Developments in statistics applied to hydrometeorology : imputation of streamflow data and semiparametric precipitation modeling

Patricia Tencaliec

▶ **To cite this version:**

Patricia Tencaliec. Developments in statistics applied to hydrometeorology : imputation of streamflow data and semiparametric precipitation modeling. Statistics [math.ST]. Université Grenoble Alpes, 2017. English. NNT : 2017GREAM006 . tel-01684069v2

HAL Id: tel-01684069

https://theses.hal.science/tel-01684069v2

Submitted on 11 Jan 2018 (v2), last revised 11 Jan 2018 (v4)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Communauté UNIVERSITÉ Grenoble Alpes

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ GRENOBLE ALPES

Spécialité : **Mathématiques Appliquées**

Arrêté ministérial :

Présentée par

Patricia Tencaliec

Thèse dirigée par **Clémentine Prieur**
et codirigée par **Anne-Catherine Favre**

préparée au sein **du Laboratoire Jean Kuntzmann**
et de **l'Ecole Doctorale Mathématiques, Sciences et Technologies de l'Information, Informatique**

Developments in statistics applied to hydrometeorology: imputation of streamflow data and semiparametric precipitation modeling

Thèse soutenue publiquement le **01/02/2017**,
devant le jury composé de :

Stéphane Girard
Directeur de recherche, Inria Grenoble Rhône-Alpes, Président
Véronique Maume-Deschamps
Professeur, Université Claude Bernard Lyon 1, Rapporteur
Valérie Monbet
Professeur, Université de Rennes 1, Rapporteur
Philippe Naveau
Directeur de recherche, LSCE CNRS, Examinateur
Benjamin Renard
Chargé de recherche, Irstea, Examinateur
Clémentine Prieur
Professeur, Université Grenoble Alpes, Directeur de thèse
Anne-Catherine Favre
Professeur, Grenoble INP, Co-Directeur de thèse
Thibault Mathevet
Ingénieur docteur, EDF-DTG, Invité

Părinților mei, pentru toată susținerea și încrederea oferită...

Vă mulțumesc!

Acknowledgements

Firstly, I would like to express my sincere appreciation to my two supervisors, Prof. Clémentine Prieur and Prof. Anne-Catherine Favre, for their patience and continuous support throughout the time of my PhD. Their constant feedback and guidance helped me grow as a researcher. Clémentine, thank you for being so organized and for always paying attention to the very last detail. Your perfectionism, even though at times overwhelming, certainly helped in keeping this work on the right path. Anne-Catherine, thank you for always having an encouraging attitude and especially for "running the last km of the PhD marathon" with me, for all the weekends and evenings spent reviewing my manuscript, they were much appreciated.

Besides my supervisors, I would like to thank Prof. Philippe Naveau for his brilliant comments and suggestions, but also for the challenging questions which pushed me to widen my research perspectives and to improve the quality of my work. I am grateful also to all the members of my dissertation jury for the careful reading of the manuscript, the corrections and suggestions that greatly improved the readability of the thesis.

My acknowledgments go also to Thibault Mathevet, for his valuable input and support received through the collaborative work during the first year of my PhD. Moreover, I would like to thank Anna Kuentz, for her remarkable work on historical streamflow reconstruction which allowed me to initiate and develop the first part of this work.

I am grateful to all the administrative staff from LJK, especially to our team assistant Anne, for taking care of all the bureaucratic aspects during my time at LJK, and thus allowing me to focus on my research.

I thank my fellow PhD students, especially Nhu, Konstantina and Federico, with whom I shared this ride from the beginning, for the stimulating discussions, all the lunches, dinners and hikes. This, without doubt, made my PhD program much more enjoyable.

I would also like to say a profound thank you to my parents, for their constant unconditional support and for always being there for me. Also, many thanks to all my family and friends, for always believing in me and encouraging me to follow my dreams. Thank you Bogdi, Ana, Daniel, Andrea, Ion, Sergiu, for visiting me several times in Grenoble these past three years. It was good to have family and friends close by when you spend your time so far away from home. Thank you Kalle for organizing all those wonderful hikes and for encouraging me to do more sport, it definitely boosted my energy and helped me clear my mind each week.

Finally, I would like to acknowledge the most important person in my life, Cătă, who encouraged me to join this journey in the first place. He has been a constant source of strength and inspiration, and without his help I would not have become who I am today. Thank you Cătă for all your support, all your love, and most important, all your delicious meals. Words cannot express the feelings and gratitude I have for you.

Patricia Tencaliec
Grenoble, France
November 2016

Abstract

Precipitation and streamflow are the two most important meteorological and hydrological variables when analyzing river watersheds. They provide fundamental insights for water resources management, design, or planning, such as urban water supplies, hydropower, forecast of flood or droughts events, or irrigation systems for agriculture.

In this PhD thesis we approach two different problems. The first one originates from the study of observed streamflow data. In order to properly characterize the overall behavior of a watershed, long datasets spanning tens of years are needed. However, the quality of the measurement dataset decreases the further we go back in time, and blocks of data of different lengths are missing from the dataset. These missing intervals represent a loss of information and can cause erroneous summary data interpretation or unreliable scientific analysis.

The method that we propose for approaching the problem of streamflow imputation is based on dynamic regression models (DRMs), more specifically, a multiple linear regression with ARIMA residual modeling. Unlike previous studies that address either the inclusion of multiple explanatory variables or the modeling of the residuals from a simple linear regression, the use of DRMs allows to take into account both aspects. We apply this method for reconstructing the data of eight stations situated in the Durance watershed in the south-east of France, each containing daily streamflow measurements over a period of 107 years. By applying the proposed method, we manage to reconstruct the data without making use of additional variables, like other models require. We compare the results of our model with the ones obtained from a complex approach based on analogs coupled to a hydrological model and a nearest-neighbor approach, respectively. In the majority of cases, DRMs show an increased performance when reconstructing missing values blocks of various lengths, in some of the cases ranging up to 20 years.

The second problem that we approach in this PhD thesis addresses the statistical modeling of precipitation amounts. The research area regarding this topic is currently very active as the distribution of precipitation is a heavy-tailed one, and at the moment, there is no general method for modeling the entire range of data with high performance. Recently, in order to propose a method that models the full-range precipitation amounts, a new class of distribution called extended generalized Pareto distribution (EGPD) was introduced, specifically with focus on the EGPD models based on parametric families. These models provide an improved performance when compared to previously proposed distributions, however, they lack flexibility in modeling the bulk of the distribution. We want to improve, through, this aspect by proposing in the second part of the thesis, two new models relying on semiparametric methods.

The first method that we develop is the transformed kernel estimator based on the EGPD transformation. That is, we propose an estimator obtained by, first, transforming the data with the EGPD cdf, and then, estimating the density of the transformed data by applying a nonparametric kernel density estimator. We compare the results of the proposed method with the ones obtained by applying EGPD on several simulated scenarios, as well as on two precipitation datasets from south-east of France. The results show that the proposed method behaves better than parametric EGPD, the MIAE of the density being in all the cases almost

twice as small.

A second approach consists of a new model from the general EGPD class, *i.e.*, we consider a semiparametric EGPD based on Bernstein polynomials, more specifically, we use a sparse mixture of beta densities. Once again, we compare our results with the ones obtained by EGPD on both simulated and real datasets. As before, the MIAE of the density is considerably reduced, this effect being even more obvious as the sample size increases.

