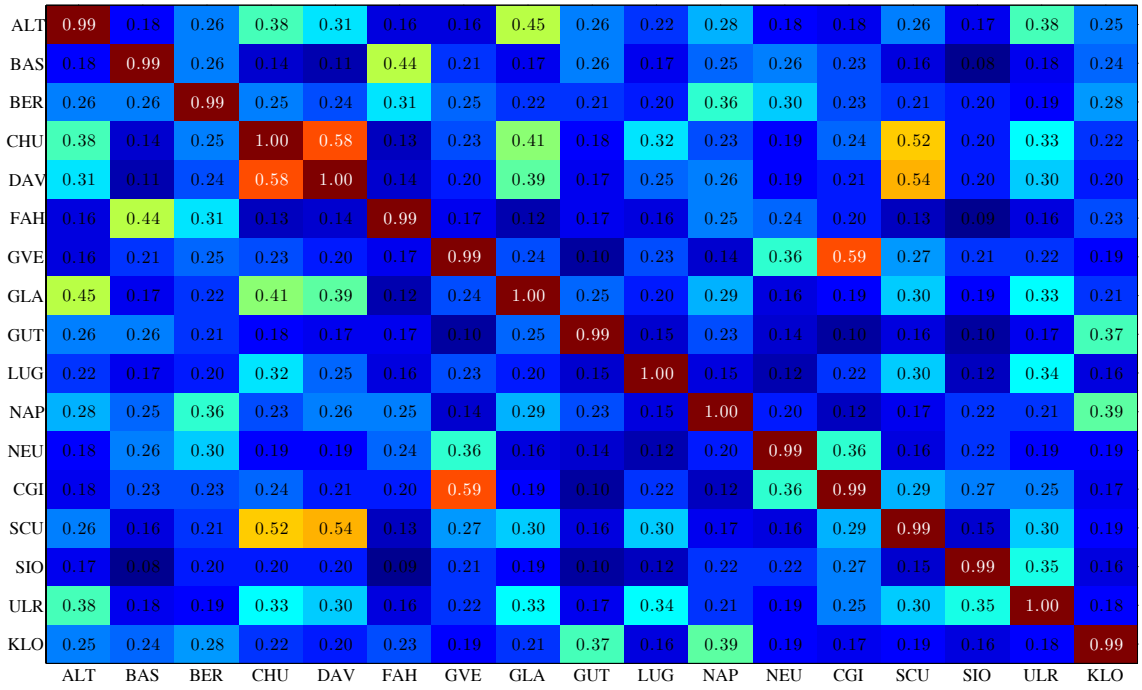
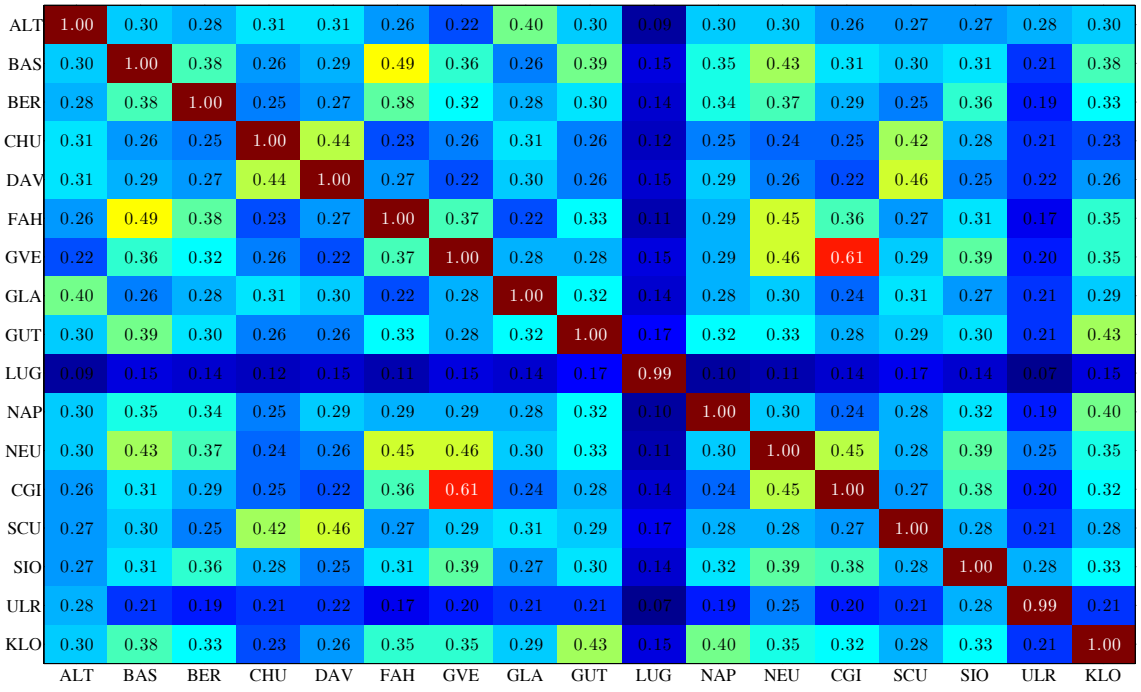


Figure 5.10. *The figures display the strength of event synchronization between the locations. Complete synchronization is ensured when the ES entry reaches the value* 1 *(the values are rounded to two decimal places).*



(a) *Event synchronization according to the first local model*



(b) *Event synchronization according to the second local model*

Figure 5.11. *The figures display the strength of event synchronization between the locations obtained from the results of FEM-BV-GPD. Complete synchronization is ensured when the ES entry reaches the value* 1 *(the values are rounded to two decimal places).*

6 Conclusion

Extreme events describe the above average behavior of a dynamical system. In order to reduce their negative social and financial impact, analysis and prediction of extreme events is significant in a wide range of areas such as civil engineering and risk management[2,42]. The main focus in data-based analysis of extreme events is to investigate their occurrence and intensity. The the latter point is studied in this thesis.

Focusing on statistical modeling of extremes, the appropriate statistical model should describe the extremes and be capable of predicting the "more extreme" extremes. The latter requirement is also known as the max-stability of a statistical model. Based on max-stability, the theoretical foundations for univariate extreme value analysis (EVA) were laid in the beginning of the 19th century, when the asymptotic distribution for sample maxima, called the Generalized Extreme Value distribution (GEV), was introduced[44,46,52]. Later on, EVA was extended towards analysis of threshold excesses: their asymptotic behavior is described by the Generalized Pareto Distribution (GPD)[35,52]. These limiting distributions were obtained under the assumption that the behavior of the underlying sample, from which the extremes are extracted, is stationary in time. The stationarity assumption can be released by incorporating the time dependence into the model parameters of the GEV and the GPD distributions[24]. The nonstationarity is often approached by expressing the model parameters as regression models[22]. The identification of the most significant covariates enables a better understanding of the causality relations in the underlying dynamics. Standard methodologies in statistical regression analysis of extremes are divided into parametric and nonparametric approaches. Parametric approaches are appropriate in the cases when either all the significant covariates are known, or when one can assume that the unresolved covariates are identically independently distributed. Those assumptions might be too strong for many real applications, for example, in climate research and economics we have often to account for the multiscale nature of the underlying process. Following, we can not always ensure that the set of the information collected about the analyzed process is complete. Standard nonparametric regression approaches include techniques based on Bayesian statistics, Mixture Modeling, or smoothing regression[7,10,74]. The first two approaches are based on a priori assumptions such as a priori known parametric distribution family for the model parameters and local stationarity. Smoothing regression can be assigned to the class of kernel smoothers, thereby inheriting the locality property.

In addition to the analysis of extremes that are extracted from one single process, a further subject of interest in EVA is the investigation of the relationship between extremes from different processes.

However, in contrast to the univariate EVA, there is no closed formulation for the asymptotic behavior of multivariate extremes. In particular, there exists no closed description of the underlying dependence structure. Standard approaches approximate the dynamics of multivariate extremes under some a priori assumptions about their dependence structure. For instances, in cases when the extremes are extracted from different spatial locations, referring to spatial extreme value analysis, results from geostatistics can be deployed. Under the assumption that the underlying dependence structure can be described by a Gaussian spatial process, Smith and Schlather max-stable processes were introduced[25,31,62]. Reich and Shaby introduced recently a nonparametric max-stable hierarchical approach, which approximates the dependence structure by a combination of predefined kernel functions[89].
In order to contribute to the current state of research in extreme value analysis, the goal of this dissertation thesis was to address primarily the following questions. First of all, the aim was to investigate what happens if some of significant covariates are systematically missing and to examine new ways of capturing the resulting nonstationarity. The second goal was to provide a nonparametric description of the underlying dependence structure, which accounts for nonstationarity and non-homogeneity. Third main task was the implementation of a framework which can appropriately address the issues of numerical instability and computational efficiency arising in the context of spatial extreme value analysis. The obtained results and conclusions are discussed in the next section.

6.1 Summary of Results and Conclusions

In the first part of this thesis, we presented an extension of the GEV and GPD models able of handling the situations with systematically missing covariates. We focused on linear regression and reduced the influence which is coming from missing covariates to an additive nonstationary offset by exploiting the Lindeberg and the Karhunen-Loève Theorems. In order to resolve the resulting nonstationarity beyond a priori assumptions, we deployed the Finite Element Time Series Analysis Methodology (FEM)[58,78]. Deployment of FEM in EVA-context allowed to approximate the underlying nonstationarity by $K \geq 1$ local EVA models and a nonstationary/nonparametric persistent hidden switching process. Following the general methodology of FEM, the persistency of the hidden switching process is ensured by assuming that it belongs to the space of functions with bounded variation (BV), where the number of switches is measured by a positive number C. In order to identify the most significant covariates for each local regression model, Lasso and Ridge shrinkage techniques were deployed. The resulting methodology is denoted as the univariate FEM-BV-EVA. The optimal FEM-BV-EVA model parameters K, C and the shrinkage parameter are chosen with respect to information criteria such as the Akaike Information Criteria. We discussed that the resulting univariate FEM-BV-EVA interpolates the underlying nonstationarity of the model parameters and goes beyond a priori assumptions typical for standard EVA approaches deploying, for instance, parametric regression, Hidden Markov Models or Local Kernel Smoothing. Additionally, the involved locally linear and stationary regression provides an easily interpretable and understandable statistical model in contrast to methods based on Neural Networks and smoothing regression.

The spatial extension of the FEM-BV-EVA formulation deployed the idea of hierarchical modeling. In order to describe the spatio-temporal variability in the dynamics of extremes, the EVA model parameters are expressed as spatial and nonstationary regression models based on resolved covariates only. The nonstationarity of the spatial regression was resolved by applying the spatial FEM formulation, where the underlying dynamics is described by a set of locally stationary models and a spatial nonstationary switching process. It was shown that the resulting spatial FEM-BV-EVA formulation is consistent with the max-stability postulate. FEM-BV-EVA describes the underlying spatial dependence structure through a data-driven nonstationary spatial clustering of extreme events.

The proposed FEM-BV-EVA methodology was integrated into the existent FEM framework - a MATLAB library containing time series analysis tools developed in the research group "Computational Time Series Analysis" of I. Horenko at ICS/USI Lugano. We extended the spatial FEM framework towards spatial regularization by additionally assuming persistent behavior in space. The implementation of FEM-BV-EVA deployed a gradient-free MCMC-based optimization technique and numerical solvers for large structured quadratic and linear problems with constraints. The performance of the FEM-BV-EVA framework was tested on various test cases and on real applications. The univariate FEM-BV-EVA approach was compared to standard approaches based on Neural Networks and smoothing regression with respect to the four criteria: (1) information content of the models; (2) ability to handle unresolved covariates; (3) computational complexity; (4) interpretability of the models. The results showed that in presence of missing covariates parametric approaches lead to biased description of the underlying dynamics. However, it was also discussed that regression analysis based on Neuronal Networks is more appropriate when the underlying dynamics is rather nonlinear then nonstationary. Further, we compared FEM-BV-GPD to methods based on smoothing regression. The latter was used for (a) resolving the nonstationarity coming from the unresolved covariates and (b) expressing the influence coming from the resolved covariates by smoothing splines. In the first case smoothing regression failed to capture the nonstationary influence coming from resolved covariates. The second case resulted in over-fitting. Additionally, based on linear regression FEM-BV-EVA provides an easily interpretable and understandable statistical model as was demonstrated on numerical examples.

Demonstration of spatial FEM-BV-EVA was performed on real data. For this, we considered daily accumulated precipitation over 17 different locations in Switzerland and defined the extremes as threshold excesses. The optimal FEM-BV-GPD description of the dynamics of extremes was obtained for two different models and a nonstationary switching process for each location. The switching process for location Lugano exhibits a completely different behavior as all the other locations. The underlying dependence structure was investigated by the nonstationary event synchronization measure revealing three major climatological regions of Switzerland: Tessin, Rhone Valley and pre-Apls. Standard stationary approach to synchronization measure inference has failed to recover this regional distinction in the analyzed data.

6.2 Future Work

Although the proposed framework is self-contained and can be directly applied for regression analysis of extremes, it could be further extended in a number of ways. In this section we address the most important ones from our point of view.
On a small scale, some of this issues were addressed directly in the thesis. For instance, in the present FEM-BV-GPD approach the threshold was chosen a priori and fixed for the entire sample, instead one could define and use a nonstationary threshold[22]. Also the consideration of a complete stochastic regression model with explicit noise terms remains for future study. In addition, FEM-BV-EVA could be extended towards nonlinear regression analysis by incorporating, for instance, Neural Networks based regression[14]. On a larger scale, analysis of the underlying dependence structure, that is of the switching process, should be approached in more details in a future research. Of particular interest is the spatial interpolation of the dynamics of extremes between observed locations and the spatio-temporal propagation of extreme events. The first issue could be, for instance, approached by referring to the switching process as a discrete process. Then its underlying dynamics could be analyzed by deploying methods such as the FEM-BV-Markov approach[60,78]. The second issue could be addressed, for example, by applying results from the complex network theory, where among others the delay behavior of a process is studied[76]. An alternative approach is to apply the FEM-BV-Causality approach for the analysis of the spatio-temporal causality interaction of discrete state models[50]. FEM-BV-EVA provides a robust regression analysis tool of spatio-temporal extremes and can be used to identify the preceding external factors. Also concepts and methods from the theory of point processes could be taken into account for a more accurate prediction of the occurrence of extremes. From computational point of view, the proposed framework could be extended towards highly-scalable applications in HPC context; the several possible levels of parallelization were discussed in detail in the thesis. The parallelization will improve the computational efficiency of the FEM-BV-EVA framework and enable regression analysis of extremes over larger regions with a finer spatial resolution.