Key words: streamflow imputation, dynamic regression models, statistical modeling of precipitation amounts, extended generalized Pareto distribution, Bernstein polynomials, nonparametric kernel estimator

Résumé

Les précipitations et les débits des cours d'eau constituent les deux variables hydrométéorologiques les plus importantes pour l'analyse des bassins versants. Ils fournissent des informations fondamentales pour la gestion intégrée des ressources en eau, telles que l'approvisionnement en eau potable, l'hydroélectricité, les prévisions d'inondations ou de sécheresses ou les systèmes d'irrigation.

Dans cette thèse de doctorat sont abordés deux problèmes distincts. Le premier prend sa source dans l'étude des débits des cours d'eau. Dans le but de bien caractériser le comportement global d'un bassin versant, de longues séries temporelles de débit couvrant plusieurs dizaines d'années sont nécessaires. Cependant les données manquantes constatées dans les séries représentent une perte d'information et de fiabilité, et peuvent entraîner une interprétation erronée des caractéristiques statistiques des données. La méthode que nous proposons pour aborder le problème de l'imputation des débits se base sur des modèles de régression dynamique (DRM), plus spécifiquement, une régression linéaire multiple couplée à une modélisation des résidus de type ARIMA. Contrairement aux études antérieures portant sur l'inclusion de variables explicatives multiples ou la modélisation des résidus à partir d'une régression linéaire simple, l'utilisation des DRMs permet de prendre en compte les deux aspects. Nous appliquons cette méthode pour reconstruire les données journalières de débit à huit stations situées dans le bassin versant de la Durance (France), sur une période de 107 ans. En appliquant la méthode proposée, nous parvenons à reconstituer les débits sans utiliser d'autres variables explicatives. Nous comparons les résultats de notre modèle avec ceux obtenus à partir d'un modèle complexe basé sur les analogues et la modélisation hydrologique et d'une approche basée sur le plus proche voisin. Dans la majorité des cas, les DRMs montrent une meilleure performance lors de la reconstitution de périodes de données manquantes de tailles différentes, dans certains cas pouvant allant jusqu'à 20 ans.

Le deuxième problème que nous considérons dans cette thèse concerne la modélisation statistique des quantités de précipitations. La recherche dans ce domaine est actuellement très active car la distribution des précipitations exhibe une queue supérieure lourde et, au début de cette thèse, il n'existait aucune méthode satisfaisante permettant de modéliser toute la gamme des précipitations. Récemment, une nouvelle classe de distribution paramétrique, appelée distribution généralisée de Pareto étendue (EGPD), a été développée dans ce but. Cette distribution exhibe une meilleure performance, mais elle manque de flexibilité pour modéliser la partie centrale de la distribution. Dans le but d'améliorer la flexibilité, nous développons, deux nouveaux modèles reposant sur des méthodes semiparamétriques.

Le premier estimateur développé transforme d'abord les données avec la distribution cumulative EGPD puis estime la densité des données transformées en appliquant un estimateur nonparamétrique par noyau. Nous comparons les résultats de la méthode proposée avec ceux obtenus en appliquant la distribution EGPD paramétrique sur plusieurs simulations, ainsi que sur deux séries de précipitations au sud-est de la France. Les résultats montrent que la méthode proposée se comporte mieux que l'EGPD, l'erreur absolue moyenne intégrée (MIAE) de la densité étant dans tous les cas presque deux fois inférieure.

Le deuxième modèle considère une distribution EGPD semiparamétrique basée sur les polynômes de Bernstein. Plus précisément, nous utilisons un mélange creuse de densités béta. De même, nous comparons nos résultats avec ceux obtenus par la distribution EGPD paramétrique sur des jeux de données simulés et réels. Comme précédemment, le MIAE de la densité est considérablement réduit, cet effet étant encore plus évident à mesure que la taille de l'échantillon augmente.

Mots clés: imputation des débits, modèles de régression dynamique, modélisation statistique des quantités de précipitations, distribution généralisée de Pareto étendue, polynômes de Bernstein, estimateur nonparamétrique par noyau

Contents

List of Tables

List of Figures

List of Abbreviations

PART 1: Dynamic regression models for the imputation of streamflow data

ACF	autocorrelation function
ADF	augmented Dickey-Fuller stationarity test
AIC	Akaike information criterion
BIC	bayesian information criterion
DRM	dynamic regression models
EDF	Électricité de France
KGE	Kling-Gupta efficiency
KPSS	Kwiatkowski-Phillips-Schmidt-Shin stationarity test
M.NS.0lag	model with no-season split and no lags for the explanatory variables
M.NS.1lag	model with no-season split and one lag for the explanatory variables
M.2S.0lag	model with 2-season split and no lags for the explanatory variables
M.2S.1lag	model with 2-season split and one lag for the explanatory variables
MLE	maximum likelihood estimation
MLR	multiple linear regression
OLS	ordinary least squares
PACF	partial autocorrelation function
PAM	partitioning around medoids
(S)ARIMA	(seasonal) autoregressive moving average model
(S)AR	(seasonal) autoregressive model
(S)MA	(seasonal) moving average model
VIF	variance inflation factor

PART 2: Flexible semiparametric approaches to model the full-range of precipitation amounts

BCTK	boundary-corrected transformed kernel
BIC	bayesian information criterion
cdf	cumulative distribution function
E1	parameter estimation approach for EGPD-BB: the case when the entire problem is optimized at once
$\mathbf{E2_{EGPD}}$	parameter estimation approach for EGPD-BB: the case when the problem is split in two subproblems and the GPD parameters are initially set to their estimated values from the EGPD model

E2$_{\text{fix}}$	parameter estimation approach for EGPD-BB: the case when the problem is split in two subproblems and the GPD parameters are initially set to their true values
E2$_{\text{grid}}$	parameter estimation approach for EGPD-BB: the case when the problem is split in two subproblems and the GPD parameters are taken from a grid
ecdf	empirical cumulative distribution function
EGPD	extended generalized Pareto distribution class
EGPD$_{Naveau}$	parametric EGPD models introduced by *Naveau et al.* (2016)
EGPD$_1$	parametric EGPD model introduced by *Naveau et al.* (2016) based on $G(u) = u^\kappa$
EGPD-BB	semiparametric EGPD model based on Bernstein-beta density estimator
ERM	empirical risk minimization
EVT	extreme value theory
GEV	generalized extreme value distribution
GIC	generalized information criteria (model selection criteria)
GP(D)	generalized Pareto (distribution)
H_ξ / h_ξ	cdf / pdf of the generalized Pareto distribution with shape parameter ξ
IAE	integrated absolute error
i.i.d.	independent and identically distributed (random variables)
LSCV	least squares cross-validation
MIAE	mean integrated absolute error
MISE	mean integrated squared error
Mix2GaGPD	mixture of two gamma and a generalized Pareto densities
Mix3GaGPD	mixture of three gamma and a generalized Pareto densities
Mix2SM	mixture of two Singh-Maddala densities
MLE	maximum likelihood estimator
MSE	mean squared error
pdf	probability density function
q_p	quantile for probability p, considering that $q_p = F^{-1}(p)$
QQ-plot	quantile-quantile plot
RMSE	root mean squared error
R$_{\text{RMSE}}$	ratio of the root mean squared errors
RMSE$_{\text{X}}$	root mean squared error of the model X
TK	transformed kernel density estimator
TK$_1$	transformed kernel density estimator based on the EGPD$_1$ transformation

General introduction

Hydrology is the study of the water cycle, more exactly its movement, distribution, and quality throughout earth. The water cycle or hydrological cycle ensures that the water is continuously moved around the earth, through different pathways and physical processes, such as precipitation, evaporation, infiltration, surface runoff, or subsurface flow.