A Appendix

A.1 Key Statistical Concepts

Let Y be a random variable, i.e., the output is uncertain. Further, let denote the set of all possible realizations of Y by the sample space Ω. In case Ω is continuos, for instance, $\Omega = \mathbb{R}$, Y describes a continues random variable. Each realization $y \in \Omega$ is assigned to Y by a probability distribution, which is given by the probability distribution function[22,79]

$$F(y) = \mathbb{P}[Y \leq y], \quad y \in \Omega. \tag{A.1}$$

Please note, if Ω is a discrete sample space, for example, $\Omega = \{0,1\}$, the probability distribution is given by the so-called probability mass function. Because in this work we focus on continuos random variables, we will omit the descriptions of respective concepts for the discrete case. For details on discrete random variables we refer the interested reader to[79]. The probability distribution in (A.1) is also denoted as the cumulative distribution function (cdf)[79]. In cases when the cdf is differentiable, the corresponding probability density function (pdf) is defined by

$$f(y) = \frac{dF(y)}{dy}, \tag{A.2}$$

such that

$$F(y) = \int_{-\infty}^{y} f(w)dw \quad \text{and} \quad \mathbb{P}[a \leq Y \leq b] = \int_{a}^{b} f(w)dw. \tag{A.3}$$

A pdf can be completely characterized/described by its main features, the so-called statistics[22,79]. The most commonly used are the expectation

$$\mathbb{E}[Y] = \int_{\Omega} yf(y)dy, \tag{A.4}$$

and the variance

$$\mathbb{V}ar(Y) = \int_{\Omega} (y - \mathbb{E}[Y])^2 f(y)dy. \tag{A.5}$$

The expectation, the value of Y on average, is also denoted as the location of a distribution. The variance, describing the spread from the expectation, is referred to as the scale parameter of a distribution. Some further statistics are the shape and the skewness parameters of a distribution[79]. In cases when the descriptive statistics contains a finite number of parameters, summarized by θ, the corresponding probability density function $f(y) = f(y;\theta)$ is parametric[29]. Then, for fixed $f(\cdot)$ and varying parameters θ the parametric family of probability distributions is defined by

$$\mathcal{F} = \{f(y;\theta)|\theta \in \Omega_\theta\}, \tag{A.6}$$

where Ω_θ is the space of all possible values for θ.

Example A.1.1 (Gaussian Distribution) *The probability density function of a Gaussian distribution with mean $\mu \in \mathbb{R}$ and variance $\sigma > 0$ is defined by*

$$f(y;\theta) = \frac{1}{\sigma\sqrt{2\Pi}} \exp\{-\frac{(y-\mu)^2}{2\sigma^2}\}. \tag{A.7}$$

The Gaussian distribution is described by $\theta = (\mu, \sigma)$, thus the family of Gaussian distributions is given by $\mathcal{F} = \{f(y;\theta)|\theta \in \mathbb{R} \times \mathbb{R}_{>0}\}$.

When the distribution $f(y)$ of a sample of random variables $Y_1, \ldots, Y_n$ is known, it can be used for estimating the descriptive statistics for summarizing the behavior of the sample, for instance, the value on average can be estimated by evaluating (A.4). Yet, in many real applications the underlying distribution of a sample $Y_1, \ldots, Y_n$ is unknown and we have to solve an inverse problem. Thereby, the aim is to find the most descriptive statistical model, i.e., an appropriate set $\{\mathcal{F}, \theta\}$. Assuming that the parametric family of probability distributions $\mathcal{F}$ is known, the inverse problem is reduced to the estimation of the optimal θ for the corresponding sample. A widely used approach for estimating the optimal θ is based on the idea that the original parameter denoted by θ_0 maximizes the likelihood of the joint occurrence of $Y_1, \ldots, Y_n$. Following, we are searching for a parameter θ which maximizes the likelihood function defined by

$$L(\theta; Y_1, \ldots, Y_N) = \mathbb{P}[Y_1, \ldots, Y_N|\theta]. \tag{A.8}$$

Assuming independence and identical distribution (i.i.d) of the sample we get

$$L(\theta; Y_1, \ldots, Y_N) = \mathbb{P}[Y_1, \ldots, Y_N|\theta] \overset{i.i.d}{=} \prod_{i=1}^{n} f(Y_i;\theta). \tag{A.9}$$

Because it is more convenient to deal with the logarithm of (A.9), we refer to the log-likelihood function

$$l(\theta; Y_1, \ldots, Y_N) = \log L(\theta; Y_1, \ldots, Y_N). \tag{A.10}$$

We get the optimal parameter by maximizing the log-likelihood function (A.10) with respect to θ

$$\theta^* = \underset{\theta}{\operatorname{argmax}}\, l(\theta; Y_1, \ldots, Y_N) \tag{A.11}$$

$$= \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{n} \log f(Y_i;\theta). \tag{A.12}$$

The function, which depends on the random sample and is used for estimating θ^*, is usually called the estimator, for instance, (A.11) corresponds to the maximum log-likelihood estimator[22,79]. Since the sample $Y_1,\ldots,Y_n$ is the result of a random process, we will get a different estimates for θ^* for different realizations of the sample. In order to measure the uncertainty of the estimate, the experiment is repeated N times (by drawing a new sample) and for each sample an estimate for θ_i^*, $i=1,\ldots,N$, is obtained. In the next step, the sample of $\theta_1^*,\ldots,\theta_N^*$ can be used for estimating the distribution of the parameters, in particular, the empirical expectation and variance. These quantities are used to measure the quality of an estimator. The bias of an estimator is defined as the deviation between the expected optimal and the original parameter

$$Bias\,(\theta^*) = \mathbb{E}[\theta^*] - \theta_0. \tag{A.13}$$

An estimator is said to be unbiased when on average the estimate θ^* is the true parameter θ_0, i.e., $Bias\,(\theta^*) = 0$. To measure the variability/uncertainty of an estimator, the mean-square error is frequently used

$$MSE(\theta^*) = \mathbb{E}[(\theta^* - \theta_0)^2] = \mathbb{V}ar[\theta^*] + Bias^2\,(\theta^*)\,. \tag{A.14}$$

In cases when the estimator is unbiased, the mean-square error is the variance of the estimator and the model parameter distribution is described by the set of parameters $\{\mathbb{E}[\theta^*], MSE(\theta^*)\}$. In cases, when the estimator is unbiased, the root of $MSE(\theta^*)$ corresponds to the standard error of an estimator. Following, hence we are interested in estimators that are more precise and less biased, we choose the ones with the smallest standard error. However in real application we often do not know the underlying distribution of the model parameters. Consequently, we can not estimate directly the corresponding variability. Moreover, in real applications, for instance, in weather and climate research, just one realization of the underlying process is available for statistical analysis. In such cases the variability of an estimator can be described by the empirical confidence intervals, which are estimated exploiting resampling/bootstrapping techniques[28].
The above approach to describe the uncertainty of the estimate in terms of confidence intervals obtained by applying repeating or resampling techniques is also called the "frequentist approach" in the literature[29].

A.2 Bayesian Inference

An alternative to distribution estimation of the model parameters, exploiting e.g., bootstrapping techniques, is provided by the "Bayesian statistics"[26,29,79]. In the Bayesian approach one assumes a priori that the model parameters are defined by some a priory fixed distribution families. Then, this information can be incorporated into the statistical inference procedure by means of the Bayes' Theorem[29,79]. The resulting conditional posterior distribution is then used for inference. In this section we briefly discuss the Bayesian approach to statistics.
Following the notation in Definition A.2.1, the a priori assumption about the distribution of the model parameters is summarized in the prior and the resulting conditional distribution of parameters

is the posterior[79]. Further, in contrast to Section A.1, where θ is an unknown constant and the probability distribution is given by $f(y;\theta)$, here θ is a realization of a random variable and we write $f(y|\theta)$.

Definition A.2.1 (Prior and posterior distributions) *The distribution $\pi(\theta)$ is called the prior distribution. The resulting density of the model parameter θ, conditioned on $\pi(\theta)$ and given that $Y = y$, is denoted by $f(\theta|y)$ and is called the posterior distribution.*

Theorem A.2.1 (Bayesian Theorem) *Given two random variables A and B with $\mathbb{P}[B] > 0$, the probability of A conditioned on B, i.e., $\mathbb{P}[A|B]$, is given by*

$$\mathbb{P}[A|B] = \frac{\mathbb{P}[B|A]\mathbb{P}[A]}{\mathbb{P}[B]}. \tag{A.15}$$

Applying the Bayesian Theorem A.2.1 we obtain the posterior distribution of parameters by

$$f(\theta|y) = \frac{f(y|\theta)\,\pi(\theta)}{f(y)}. \tag{A.16}$$

The posterior distribution is proportional to the likelihood and the prior. The proportionality constant is given by

$$f(y) = \int_{\Theta} f(y|\theta)\,\pi(\theta)\,d\theta, \tag{A.17}$$

and is also called the marginal distribution, which is the distribution of the observed data averaged over the prior distribution of all values of θ. The random behavior arising from θ is "integrated out"[13]. In some special cases the integral in (A.17) can be evaluated and we obtain a closed expression for the posterior. In particular, this is the case when the prior and the posterior are conjugate, i.e., coming from the same family of distributions[4,29]. The posterior distribution $f(\theta|y)$ accounts for the a priori knowledge and can be used for postprocessing, for example, the estimation of the expectation and the variance for θ. However, in general there is no closed expression of (A.17). To obtain this distribution practically in numerical simulation, for instance, based on Monte Carlo methods, can be deployed[22,29,79]. One of the main issues in Bayesian inference, is the choice of an appropriate prior; a wrong a priori assumptions about the distribution of the model parameters might lead to biased conclusions about the underlying dynamics. In case no a priori knowledge is available the easiest way is to refer to a prior where the values for θ are identically distributed, such a prior is called uninformative[4]. For more detailed discussions please refer to[4,22,26,29,79].

A.3 FEM Spatial Regularization

In the spatial FEM setting, first introduced for spatio-temporal Markov regression analysis of discrete/categorical dynamical processes[38], the following objective functional

$$\mathcal{L}(\Theta,\Gamma(s,t)) = \sum_{i=1}^{N_S}\sum_{j=1}^{N_T}\sum_{k=1}^{K}\gamma_k(s_i,t_j)g_{Model}\left(Y_{s_i,t_j},\theta_k\right), \tag{A.18}$$

is minimized with respect to the model parameters Θ and $\Gamma(s,t)$. Thereby, $g_{Model}(.)$ describes the "distance measure" between the data $Y_{s,t}$ and the appropriate $Model$, which is characterized by Θ. For example, in context of this thesis $Model$ corresponds either to GEV or to GPD distributions and $g_{Model}(.)$ to the negative log-likelihood. Further, the switching process $\Gamma(s,t)$ has to fulfill the convexity constraints

$$\sum_{k=1}^{K} \gamma_k(s,t) = 1 \quad \text{and} \quad \gamma_k(s,t) \geq 0, \tag{A.19}$$

for $t = t_1, \ldots, t_{N_T}$, $s = s_1, \ldots, s_{N_S}$ and $k = 1, \ldots, K$ and the temporal persistency constraint

$$||\gamma_k(s,t)||_{\mathcal{R}([1,N_T])} \leq C(N_T), \quad s = s_1, \ldots, s_{N_S}, \quad k = 1, \ldots, K, \tag{A.20}$$

with $\mathcal{R}([a,b]) = BV([a,b])$ or $\mathcal{R}([a,b]) = H^1([a,b])$. Additionally we can impose spatial persistent behavior involving

$$||\gamma_k(s,t)||_{\mathcal{R}([1,N_S])} \leq C(N_S), \quad t = t_1, \ldots, t_{N_T}, \quad k = 1, \ldots, K, \tag{A.21}$$

please compare Section 4.4 for a detailed discussion. In the following sections we derive the corresponding linear and quadratic problems, referring to a spatio-temporal BV and spatio-temporal H1 regularization, respectively.

A.3.1 H1 regularization in time and space

In the following, we extend FEM[38,58,78] towards $H1$ regularization in space, i.e., we consider the following spatio-temporal persistency constraints for the switching process:

$$||\gamma_k(s,t)||_{H_1(t_1,t_{N_T})} \leq C(N_T), \quad s = s_1, \ldots, s_{N_S}, \quad k = 1, \ldots, K, \tag{A.22}$$

$$||\gamma_k(s,t)||_{H_1(s_1,s_{N_S})} \leq C(N_S), \quad t = t_1, \ldots, t_{N_T}, \quad k = 1, \ldots, K. \tag{A.23}$$

In line with the FEM methodology[58], where the temporal $H1$ regularization is incorporated by a Lagrange multiplier ε_1^2, using an additional Lagrange multiplier ε_2^2 we insert the constraint (A.23) and obtain the spatio-temporal $H1$ regularized average distance functional:

$$\mathcal{L}\left(\Theta, \Gamma(s,t), \varepsilon_1^2, \varepsilon_2^2\right) = \sum_{i=1}^{N_J} \sum_{j=1}^{N_T} \sum_{k=1}^{K} \gamma_k(s_i,t_j) g_{Model}\left(Y_{s_i,t_j}, \theta_k\right) + \varepsilon_1^2 \sum_{k=1}^{K} ||\gamma_k(s,t)||_{H^1(t_1,t_{N_T})} \tag{A.24}$$

$$+\varepsilon_2^2 \sum_{k=1}^{K} ||\gamma_k(s,t)||_{H^1(s_1,s_{N_S})}. \tag{A.25}$$

The above functional can be reformulated in a quadratic one. For this purpose, the switching process is discretized by the Finite Elements as was proposed for the original FEM-H1 formulation[58]. Let

us first consider $\varepsilon_2^2 = 0$, then we can directly adapt the formulation in[58] and after some algebraic reformulation obtain

$$\mathcal{L}\left(\Theta, \Gamma(s,t), \varepsilon_1^2\right) = \sum_{k=1}^{K} \left(a(\theta_k)^\dagger \overline{\gamma}_k + \underbrace{\varepsilon_1^2 \overline{\gamma}_k^\dagger \mathcal{H}_{1k} \overline{\gamma}_k}_{\text{time regularization}} \right), \tag{A.26}$$

with the vector

$$\begin{aligned} a(\theta_k) = (\, & g_{Model}\left(Y_{s_1,t_1}, \theta_k\right), \ldots, g_{Model}\left(Y_{s_1,t_{N_T}}, \theta_k\right), && \text{(A.27)} \\ & g_{Model}\left(Y_{s_2,t_1}, \theta_k\right), \ldots, g_{Model}\left(Y_{s_2,t_{N_T}}, \theta_k\right), && \text{(A.28)} \\ & \vdots && \text{(A.29)} \\ & g_{Model}\left(Y_{s_{N_S},t_1}, \theta_k\right), \ldots, g_{Model}\left(Y_{s_{N_S},t_{N_T}}, \theta_k\right), && \text{(A.30)} \\ & && \text{(A.31)} \end{aligned}$$

and $\overline{\gamma}_k^\dagger$ is the spatio-temporal discretized switching process with

$$\begin{aligned} \overline{\gamma}_k^\dagger = (\, & \tilde{\gamma}_k(s_1,t_1), \ldots, \tilde{\gamma}_k(s_1,t_{N_T}), && \text{(A.32)} \\ & \tilde{\gamma}_k(s_2,t_1), \ldots, \tilde{\gamma}_k(s_2,t_{N_T}), && \text{(A.33)} \\ & \vdots && \text{(A.34)} \\ & \tilde{\gamma}_k(s_{N_S},t_1), \ldots, \tilde{\gamma}_k(s_{N_S},t_{N_T})\,). && \text{(A.35)} \end{aligned}$$