Water is essential for human survival and well-being, ecosystems endurance, but also it is important for many economical sectors. However, water resources are unevenly distributed in space and time. Moreover, due to the pressure exerted by humans on the environment, such as population growth, urbanization, higher living standards, pollution, or deforestation, water resources demands are increasing every day. At the same time, extreme weather events are more frequent and catastrophic around the world, having a major impact on water availability and quality.

As a result, the growing demand for the limited water resources requires a deeper understanding of the underlying hydrological processes. A rigorous analysis of the hydrological variables, their risk assessment (*e.g.*, floods, droughts, erosion), or forecasting, all depend on reliable information about the quality and quantity of water available, but also how this availability changes in time and space.

The work presented in this PhD thesis addresses practical topics that arise in the process of statistical modeling and analyzing observed data related to the hydrological processes, such as streamflow and precipitation. While there are numerous topics that can be approached around this subject, in the following two sections we highlight the problems that we were faced with and the methods that we chose to tackle them.

Streamflow

Streamflow is one of the important variables when performing hydrological analysis of a watershed. By definition, streamflow refers to the flow of water in streams, rivers, and other water paths, and is a major element of the water cycle.

Streamflow is a highly variable quantity and it can vary from very small values in periods of extended drought, to extremely large values during the rainy seasons or when the mountain snow melts in the spring. This variability can result in episodes such as, long shortages of fresh water supply or, at the other extreme, flooding, both of them with catastrophic impact on population, economy, as well as bio-ecosystems.

Therefore, water resources management and planning is a task that needs to be carefully approached. To be able to do so, one needs access to reliable datasets providing measurements over long periods of time. Besides water resources management, extreme flood/drought prediction, streamflow forecast and climate variability analysis, all require reliable time series. Since extreme events are seldom by definition, long and continuous time series spanning tens of years are necessary, as they allow for a more accurate characterization of the watershed operation.

For example, Figure G.1 shows the volume of water flowing down the Durance river as measured at one of the stations used in our study, *i.e.*, Durance at Val-des-Près. The data span

seven years, from 2002 to 2008, and as we can see from the figure, there is a high variability between observations. The overall pattern, or hydrological cycle, appears to be the same for every year, *i.e.*, high values in late spring and beginning of summer, and smaller ones in the winter, which fits the profile of the Durance as its flow is highly influenced by the snow melts. However, while the yearly pattern analyzed from a bird's eye view might present similarities, when analyzed closely we can see that the differences between actual values from different years cannot be neglected. For example, one can see a large difference between year 2007 and 2008 in Figure G.1.

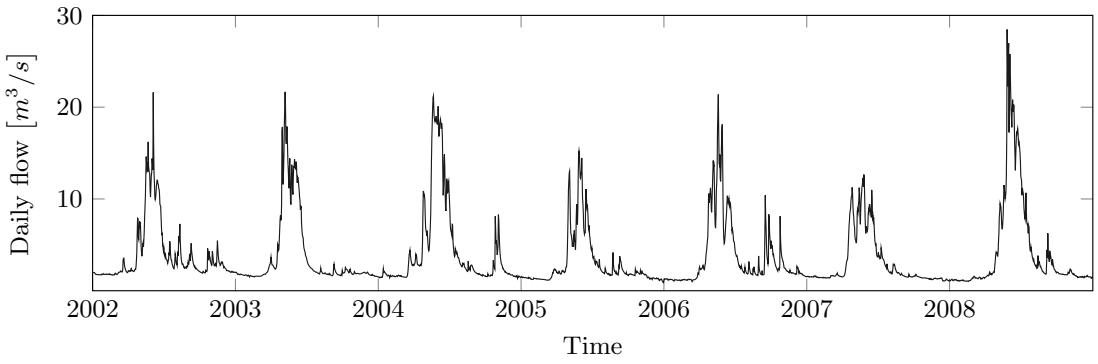

Figure G.1: Observed streamflow measurements of the Durance river between 2002 and 2008 at the Durance Val-des-Près station

In streamflow analysis a dataset of seven years is considered rather short and does not contain sufficient information to characterize the overall behavior of the watershed. This is especially true for extreme events, such as, severe droughts and floods, that might not even appear in such a short time span. That is why, it is usually required to work with longer datasets, usually spanning tens of years. However, as we go further back in time, the reliability of the measurement dataset decreases. In practice, one should expect the actual dataset to look more like the one in Figure G.2. Here, data from the same measurement station is presented, but now going back as far as 1904. We can notice that there are blocks of data of different lengths (from a few observations to entire years) missing from the dataset. These missing intervals in the time series represent a loss of information and can cause erroneous summary data interpretation or unreliable scientific analysis.

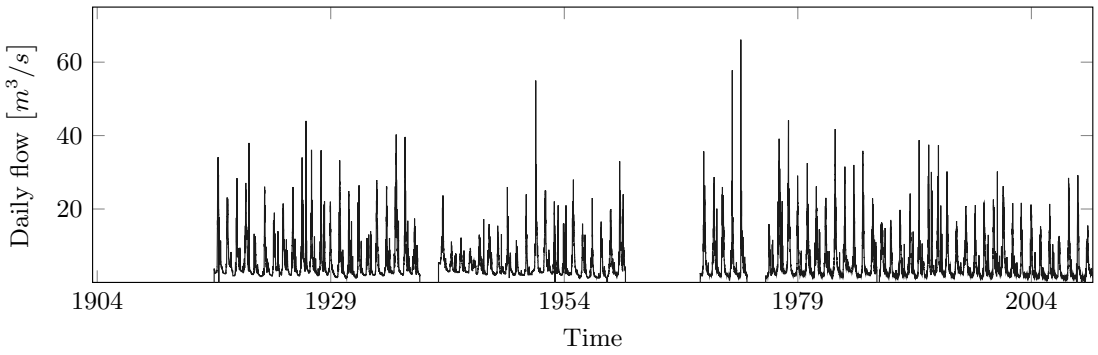

Figure G.2: Observed streamflow measurements with missing intervals of the Durance river between 1904 and 2010 at the Durance Val-des-Près station

Missing data blocks of very short length are not a serious impediment for data analysis, as they can be easily corrected and they do not influence the overall statistical image of the dataset. However, longer periods of missing data are a serious issue, as analysis performed using these data may be biased. There are several reasons for missing periods of measurement data. One of them is when an extreme weather event damages the measuring equipment. In this case, important data that characterize rare events, and that would be of out most importance in certain studies, is lost. Other causes for missing data could be due to equipment failure either at collection or at storage, or they might be human-induced, *e.g.,* incorrect handling of data by field personnel or as a result of devastating events, such as wars. Reconstructing these datasets becomes a challenging prerequisite to any hydrological study.

Naive methods for dealing with incomplete datasets could be used. However, they are rather limited in their application and present several disadvantages. For example, the easiest approach is to ignore the missing data blocks. This could work for a few missing random observations, but it will certainly fail for extended missing periods. Other basic methods would be to replace the missing observation with the mean of previous observations, or to carry the last observation forward. Once again, these methods might work for time series with relatively small gaps of missing observations, but in the case of large blocks of missing data, the variance of the substituted observations will be underestimated.