The discretization matrix $\mathcal{H}_{1k}$ is given by

$$H_{1k} = \begin{pmatrix} H_{s_1} & & \\ & \ddots & \\ & & H_{s_{N_S}} \end{pmatrix} \in \mathbb{R}^{N_S N_T \times N_S N_T}, \tag{A.36}$$

where each block matrix H_s, $s = s_1, \ldots, s_{N_S}$, corresponds to the temporal discretization for each location s

$$H_s = \begin{pmatrix} 1 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 1 \end{pmatrix} \in \mathbb{R}^{N_T \times N_T}. \tag{A.37}$$

By applying a similar idea for spatial regularization, we discretize also the spatial dimension and obtain

$$\mathcal{L}\left(\Theta, \Gamma(s,t), \varepsilon_1^2, \varepsilon_2^2\right) = \sum_{k=1}^{K} \left(a(\theta_k)^\dagger \overline{\gamma}_k + \underbrace{\varepsilon_1^2 \overline{\gamma}_k^\dagger \mathcal{H}_{1k} \overline{\gamma}_k}_{\text{time regularization}} + \underbrace{\varepsilon_2^2 \overline{\gamma}_k^\dagger \mathcal{H}_{2k} \overline{\gamma}_k}_{\text{space regularization}} \right), \tag{A.38}$$

where $\overline{\gamma}_k$ and $\mathcal{H}_{1k}$ are defined in (A.32) and (A.36), respectively. The matrix $H_{2k} \in \mathbb{R}^{N_S N_T \times N_S N_T}$ describes the spatial pairwise connectivity among all locations, referring thereby, for instance, to the classical correlation or the cross-correlation matrix of the observed series among all locations. In cases if we refer to the classical correlation, then each pair of s_i, s_v is correlated equally for each time step and the computation of H_{2k} is given by the Algorithm 5. Matrix H_{1k} is symmetric and we

Algorithm 5: An example for computation of H_{2k}

Initialize: compute correlation matrix $corr(Y_{s,t})$ for a given series $Y_{s,t}$;

for $t = 1 : N_T$ **do**

$H_{2k}(t : N_T : end, t : N_T : end) = corr(Y_{s,t})$;

can also ensure symmetry for H_{2k} by considering only symmetric connectivity. By summing up the matrices $\mathcal{H}_k = \varepsilon_1^2 \mathcal{H}_{1k} + \varepsilon_2^2 \mathcal{H}_{2k}$ and using $a(\Theta) = (a(\theta_1), \ldots, a(\theta_K))$ and $\mathcal{H} = diag(\mathcal{H}_1, \ldots, \mathcal{H}_K)$, i.e., H is a block-diagonal matrix, we can rewrite (A.24) into

$$\mathcal{L}\left(\Theta, \Gamma(t), \varepsilon_1^2, \varepsilon_2^2\right) = \overline{\Gamma}^{\dagger} \mathcal{H} \overline{\Gamma} + a^{\dagger}(\Theta) \overline{\Gamma} \tag{A.39}$$

$$\text{with } Aeq\overline{\Gamma} = beq \text{ and } \overline{\Gamma} \geq 0. \tag{A.40}$$

The equality constraint in (A.40) summarizes the equality constraint in (A.19) for all locations at each time step. Standard methods for quadratic programming [19,45] can be applied for minimizing (A.39-A.40).

A.3.2 BV regularization in time and space

The class of function with bounded variation allows to preserve sharp transitions of the switching process, while smooth transitions are not excluded since: $H^1([0, N_T]) \subset BV([0, N_T])$ [78]. In this section FEM [38,58,78] is extended towards BV regularization in space, i.e., we refer to the following constraints:

$$||\gamma_k(s,t)||_{BV(t_1, t_{N_T})} \leq C(N_T), \quad s = s_1, \ldots, s_{N_S}, \quad k = 1, \ldots, K, \tag{A.41}$$

$$||\gamma_k(s,t)||_{BV(s_1, s_{N_S})} \leq C(N_S), \quad t = t_1, \ldots, t_{N_T}, \quad k = 1, \ldots, K. \tag{A.42}$$

In the next two steps we adapt the strategy of reformulating the temporal regularization constraint by using slack variables as was proposed in [78]: (A.41) can be rewritten for each model k and each location s towards

$$\sum_{t=j}^{N_T - 1} \eta_k(s, t_j) \leq C(N_T), \tag{A.43}$$

$$\gamma_k(s, t+1) - \gamma_k(s,t) - \eta_k(s,t) \leq 0, \tag{A.44}$$

$$-\gamma_k(s, t+1) + \gamma_k(s,t) - \eta_k(s,t) \leq 0, \tag{A.45}$$

$$\eta_k(s,t) \geq 0, \tag{A.46}$$

for $t = t_1, \ldots, t_{N_T-1}$. Further, the spatial *BV* regularization (A.42) imposes

$$\sum_{i=1}^{N_S-1} |\gamma_k(s+1,t) - \gamma_k(s,t)| \leq C(N_S), \tag{A.47}$$

for each model k and each time step t. Analogue, exploiting slack variables constraint (A.42) can be reformulated

$$\sum_{i=1}^{N_S-1} \zeta_k(s,t) \leq C(N_S), \tag{A.48}$$

$$\gamma_k(s+1,t) - \gamma_k(s,t) - \zeta_k(s,t) \leq 0, \tag{A.49}$$

$$-\gamma_k(s+1,t) + \gamma_k(s,t) - \zeta_k(s,t) \leq 0, \tag{A.50}$$

$$\zeta_k(s,t) \geq 0, \tag{A.51}$$

for $s = s_1, \ldots, s_{N_S-1}$. However, constraints (A.49-A.50) imply that only direct adjacent locations are connected. In order to account for interactions among all locations, consider a symmetric matrix $W \in \mathbb{R}^{N_S \times N_S}$ where the weights w_{iv} describe the connection between the location s_i and location s_v, for $i, v = 1, \ldots, N_S$. For example, the matrix W can be chosen as the classical correlation or the cross-correlation matrix of the observed series (further discussion how to choose W can be found in Section 4.4). Then, instead of (A.47) we consider the weighted *BV* regularization

$$\sum_{i=1}^{N_S} \sum_{v=1}^{N_S} \frac{1}{2} w_{iv} |\gamma_k(s_i,t) - \gamma_k(s_v,t)| \leq C(N_S). \tag{A.52}$$

The factor $\frac{1}{2}$ is due to the symmetry of matrix W. Constraint (A.52) has to be adopted in line with (A.48-A.49-A.50-A.51); taking the symmetry into consideration (thus the factor $\frac{1}{2}$ is vanishing), we obtain

$$\sum_{i=1}^{N_S-1} \sum_{v=i+1}^{N_S} \zeta_k(s_i,s_v,t) \leq C(N_S), \tag{A.53}$$

$$w_{iv}\left(\gamma_k(s_i,t) - \gamma_k(s_v,t)\right) - \zeta_k(s_i,s_v,t) \leq 0, \tag{A.54}$$

$$-w_{iv}\left(\gamma_k(s_i,t) + \gamma_k(s_v,t)\right) - \zeta_k(s_i,s_v,t) \leq 0, \tag{A.55}$$

$$\zeta_k(s_i,s_v,t) \geq 0. \tag{A.56}$$

Concluding, when referring to spatio-temporal persistent behavior the optimal $\Gamma(s,t)$ for fixed Θ is obtained by linear minimization of (A.18) with constraints (A.19), (A.43-A.44-A.45-A.46) and (A.53-A.54-A.55-A.56). In order to rewrite this minimization problem as a closed linear constrained problem, we define in line with[78] the vector of all the unknowns, i.e., $\Gamma(s,t)$ and the involved slack variables, by

$$\Omega = (\omega_1, \ldots, \omega_K, \overline{\omega}_1, \ldots, \overline{\omega}_K, \tilde{\omega}_1, \ldots, \tilde{\omega}_K), \tag{A.57}$$

with

$$\omega_k = (\gamma_k(s_1,t_1),\ldots,\gamma_k(s_1,t_{N_T}),\ldots,\gamma_k(s_{N_S},t_1),\ldots,\gamma_k(s_{N_S},t_{N_T})) \in \mathbb{R}^{N_S N_T \times 1}, \tag{A.58}$$

$$\overline{\omega}_k = (\eta_k(s_1,t_1),\ldots,\eta_k(s_1,t_{N_T-1}),\ldots,\eta_k(s_{N_S},t_1),\ldots,\eta_k(s_{N_S},t_{N_T-1})) \in \mathbb{R}^{N_S(N_T-1)\times 1}, \tag{A.59}$$

$$\tilde{\omega}_k = (\zeta_k(s_1,v_2,t_1),\ldots,\zeta_k(s_1,v_2,t_{N_T}),\ldots,\zeta_k(s_{N_S-1},v_{N_S},t_1),\ldots,\zeta_k(s_{N_S-1},v_{N_S},t_N)) \in \mathbb{R}^{0.5N_T(N_S^2-N_S)\times 1}. \tag{A.60}$$

Further, we define the vector of NLL values for all time steps and all locations by $\Phi = (\phi_1,\ldots,\phi_K)$ with

$$\phi_k = \left(g_{Model}\left(Y_{s_1,t_1},\theta_k\right),\ldots,g_{Model}\left(Y_{s_1,t_{N_T}},\theta_k\right),\ldots, \tag{A.61}$$

$$g_{Model}\left(Y_{s_{N_S},t_1},\theta_k\right),\ldots,g_{Model}\left(Y_{s_{N_S},t_{N_T}},\theta_k\right)\Big) \in \mathbb{R}^{1\times N_S N_T}, \tag{A.62}$$

for $k = 1,\ldots K$. Then, we can rewrite the functional (A.18) as

$$\mathcal{L}(\Theta,\Gamma(s,t)) = \langle \Phi(\Theta),\Omega\rangle, \tag{A.63}$$

and the constraints (A.19), (A.43-A.44-A.45-A.46) and (A.53-A.54-A.55-A.56) as equality and inequality constraints

$$Aeq\Omega = beq, \quad Aneq\Omega \leq bneq, \quad \Omega \geq 0. \tag{A.64}$$

To minimize (A.63-A.64) standard methods for linear programming can be applied, for instance, the simplex method[19,45] which is available, for example, in the GNU Scientific Library and also in MATLAB.

Bibliography

[1] Akaike, H. 1974. A new look at the statistical model identification, *IEEE Transactions on Automatic Control* **19**(6).

[2] Albeverio, S., Jentsch, V. and Kantz, H. 2006. *Extreme Events in Nature and Society*, 1th edn, Springer New York.

[3] Andrieu, C., De Freitas, N., Doucet, A. and Jordan, M. I. 2003. An introduction to mcmc for machine learning, *Machine Learning* **50**(1): 5–43.

[4] Aster, R. C., Borchers, B. and Thurber, C. H. 2013. *Parameter estimation and inverse problems*, Academic Press.

[5] Bacro, J.-N. and Gaetan, C. 2013. Estimation of spatial max-stable models using threshold exceedances, *Statistics and Computing* pp. 1–12.

[6] Beirlant, J., Goegebeur, Y., Segers, J. and Teugels, J. 2006. *Statistics of extremes: theory and applications*, Wiley. com.

[7] Betrò, B., Bodini, A. and Cossu, A. Q. 2008. Using a hidden markov model to analyse extreme rainfall events in central-east sardinia, *Environmetrics* **19**(7): 702–713.

[8] Billings, S. A. 2013. *Nonlinear system identification: NARMAX methods in the time, frequency, and spatio-temporal domains*, John Wiley & Sons.

[9] Bishop, C. M. et al. 2006. *Pattern recognition and machine learning*, Vol. 1, springer New York.

[10] Boldi, M.-O. and Davison, A. C. 2007. A mixture model for multivariate extremes, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **69**(2).

[11] Brooks, S., Gelman, A., Jones, G. and Meng, X.-L. 2011. *Handbook of Markov Chain Monte Carlo (Chapman & Hall/CRC Handbooks of Modern Statistical Methods)*, 1 edn, Chapman and Hall/CRC.

[12] Burnham, K. P. and Anderson, D. R. 2002. *Model selection and multimodel inference: a practical information-theoretic approach*, Springer.

[13] Cameron, A. C. and Trivedi, P. 2013. *Regression analysis of count data*, Vol. 53, Cambridge University Press.

[14] Cannon, A. J. 2010. A flexible nonlinear modelling framework for nonstationary generalized extreme value analysis in hydroclimatology, *Hydrological Processes* **24**(24): 673–685.

[15] Cannon, A. J. 2011. Gevcdn: an r package for nonstationary extreme value analysis by generalized extreme value conditional density estimation network, *Computers Geosciences* **37**: 1532–1533.

[16] Cannon, A. J. 2012. Neural networks for probabilistic environmental prediction: Conditional density estimation network creation and evaluation (cadence) in r, *Computers & Geosciences* **41**: 126–135.

[17] Cassou, C., Terray, L. and Phillips, A. S. 2005. Tropical atlantic influence on european heat waves, *Journal of climate* **18**(15): 2805–2811.

[18] Chavez-Demoulin, V. and Davison, A. C. 2005. Generalized additive modelling of sample extremes, *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **54**(1): 207–222.

[19] Chinneck, J. W. 2007. *Feasibility and Infeasibility in Optimization:: Algorithms and Computational Methods*, Vol. 118, Springer.

[20] Chorin, A. J. and Hald, O. H. 2006. *Stochastic Tools in Mathematics and Science*, Springer.

[21] Clarke, B. S., Fokouâe, E. and Zhang, H. H. 2009. *Principles and theory for data mining and machine learning*, Springer Series in Statistics.

[22] Coles, S. G. 2001. *An Introduction to Statistical Modelling of Extreme Values*, Springer, Springer series in statistics.

[23] Coles, S. G. and Dixon, M. J. 1999. Likelihood-based inference for extreme value models, *Springer, Extremes* **2**(1): 5–23.

[24] Coles, S. G. and Powell, E. A. 1996. Bayesian methods in extreme value modelling: A review and new developments, *International Statistical Review* **64**(1).

[25] Cooley, D., Cisewski, J., Erhardt, R. J., Jeon, S., Mannshardt, E., Omolo, B. O. and Sun, Y. 2012. A survey of spatial extremes: measuring spatial dependence and modeling spatial effects, *REVSTAT–Statistical Journal* **10**(1): 135–165.

[26] Cox, D. R. 2006. *Principles of statistical inference*, Cambridge University Press.

[27] Cressie, N. and Wikle, C. K. 2011. *Statistics for spatio-temporal data*, Wiley. com.

[28] Davison, A. C. 1997. *Bootstrap methods and their application*, Vol. 1, Cambridge university press.

[29] Davison, A. C. 2003. *Statistical models*, number 11, Cambridge University Press.

[30] Davison, A. C. and Gholamrezaee, M. M. 2012. Geostatistics of extremes, *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Science* **468**(2138): 581–608.