Giving the importance of dataset reconstruction and the limited performance of naive methods, many researchers became interested in this topic. Several methods of various complexity, such as averaging nearest neighbors, regression, autoregressive models, state-space models, or most recently, artificial intelligence or machine learning are among the methods that have been previously applied to tackle this problem. However, the majority of these methods depend on the existence of additional variables such as rainfall or temperature. This represents a major drawback if measurement of these additional variables do not exist or one does not have access to them.

One of the main objectives of this thesis is to reconstruct observed streamflow data from several correlated hydrometric stations situated along the Durance watershed in the south-east part of France. The dataset contains missing periods of various lengths, and we address the case when one has available for the analysis only streamflow data. This consideration is not far from reality, as frequently, we do not have access to long-historical data for other variables, like for example precipitation.

Our approach is based on dynamic regression models (DRMs), more specifically, a multiple linear regression with ARIMA (autoregressive integrated moving average) residual modeling. Unlike previous studies that address either the inclusion of multiple explanatory variables or the modeling of the residuals from a simple linear regression, the use of DRMs allows to take into account both aspects, and thus improves the performance of the model without adding excessive complexity.

Therefore, DRM is the technique that we employ in order to approach the problem of missing data reconstruction, and as it can be seen in the case studies presented in Part I of the thesis, the performance of this model is relatively high. We compared the results of our model with the ones obtained from a complex model based on analogs and a hydrological model (ANATEM), as well as with the results obtained from a nearest-neighbor approach. In the majority of the cases, DRMs displayed better estimation error when reconstructing missing values blocks of various lengths, in some of the cases ranging up to 20 years.

Precipitation

Another important variable in the water cycle is precipitation. Precipitation measurements are widely used as input in hydrological models. They are often needed in many applications

regarding water resources management, design, or planning, such as urban water supplies, hydropower, forecast of flood or droughts events, irrigation systems, agriculture, etc.

One of the active topics of research in hydrology regarding precipitation is to find a probability distribution that can describe the overall behavior of precipitation at a station. The most common approach in developing such a stochastic model is to firstly describe the process of rainfall occurrence and, then, to employ a probability distribution function in order to characterize rainfall amount on wet days. Both steps in this procedure are important in order to get a valid model, but different statistical tools are used to approach each of them. The process of rainfall occurrence is a discrete process, meanwhile, rainfall amount is as a continuous process. In the work presented in this thesis we are focused only on the second step of this process, *i.e,* statistical modeling of rainfall amount.

Establishing a probability distribution that provides a good fit for the rainfall amount has proven to be a challenging task, mainly due to the fact that rainfall amounts are heavily skewed to the right. Different distributions, such as Weibull, gamma, exponential, or lognormal have been considered as possible candidates, with gamma and exponential being typically the preferred choices. Figure G.3 shows the histogram of hourly rainfall measurements from the Lyon station located in the south-east region of France from 1996 to 2011. It also illustrates how a gamma distribution fits these data. What can be noticed from the quantile-quantile plot (QQ-plot) is that, while the lower tail and the first part of the bulk (center part of the distribution) are properly estimated, the performance of the model decreases significantly towards the upper tail.

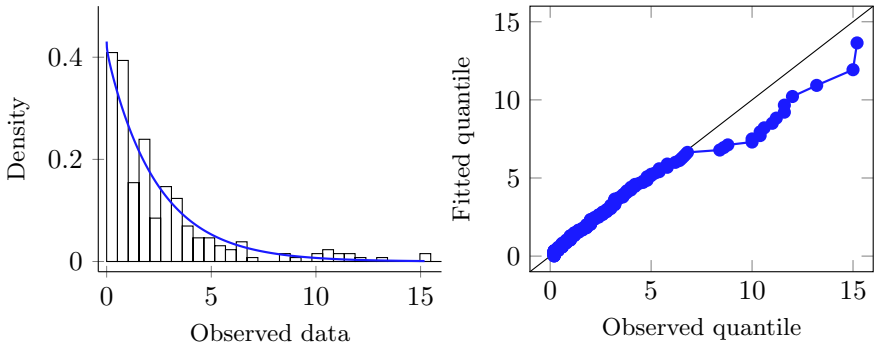

Figure G.3: Normalized histogram with gamma density fit and QQ-plot of hourly rainfall data from the summer months (June, July, August) at the Lyon station from 1996 to 2011

As the upper tail of the distribution holds crucial information that characterizes extreme events, many researchers became interested in studying only the behavior of the largest rainfall intensities. The popular framework of Extreme Value Theory (EVT), more exactly the Generalized Pareto Distribution (GPD) has been quickly adopted in this sense. However, beside the considerable reduction in sample size, a major drawback of this approach is the need of a threshold selection, *i.e,* the limit that differentiates between large and moderate rainfall amount. Defining this threshold is a delicate task in the field of EVT, since it has a major impact on the capability of the model in describing the extreme events.

While characterizing extreme events is essential, also one cannot totally neglect the remaining values. Several applications such as water resource management requires not only a clear understanding of extreme events, but rather a global assessment of rainfall amount. Extreme mixture models have been proposed in the literature in this sense, the latest being the extended generalized Pareto distribution (EGPD) based on parametric families, published by *Naveau et al.* (2016). We illustrate the formulation of the cdf and pdf of this model in (G.1).

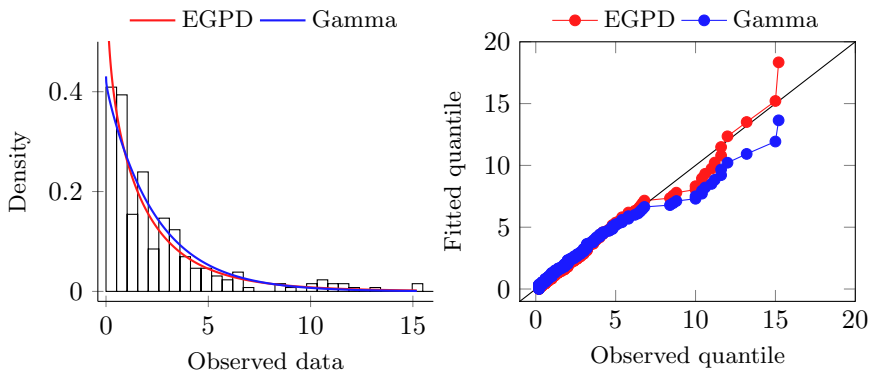

Figure G.4: Normalized histogram with density fit and QQ-plot for both gamma and parametric EGPD model, based on hourly rainfall data from the summer months (June, July, August) at the Lyon station from 1996 to 2011

$$\begin{aligned} F(x) &= G\left\{H_\xi(x)\right\}, \\ f(x) &= g\left\{H_\xi(x)\right\} \cdot h_\xi(x). \end{aligned} \tag{G.1}$$

Here, h_ξ and H_ξ represent the pdf and cdf of the GPD, while g and G denote a continuous pdf and cdf on the unit interval. This model is in compliance with the EVT for both lower and upper tails, and at the same time, it allows a smooth transition between the two ends through the G function. Figure G.4 shows the improvement that the parametric EGPD brings when compared to the gamma distribution in the modeling of the entire distribution.

However, EGPD lacks flexibility in modeling the bulk of the distribution. More specifically, the parametric families employed until now for the G function (*e.g.*, power function) are too stringent, *i.e.*, they do not allow for a flexible modeling of the bulk. This weakness of the model is studied and improved through the two methods proposed in the second part of the thesis, both of them relying on semiparametric approaches.

The first method that we develop is the transformed nonparametric kernel estimator based on EGPD transformation. We propose an estimator obtained by, first, transforming the data with the EGPD cdf from (G.1), and then, estimating the density of the transformed data by applying a nonparametric kernel density estimator. We compare the results of the proposed method with the ones obtained by applying the parametric EGPD model on several simulated scenarios from mixture of distributions that account for different degrees of "bumpiness" in the bulk, as well as on two precipitation datasets (one from the Durance watershed and one from Lyon). The results show that the proposed method performs better than the parametric EGPD, since the mean integrated absolute error (MIAE) of the density is nearly 50% smaller in all the test cases investigated.

As an alternative approach for introducing more flexibility in the EGPD model for the bulk of the distribution, we consider the case when G comes from a different family of distribution than the parametric one, *i.e.*, a semiparametric model based on Bernstein polynomials. More specifically, relying on the relationship between the beta distribution and Bernstein polynomials, we use a sparse mixture of beta densities. Once again, we compare our results with the ones obtained by the parametric EGPD on both simulated and real datasets. As before, the MIAE of the density is considerably reduced. This effect is more obvious as the sample size increases. For medium size and large datasets the MIAE of the proposed method was up to five times smaller than the one of the parametric EGPD.

Research objectives and outline

In the previous two sections we gave a brief introduction into current research topics related to hydrology, also investigated in this thesis. Our work is focused on two main axis that also guided this manuscript to be organized in two distinct parts.

Part 1. Dynamic regression models for the imputation of streamflow data

In Chapter 1 we present the problem background and make a review of the previous research. We follow up in Chapter 2 with a detailed presentation and exploratory analysis of the measurements dataset used in this thesis, and its particularities. This chapter is essential, as it was used to define the main research questions and problems. We were faced with large blocks of missing data, and the only variables that we had access to in order to reconstruct these data were the streamflow measurements from the different stations of the watershed.

We decided to approach the problem of data imputation by using dynamic regression models (DRM), for which we present the theoretical backgrounds in Chapter 3. This type of models have been previously used in other fields, but up until now, to the best of our knowledge, there was no attempt to apply them in hydrology. This method incorporates the advantages of multiple linear regression and residual modeling, and as can be seen from the case study presented in Chapter 4, it provides superior results when compared to other existing methods.

Chapter 5 ends this part of the thesis by summarizing the main conclusions and possible future work.

The main results obtained in this part were published in:

Tencaliec, P., Favre, A. C., Prieur, C., & Mathevet, T. (2015). *Reconstruction of missing daily streamflow data using dynamic regression models.* Water Resources Research, 51(12), 9447-9463.

Part 2. Flexible semiparametric approaches to model the full-range of precipitation amounts

The second part of this thesis focuses on improvements that can be brought to the parametric EGPD model in the context of rainfall amounts modeling. We focused on EGPD as it is a novel and a promising approach for modeling entire rainfall amounts distributions. Although it provides improved performance when compared to previous models, the parametric EGPD lacks flexibility in modeling the bulk of the rainfall distribution. This is where we concentrated our efforts in order to improve the estimates of EGPD.

As in the previous part, we start with setting the background and presenting the state of the art in Chapter 1.

Chapter 2 presents the first approach that we propose for improving the behavior of the parametric EGPD, namely a method based on a nonparametric kernel density estimator. The results presented in this chapter show that the proposed method performs better than the parametric EGPD, as the MIAE of the density is reduced by half.

As an alternative model, we introduce a semiparametric EGPD model based on Bernstein polynomials, more specifically, we rely a sparse mixture of beta densities. This method is presented in Chapter 3, and the comparison with the parametric EGPD shows that the MIAE of the density is considerably reduced.

Chapter 4 is dedicated to summarizing the main conclusions of this part, as well as for presenting some future work ideas.

The main results of this part are currently being developed into two articles, intended to be submitted in the next two months.

1. Tencaliec, P., Naveau, P., Favre, A.-C., Prieur, C. *Kernel density estimation with EGPD transformation for modeling the full range of rainfall amount*, (to be submitted)

2. Tencaliec, P., Naveau, P., Prieur, C., Favre, A.-C. *Modeling full range rainfall amount using sparse semiparametric mixture models,* (to be submitted).

Part I

Dynamic regression models for the imputation of streamflow data

Chapter 1

Introduction

River discharge is one of the most important quantities in hydrology. It provides fundamental records for water resources management and climate change monitoring, either as indicator of past hydrological variability or as contributor to future behavior prediction. Even very short data-gaps in this information can cause extremely different analysis results. Therefore, reconstructing missing data of incomplete datasets is an important step and it can affect the performance of the environmental models, engineering and research applications. In Section 1.1 we review some of the most applied approaches for streamflow times series reconstruction. Then, in Section 1.2 we shift the focus towards our proposed technique for streamflow imputation, i.e., dynamic regression model, and give a background on some previous applications of this model.

1.1 Background on streamflow reconstruction

Hydrology is the study of water, more exactly its movement, distribution, and quality throughout earth. Hydrology is concerned with the water cycle, water resources and environmental sustainability. The water cycle ensures that the water is continuously cycled around the earth, through different pathways and physical processes, such as evaporation, condensation, precipitation, infiltration, surface runoff, and subsurface flow.

Water is essential for human survival and well-being, ecosystems endurance, but also it is important for many economical sectors. However, water resources are unevenly distributed in space and time and, moreover, due to different factors such as population growth, urbanization, higher living standards, pollution, deforestation or climate change, water resources are facing serious threats.

The current growing demand for these limited resources continues to rise as the population increases and shifts. A solid and sustainable management relies on reliable information about the quality and quantity of water available, but also how this availability changes in time and space. Therefore, it is important to have a clear understanding of all the elements of the water cycle and how they interact.

The main variables that describe the hydrological functioning of water bodies are air temperature, precipitation, soil moisture and streamflow. Numerous research and operational applications, such as water resources management and planning, extreme flood or drought anticipation, streamflow forecast and climate variability analysis, require reliable time series. Since extreme events are seldom by definition, long and continuous time series are necessary, allowing a more accurate analysis of watershed operation.

There are several causes that can create discontinuities in data series. They might be missing due to equipment failure either at collection or at storage, as consequence of extreme weather, or

there might be human-induced causes, like wars, incorrect handling of data by field personnel, etc. These missing intervals in the time series represent a loss of information and can cause erroneous summary data interpretation or unreliable scientific analysis.

Consequently, in order to obtain reliable and accurate information from the data, these gaps must be filled. Despite the vast research on this subject, the estimation of missing intervals, also known in the literature as imputation (*Schneider* (2001)), infilling (*Harvey et al.* (2012)) or reconstruction (*Kim and Pachepsky* (2010)), represents a great challenge in hydrology and geosciences in general.

Numerous, yet contrasting, reconstruction methods have been proposed for streamflow data, such as deterministic and stochastic, parametric and nonparametric, linear and nonlinear, etc. However, we can classify them in two main groups: stochastic models (*i.e.*, models based on data connected through statistical and mathematical concepts) and hydrological or process-based model (*i.e.*, models that represent the physical processes from the real world). Our interest is on the former one, *i.e.*, stochastic models.