[31] Davison, A. C., Padoan, S. A. and Ribatet, M. 2012. Statistical modeling of spatial extremes, *Statistical Science* **27**(2): 161–186.

[32] Davison, A. C. and Ramesh, N. 2000. Local likelihood smoothing of sample extremes, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **62**(1): 191–208.

[33] Davison, A. C. and Smith, R. L. 1990. Models for exceedances over high thresholds, *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 393–442.

[34] De Haan, L. 1984. A spectral representation for max-stable processes, *The Annals of Probability* pp. 1194–1204.

[35] de Haan, L. and Ferreira, A. 2006. *Extreme value theory*, Springer.

[36] de Saint-Exupéry, A. 1943. *Le petit prince*, Vol. 1, Oscar Gm, Sergio Pg.

[37] de Wiljes, J., Majda, A. J. and Horenko, I. 2013. An adaptive markov chain monte carlo approach to time series clustering of processes with regime transition behavior, *Multiscale Modeling & Simulation* **11**(2): 415–441.

[38] de Wiljes, J., Putzig, L. and Horenko, I. 2014. Discrete nonhomogeneous and nonstationary logistic and markov regression models for spatiotemporal data with unresolved external influences, *Communications in Applied Mathematics and Computational Science* **9**(1): 1–46.

[39] Diaz, H. F. and Murnane, R. J. 2008. *Climate extremes and society*, Cambridge University Press.

[40] Embrechts, P., Klüppelberg, C. and Mikosch, T. 2001. *Modeling Extremal Events*, 8th edn, Springer, Stochastic Modelling and Applied Probability.

[41] Engl, H. W., Hanke, M. and Neubauer, A. 1996. *Regularization of inverse problems*, Vol. 375, Springer.

[42] Fasen, V., Klüppelberg, C. and Menzel, A. 2014. *Quantifying extreme events*, Springer, Heidelberg. In: Klüppelberg, C., Straub, D. and Welpe, I. (Eds.), Risk - A Multidisciplinary Introduction.

[43] Feudale, L. and Shukla, J. 2011. Influence of sea surface temperature on the european heat wave of 2003 summer. part i: an observational study, *Climate dynamics* **36**(9-10): 1691–1703.

[44] Fisher, R. A. and Tippett, L. H. C. 1928. Limiting forms of the frequency distribution of the largest or smallest member of a sample, *Mathematical Proceedings of the Cambridge Philosophical Society*, Vol. 24, Cambridge Univ Press, pp. 180–190.

[45] Floudas, C. A. and Pardalos, P. M. 2008. *Encyclopedia of optimization*, Vol. 1, Springer.

[46] Fréchet, M. 1927. Sur la loi de probabilité de lâĂŹécart maximum, *Annales de la societe Polonaise de Mathematique*, Vol. 6, Bibliothèque des Sciences Humaines, Editions Gallimard, pp. 93–116.

[47] Fröhlich, C. 2000. Observations of irradiance variations, *Space Science Reviews* **94**: 15–24.

[48] Fröhlich, C. 2006. Solar irradiance variability since 1978: Revision of the PMOD composite during solar cycle 21, *Space Science Reviews* **125**(1-4).

[49] Fuentes, M., Henry, J. and Reich, B. 2012. Nonparametric spatial models for extremes: Application to extreme temperature data, *Extremes* pp. 1–27.

[50] Gerber, S. and Horenko, I. 2014. On inference of causality for discrete state models in a multiscale context, *Proceedings of the National Academy of Sciences* (41): 14651–14656.

[51] Ghosh, S. and Mallick, B. K. 2010. Spatio-temporal modeling of extreme precipitation data: A bayesian approach, *Department of Statistics, Texas A&M University College Station* .

[52] Gumbel, E. J. 1958. *Statistics of extremes*, DoverPublications. com.

[53] Gurobi Optimization, I. 2013. Gurobi optimizer reference manual.
URL: *http://www.gurobi.com*

[54] Hadamard, J. S. 1902. Sur les problèmes aux dérivées partielles et leur signification physique, *Princeton University Bulletin,* **13**: 49–52.

[55] Häkkinen, S., Rhines, P. B. and Worthen, D. L. 2011. Atmospheric blocking and atlantic multidecadal ocean variability, *Science* **334**(6056): 665–659.

[56] Hastie, T. J. and Tibshirani, R. J. 1990. *Generalized additive models*, Vol. 43, CRC Press.

[57] Hastie, T., Tibshirani, R. and Friedman, J. 2009. *The Elements of Statistical Learning*, 2nd edn, Springer.

[58] Horenko, I. 2010a. Finite element approach to clustering of multidimensional time series, *SIAM Journal on Scientific Computing* **32**(1): 62–83.

[59] Horenko, I. 2010b. On identification of nonstationary factor models and its application to atmospherical data analysis, *AMS, Journal of the Atmospheric Sciences* **67**(5): 1559–1574.

[60] Horenko, I. 2011. Nonstationarity in multifactor models of discrete jump processes, memory and application to cloud modeling, *AMS, Journal of the Atmospheric Sciences* **68**(7): 1493–1506.

[61] Hurvich, C. M. and Tsai, C.-L. 1989. Regression and time series model selection in small samples, *Biometrika* **76**(2): 297–307.

[62] Huser, R. and Davison, A. C. 2014. Space-time modelling of extreme events, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **76**(2): 439–461.

[63] Kabluchko, Z., Schlather, M. and de Haan, L. 2009. Stationary max-stable fields associated to negative definite functions, *The Annals of Probability* pp. 2042–2065.

[64] Kaiser, O. and Horenko, I. 2014. On inference of statistical regression models for extreme events based on incomplete observation data, *Communications in Applied Mathematics and Computational Science* **9**(1).

[65] Kaiser, O., Igdalov, D. and Horenko, I. 2014. Statistical regression analysis of threshold excesses with systematically missing covariates, *submitted to Multiscale Modeling and Simulation* .

[66] Kane, R. P. and de Paula, E. R. 1996. Atmospheric co2 changes at mauna loa, hawaii, *Journal of Atmospheric and Terrestrial Physics* **58**(15): 1673 – 1681.

[67] Knippertz, P., Ulbrich, U., Marques, F. and Corte-Real, J. 2003. Decadal changes in the link between el niño and springtime north atlantic oscillation and european–north african rainfall, *International journal of climatology* **23**(11): 1293–1311.

[68] Kotz, S. and Nadarajah, S. 1988. *Multivariate Extreme-Value Theory*, Wiley Online Library.

[69] Kozek, W. 1993. Optimally karhunen-loeve-like stft expansion of nonstationary processes, **4**: 428–431 vol.4.

[70] Liang, F., Liu, C. and Carroll, R. 2010. *Advanced Markov Chain Monte Carlo methods: learning from past samples*, 1 edn, John Wiley and Sons Ltd.

[71] Lima, A. R., Cannon, A. J. and Hsieh, W. W. 2012. Nonlinear regression in environmental sciences by support vector machines combined with evolutionary strategy, *Computers & Geosciences* .

[72] Lindeberg, J. W. 1922. Eine neue herleitung des exponentialgesetzes in der wahrscheinlichkeitsrechnung, *Mathematische Zeitschrift* **15**(1): 211–225.

[73] Loève, M. 1978. *Probability Theory*, Vol. II of Graduate Texts in Mathematics, 4 edn, Springer-Verlag.

[74] MacDonald, A., Scarrott, C. J., Lee, D., Darlow, B., Reale, M. and Russell, G. 2011. A flexible extreme value mixture model, *Computational Statistics & Data Analysis* **55**(6): 2137–2157.