Hydrological models are usually of increased complexity and require knowledge on various processes related to streamflow, such as precipitation, evaporation, or temperature, among others. They are tuned and constructed locally, *i.e.*, for specific catchment areas, so it is usually difficult to generalize these models over different sites. Beside this, the fact that they require other collections of related meteorological variables is another possible drawback.

Considering the complexity of the hydrological models, many researches have become interested in statistical models. There are several methods reported in the literature from this category. Among these, we remind the works of *Hirsch* (1979), that discuss multiple infilling approaches for daily data using data from the nearby station(s), along with their basin characteristics, such as drainage area, river length, basin elevation, etc. They proved that useful reconstructions can be obtained even with few data at the target station, but by making use of the watershed information from neighbor stations. The work of *Wallis et al.* (1991) is similar, they replace the missing value of a target station with the weighted daily streamflow data from several neighbor stations. The weights for each neighbor station are computed as the ratio between the monthly mean flow of the target and neighbor station.

Different approaches based on simple linear regression or regression with residuals modeling are presented in *Raman et al.* (1995). Although their models are simple to apply, they require as well the existence of rainfall data series which is used in the regression model. This can be a disadvantage if one does not have access to such series. On the same line, *Woodhouse et al.* (2006) works with a multiple linear regression model with forward stepwise predictor selection.

Reviews studies by *Gyau-Boakye and Schultz* (1994) and *Harvey et al.* (2012) summarize and compare several methods used for infilling flow data. *Gyau-Boakye and Schultz* (1994) compare ten widely known techniques including interpolation, recursive models, autoregressive models, regression and nonlinear models. Their results show that the model choice is influenced by the length of the estimation period or by the season, but on average, interpolation and multiple regression models yield good results. *Harvey et al.* (2012) propose an extended description of approaches used in hydrology for missing data imputation or prediction, along with an applied comparison of simple and multiple regression models. It was proved that one can have a better accuracy if multiple explanatory variables are included.

More recent studies present procedures for filling missing hydrological data by using state-space models with Estimation-Maximization (EM) algorithm as in *Amisigo and van de Giesen* (2005). The authors make use of both spatial and temporal information and create a dynamic model. An important disadvantage of their approach is that it is suitable for short and medium term missing data, *i.e.*, from days up to a month.

Many researches use linear models in order to infill missing data in streamflow time series. These models are simple to apply and, with enough explanatory variables, they provide suitable estimates. As mentioned in *Elshorbagy et al.* (2002), in many applications nonlinear

models might provide better results. For example, the superiority of artificial neural networks (ANN) over autoregressive moving average models with explanatory variables was proved in *Hsu et al.* (1995) or over linear regression in *Elshorbagy et al.* (2000), *Khalil et al.* (2001), *Panu et al.* (2000). However, this class of models requires a deeper knowledge in computer sciences. Moreover, ANN models have no strong theoretical assumptions and the output is difficult to interpret, being often considered black-box models.

Currently, there is no precise method that is generally applied in streamflow imputation. The choice of the method depends on several factors, such as, the number and nature of missing observations, availability of data from correlated neighbor stations or from other meteorological variables. An imputation method should be used only after a careful analysis of the available data and of the missing data pattern.

1.2 Streamflow imputation and dynamic regression models

We have seen earlier that previous works addressed the infilling of flow data by using the multiple linear regression (*Gyau-Boakye and Schultz* (1994), *Harvey et al.* (2012), *Woodhouse et al.* (2006)) or even simple linear regression with residual modeling (*Raman et al.* (1995)), but none approached the problem as a multiple linear regression with residual modeling. While the streamflow models found in the literature address only one aspect of the prior problem formulation, *i.e.*, either the inclusion of multiple inputs or the modeling of the residuals from a regression with only one input, the use of dynamic regression models (DRMs) allow for both aspects to be taken into account.

Our main objective is to reconstruct streamflow data from several correlated hydrometric stations that contain missing intervals of various lengths. We address the case when one has available for the analysis only streamflow data. This consideration is founded as, frequently, we do not have access to long-historical data for other variables, like for example precipitation. Consequently, we consider DRMs a promising candidate for solving this problem.

Generally, a DRM is a system where an output A at one time step (*i.e.,* streamflow data at time t) can be linked to the output A at some past time, or to other variables (*i.e.,* streamflow data from other correlated stations) from the same period (*i.e.,* time t) or from a past time (*i.e.,* $t-1, t-2...$). Beside this, it also adjusts the correlation from the remainder part (residuals) by fitting an autoregressive integrated moving average (ARIMA) structure. More details on this matter are presented in Chapter 3, Section 3.1.

The DRMs have been used before by *Tsay* (1984) to model the monthly highway traffic volume in Taiwan, by *Greenhouse et al.* (1987) to fit biological rhythm data, by *Miaou* (1990) to estimate the water demand in some states of the USA, or, more recently, by *Bercu and Proïa* (2013) to forecast energy consumption in France or by *Nogales et al.* (2002), *Vagropoulos et al.* (2016) to forecast next-day electricity prices in Spain and USA and PV generation in Greece, respectively.

Chapter 2

Data presentation and exploratory analysis

The reconstruction of missing streamflow data is a complex process which requires careful attention not only in the modeling step, but also in pre-modeling. Understanding the origin and behavior of the data might help to construct and comprehend the performance of the model and derive more pertinent conclusions about the results. In Section 2.1 we present the streamflow time series that will be reconstructed and also analyze the pattern of the missing data. Then, in Section 2.2 we carry out an exploratory analysis and particularly look at the correlation and similarities between the stations used in the study by applying statistical and hydrological analyses.

2.1 Data presentation

2.1.1 Durance watershed

The application study of this work is done on the Durance watershed. Situated in the south-east region of France, the Durance river is the second largest tributary of the Rhône, after Saône. It has a length of more than 300[km] and a catchment area of more than $14\,000\,[\mathrm{km}^2]$. It has the source in the massif of Écrins at an altitude of 4102[m], and it flows into the Rhône river, near Avignon.

The Durance watershed is divided into three geographical areas: upper, middle and lower basin. The upper Durance is characterized by a mountainous area with abrupt valleys, while the middle part has a lower altitude and the valleys are wider, about 60% of its drainage area is under 1000[m]. The lower Durance is the smallest, with a catchment area of no more than $3600\,[\mathrm{km}^2]$. It is composed mainly of dry lowland, but it still remains mainly in a hilly area. In Figure 1.2.1 we illustrate the location of the watershed.

There are more than 50 hydrometric stations within the watershed managed either by Électricité de France (EDF) or by the Regional Department of Environment, Planning and Housing of Provence-Alpes-Côte-d'Azur region (DREAL PACA). For this study we selected eight stations situated in the upper and middle regions of the Durance. Our selection is based on the length of the time series, *i.e.*, all these stations have a data sample longer than 100 years, which is seldom in hydrology. The exact location of each station is showed in Figure 1.2.2, while their main characteristics are presented in Table 1.2.1.

The Durance watershed is defined by its many uses, which makes it one of the most important rivers in southern France. It offers many purposes like hydropower generation, irrigations, water

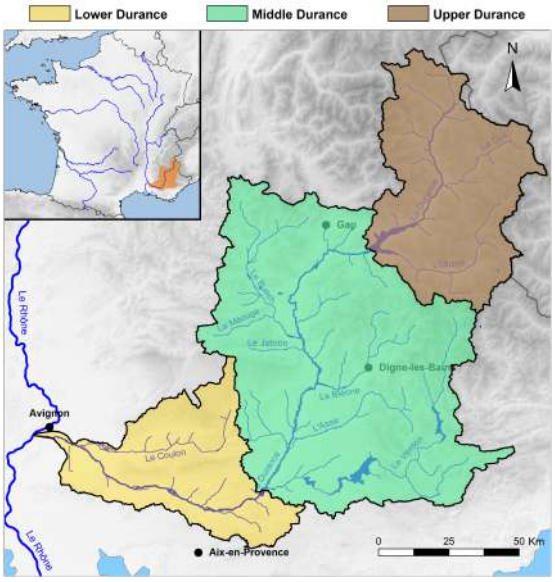

Figure 1.2.1: Location and drainage area of the Durance watershed (source: *Kuentz* (2013))

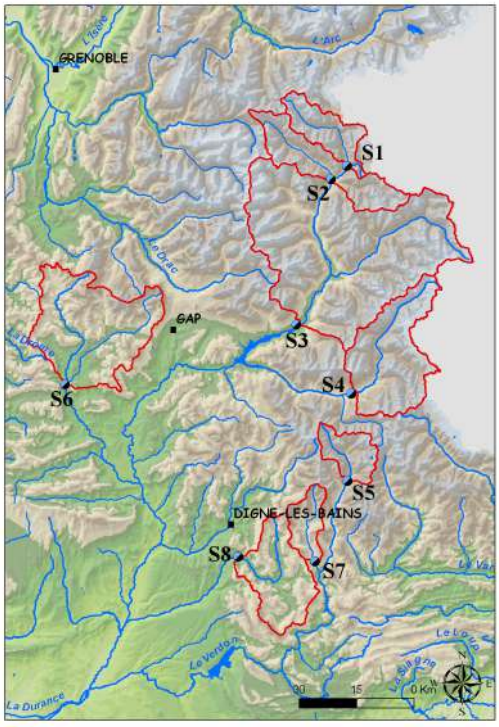

Figure 1.2.2: Location and drainage area of the eight stations of interest from the Durance watershed

Table 1.2.1: Main characteristics of the eight stations of interest from the Durance watershed

Code	Station name	In service from	Location	Altitude [m]	Area [km^2]
S1	Durance (Val-des-Près)	1917	Upper	1360	203
S2	Durance (Briançon)	1905	Upper	1187	548
S3	Durance (La Clapière)	1903	Upper	787	2170
S4	Ubaye (Barcelonnette)	1903	Upper	1132	549
S5	Verdon (Colmars)	1903	Middle	1230	158
S6	Buech (Chambons)	1905	Middle	662	723
S7	Issole (Saint-André-les-Alpes)	1904	Middle	931	137
S8	Asse (Clue de Chabrières)	1906	Middle	605	375

supply for cities like Marseille and Aix-en-Provence or tourism near the lakes. Furthermore, due to its mixed climatological environment (from a nival regime in the north-east area to a mediterranean-pluvial in the south area), along with the geographical and functional complexity, the analysis of the Durance river is challenging.

2.1.2 Durance data

The observations for the flow data are provided by Électricité de France (EDF) or are taken from the HYDRO database (http://www.hydro.eaufrance.fr/), depending on the station and period. We used in this study the daily flow measurements starting from 1904 until 2010, thus 107 years.

The measurement installations are situated on the rivers and most of them provide natural flow data. In the early period (from 1904 to ∼1950, depending on the stations), the river stage measurements were made by daily human observation, then between ∼1950 to ∼1980 by using a limnigraph (device for automatically recording the water level) and lastly, since ∼1980, by using an electronic data-logger. These stations were installed at the beginning of the 20^{th} century in order to help the French administration issue flood alerts (*Imbeaux* (1892)) and improve the understanding of the hydroelectric potential of the Durance watershed. An extensive part of these streamflow time series (*i.e.,* the early decades) had to be restored from different archives through a documentary research, see *Kuentz* (2013), *Kuentz et al.* (2013, 2014) for details. These studies provide an extended characterization of the hydrometeorological variability of the Durance watershed during the last century, and also give an historical review about the measurement procedures at each station.

2.1.2.1 Homogeneity study

In time series analysis, we are concerned with the stability or homogeneity of the stochastic process over time. There are several factors that can drive a hydrological random variable to become heterogeneous (not stable over time), such as climate change, relocation of the stations, or, as seen earlier, changes in the measurement techniques, among others. This difference in measurements can create heterogeneity and, thus, have an impact on the analysis of the streamflow data.

A test for homogeneity is equivalent to a test of a statistical distribution, *i.e.*, we want to detect possible shifts in time of the mean and the variance of the random variable. To address this aspect we followed the two-step approach presented in *Wijngaard et al.* (2003) for testing homogeneity in daily temperature and precipitation series. Same workflow was applied later by *Vezzoli et al.* (2012) for daily discharge data, or by *Kang and Yusof* (2012) for a hydrometeorological dataset with missing values. The approach consists of:

Table 1.2.2: Test statistic values for SNH, Buishand, Pettitt, and Von Neumann homogeneity tests and the overall classification into the three classes: useful, doubtful, suspect, for each of the eight stations of interest from the Durance watershed. In the squared brackets, we indicate if the null hypothesis is accepted (A) or rejected (R, in red), meanwhile last row contains the 1% critical values for each test.

Station	SNH	B	P	VN	Classification
S1	$3.99_{[A]}$	$1.58_{[A]}$	$349_{[A]}$	$1.36_{[R]}$	useful
S2	$11.85_{[A]}$	$1.69_{[A]}$	$811_{[A]}$	$1.58_{[A]}$	useful
S3	$4.19_{[A]}$	$1.24_{[A]}$	$545_{[A]}$	$1.67_{[A]}$	useful
S4	$6.47_{[A]}$	$1.25_{[A]}$	$589_{[A]}$	$2.03_{[A]}$	useful
S5	$3.74_{[A]}$	$1.78_{[A]}$	$462_{[A]}$	$1.29_{[R]}$	useful
S6	$2.49_{[A]}$	$1.03_{[A]}$	$322_{[A]}$	$2.03_{[A]}$	useful
S7	$5.50_{[A]}$	$1.12_{[A]}$	$385_{[A]}$	$1.92_{[A]}$	useful
S8	$6.80_{[A]}$	$1.45_{[A]}$	$680_{[A]}$	$1.46_{[R]}$	useful
critical values	12.32	1.86	841	1.54	

1. Applying four homogeneity tests: standard normal homogeneity test (SNH), Buishand range test (B), Pettitt test (P), and von Neumann ratio test (VN) to evaluate the series.

2. Classifying these tests results into three classes:

   - useful (homogeneous data): at most one test rejects the null hypothesis at the 1% level (*i.e.*, the test statistic is above the critical 1% level value)
   - doubtful: two tests reject the null hypothesis at the 1% level
   - suspect (inhomogeneous data): three or four tests reject the null hypothesis at the 1% level

As suggested in the works mentioned above, the variable to be tested is the annual maximum series of daily streamflow data. We take into consideration only the years that have no missing data. All tests assume under the null hypothesis that the annual values are independent and identically distributed, *i.e.*, no break in the mean. Under the alternative hypothesis, the first three tests assume a break in the mean, while the last one assumes that the series is not randomly distributed, *i.e.*, the observations are correlated. Details about each test can be found in the references mentioned earlier.