[75] Majda, A. J., Abramov, R. V. and Grote, M. J. 2005. *Information theory and stochastics for multiscale nonlinear systems*, American Mathematical Society.

[76] Malik, N., Bookhagen, B., Marwan, N. and Kurths, J. 2012. Analysis of spatial and temporal extreme monsoonal rainfall over south asia using complex networks, *Climate dynamics* **39**(3-4): 971–987.

[77] Meerbach, E., Dittmer, E., Horenko, I. and Schütte, C. 2006. Multiscale modelling in molecular dynamics: Biomolecular conformations as metastable states, *Computer Simulations in Condensed Matter Systems: From Materials to Chemical Biology Volume 1*, Springer, pp. 495–517.

[78] Metzner, P., Putzig, L. and Horenko, I. 2012. Analysis of persistent nonstationary time series and applications, *Communications in Applied Mathematics and Computational Science* **7**(2): 175–229.

[79] Mood, A. M., Graybill, F. A. and Boes, D. 1974. *Introduction to the theory of statistics*, 3rd ed edn, New York : McGraw-Hill.

[80] Nelsen, R. B. 2006. *An introduction to copulas*, Springer.

[81] Neuneier, R., Hergert, F., Finnoff, W. and Ormoneit, D. 1994. Estimation of conditional densities: A comparison of neural network approaches, *ICANNâĂŹ94*, Springer, pp. 689–692.

[82] Neville, S. E., Palmer, M. J. and Wand, M. P. 2011. Generalized extreme value additive model analysis via mean field variational bayes, *Australian New Zealand Journal of Statistics* **53**(3): 305–330.

[83] Nolan, J. P. 2013. *Stable Distributions - Models for Heavy Tailed Data*, Birkhauser, Boston. In progress, Chapter 1 online at academic2.american.edu/∼jpnolan.

[84] Ouachani, R., Bargaoui, Z. and Ouarda, T. 2013. Power of teleconnection patterns on precipitation and streamflow variability of upper medjerda basin, *International Journal of Climatology* **33**(1): 58–76.

[85] Padoan, S. A., Ribatet, M. and Sisson, S. A. 2010. Likelihood-based inference for max-stable processes, *Journal of the American Statistical Association* **105**(489): 263–277.

[86] Padoan, S. and Wand, M. 2008. Mixed model-based additive models for sample extremes, *Statistics & Probability Letters* **78**(17): 2850–2858.

[87] Pierrehumbert, R. T. 2001. Energy balance models, *Technical report*.
URL: *https://www.whoi.edu/fileserver.do?id=21420pt=10p=17292*

[88] R Core Team 2013. *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
URL: *http://www.R-project.org/*

[89] Reich, B. J. and Shaby, B. A. 2012. A hierarchical max-stable spatial model for extreme precipitation, *The Annals of Applied Statistics* **6**(4): 1430–1451.

[90] Resnick, S. I. 2007. *Extreme values, regular variation, and point processes*, Springer.

[91] Ribatet, M., Cooley, D. and Davison, A. C. 2009. Bayesian inference from composite likelihoods, with an application to spatial extremes, *arXiv preprint arXiv:0911.5357* .

[92] Ribatet, M. and Sedki, M. 2013. Extreme value copulas and max-stable processes, *Journal de la Société Française de Statistique* **154**(1): 138–150.

[93] Rodó, X., Baert, E. and Comin, F. 1997. Variations in seasonal rainfall in southern europe during the present century: relationships with the north atlantic oscillation and the el niño-southern oscillation, *Climate Dynamics* **13**(4): 275–284.

[94] Romano, J. P. and Siegel, A. F. 1986. *Counterexamples in probability and statistics*, CRC Press.

[95] Sang, H. and Gelfand, A. E. 2010. Continuous spatial process models for spatial extreme values, *Journal of agricultural, biological, and environmental statistics* **15**(1): 49–65.

[96] Scarrott, C. and macdonald2011flexible, A. 2012. A review of extreme value threshold estimation and uncertainty quantification, *REVSTAT–Statistical Journal* **10**(1): 33–60.

[97] Schlather, M. 2002. Models for stationary max-stable random fields, *Extremes* **5**(1): 33–44.

[98] Smith, R. L. 1985. Maximum likelihood estimation in a class of nonregular cases, *Biometrika* **72**(1): 67–90.

[99] Stephenson., A. G. 2012. *ismev: An Introduction to Statistical Modeling of Extreme Values*. Original S functions written by Janet E. Heffernan, R package version 1.39.
URL: *http://CRAN.R-project.org/package=ismev*

[100] Tank, A. M. G. K., Wijngaard, J. B., Können, G. P., Böhm, R., Demarée, G., Gocheva, A., Mileta, M., Pashiardis, S., Hejkrlik, L., Kern-Hansen, C., Heino, R., Bessemoulin, P., Müller-Westermeier, G., Tzanakou, M., Szalai, S., Pálsdóttir, T., Fitzgerald, D., Rubin, S., Capaldo, M., Maugeri, M., Leitass, A., Bukantis, A., Aberfeld, R., van Engelen, A. F. V., Forland, E., Mietus, M., Coelho, F., Mares, C., Razuvaev, V., Nieplova, E., Cegnar, T., Antonio López, J., Dahlström, B., Moberg, A., Kirchhofer, W., Ceylan, A., Pachaliuk, O., Alexander, L. V. and Petrovic, P. 2002. Daily dataset of 20th-century surface air temperature and precipitation series for the european climate assessment, *International Journal of Climatology* **22**(12): 1441–1453.

[101] Tarantola, A. 2005. *Inverse problem theory and methods for model parameter estimation*, SIAM.

[102] Teng, H., Branstator, G., Wang, H., Meehl, G. A. and Washington, W. M. 2013. Probability of us heat waves affected by a subseasonal planetary wave pattern, *Nature Geoscience* **6**(12): 1056–1061.

[103] The MathWorks, I. 2013. Global optimization toolbox, user's giude 2013a.
URL: *http://www.mathworks.com*

[104] Thibaud, E., Mutzner, R. and Davison, A. C. 2013. Threshold modeling of extreme spatial rainfall, *Water Resources Research* .

[105] Tibshirani, R. 1996. Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 267–288.

[106] Trenberth, K. E. 1997. The definition of el ninõ, *Bulletin of the American Meteorological Society* **78**.

[107] Uchaikin, V. V. and Zolotarev, V. M. 1999. *Chance and Stability: Stable Distributions and their applications*, Walter de Gruyter.

[108] Umbricht, A., Fukutome, S., Liniger, M. A., Frei, C. and Appenzeller, C. 2013. Seasonal variation of daily extreme precipitation, *Technical report*.

[109] Varin, C., Reid, N. and Firth, D. 2011. An overview of composite likelihood methods, *Statistica Sinica* **21**(1): 5–42.

[110] Wahba, G. 1990. *Spline models for observational data*, Vol. 59, SIAM.

[111] Wood, S. 2006. *Generalized Additive Models: An Introduction with R*, Chapman and Hall/CRC.

[112] Wood, S. N. 2003. Thin-plate regression splines, *Journal of the Royal Statistical Society (B)* **65**(1): 95–114.

[113] Wood, S. N. 2011. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models, *Journal of the Royal Statistical Society (B)* **73**(1): 3–36.

[114] Yee, T. W. and Stephenson, A. G. 2007. Vector generalized linear and additive extreme value models, *Extremes* **10**(1-2): 1–19.

[115] Zhang, Z. and Smith, R. L. 2010. On the estimation and application of max-stable processes, *Journal of Statistical Planning and Inference* **140**(5): 1135–1153.

[116] Zolotarev, V. M. 1986. *One-dimensional stable distributions*, Vol. 65, American Mathematical Soc.