In Table 1.2.2 we present the results of these tests for the eights stations from the Durance watershed. An "A" label means that the test statistic is not significant at a 1% level and the null hypothesis of no break in the mean is accepted, while a label "R" means the test statistic is significant, thus the null hypothesis is rejected. We have three stations (S1, S5 and S8) that reject the von Neumann ratio test, meaning that these time series might show some time correlation. However, the overall classification of all the stations is as "useful" at 1% significance level, thus we can say that the data are homogeneous.

Remark 1.2.1. As shown in *Yozgatligil and Yazici* (2016), the homogeneity of a time series can be also studied by means of the stationarity tests such as Augmented Dickey-Fuller (ADF) or Kwiatkowski-Phillips-Schmidt-Shin (KPSS), not commonly used in homogeneity detection analysis. Similarly to the von Neumann ratio test used in this study, these tests are not location-specific tests, *i.e.*, there is no indication of the exact change point, but they test the existence of trend in a time series.

2.1.2.2 Overview of the missing data pattern

The missing data for the Durance watershed are mainly due to absence of human reading (early period), extreme weather events, technical/maintenance issues, or disturbances during the 2nd World War (*Kuentz et al.* (2014)). Consequently, these data contain a large number of missing points, especially at the beginning of the period and around 1940-1960.

Table 1.2.3: Main characteristics of the eight stations of interest from the Durance watershed

Code	Station name	Missing data	
		#	%
S1	Durance (Val-des-Près)	9217	24%
S2	Durance (Briançon)	4900	13%
S3	Durance (La Clapière)	5903	15%
S4	Ubaye (Barcelonnette)	1207	3%
S5	Verdon (Colmars)	3340	9%
S6	Buech (Chambons)	5473	14%
S7	Issole (Saint-André-les-Alpes)	9711	25%
S8	Asse (Clue de Chabrières)	7067	18%

The percentage of missing data for the eight stations ranges from 3% to 25%, as shown in Table 1.2.3. In Figure 1.2.3 one can find the pattern of the missing data for each station for the entire period 1904-2010. We can notice that for some intervals, such as 1948-1951, out of the total number of 11 688 daily observations for the eight stations, only 4383 observations are available, *i.e.*, less than 40%.

2.2 Exploratory analysis

To determine the relationships and correlations between the eight stations, an exploratory data analysis was used to determine possible similarities among variables (stations) and, eventually, to group them based on their characteristics. This part is important as it offers an initial selection for the input variables in the regression models. We used both hydrological and statistical tools, such as hydrological regimes, correlation or clustering analyses.

2.2.1 Hydrological regimes

First, we look at the monthly mean flow (hydrological regime) and observe the behavior of the station. An illustration of the hydrological regimes for the eight stations is shown in Figure 1.2.4. It can be seen that each station has two periods of high flow: one in the autumn and one in spring or summer depending on the station.

For the stations from upper Durance (S1-S4), the autumn peak (*i.e.*, October) is due to strong rainfall events, while the other one, much higher this time, is located at the beginning of the summer in June and it occurs because of the snowmelt from the mountainous areas. Meanwhile, the stations from middle Durance (S5-S8) have one peak at the end of autumn, again due to rainfall accumulations, and the second one around middle/end spring due to an early snowmelt.

These results are consistent with the climate of the area and the elevation ranges of the watersheds. The stations from upper Durance are located in a rocky mountain area at altitudes ranging from more than 4000[m] to around 600[m], where, besides the rainfall events in autumn, most of the precipitations fall as snow, from the end of fall to the beginning of spring. For this reason, we will have a snow regime characterized by very high flow at the beginning of the summer due to snow and glaciers melt and dry winter (low flow). Moreover, the stations from

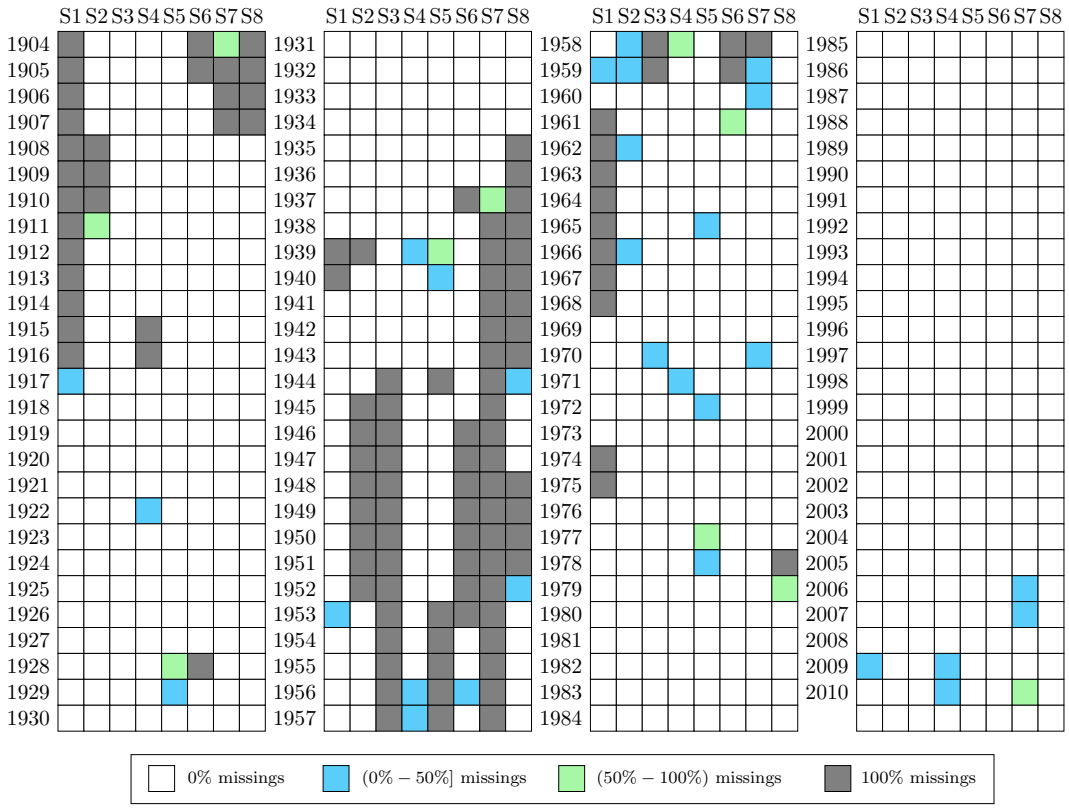

Figure 1.2.3: Missing data pattern from 1904-2010 for the eight stations of interest from the Durance watershed

middle Durance are located at altitudes lower than 2500[m], with more than 60% of them being at under 1000[m]. Thus, we will observe, on the one hand that the autumn rain is increasing and lasting until the first part of the winter, but on the other hand the snowmelt process is starting earlier, like in May for S5 and in April for S6, S7, S8. This specific behavior is called rain-snow regime.

Therefore, each regime displays two seasons: autumn-winter (referred so far as *cold season*), defined by rain (less in upper Durance, more in middle Durance), and spring-summer (referred so far as *warm season*), defined by snowmelt (earlier or later depending on the altitude).

2.2.2 Correlation analysis

The above statements were also validated by statistical analysis, *i.e.*, correlation analysis. The correlation matrix of the daily flow data is computed using only the complete cases of the dataset (only days with information available for all the stations, *i.e.,* 49.7% of the data). The chosen criterion is Spearman's rank correlation coefficient. It is a nonparametric rank statistic, which assesses how well an arbitrary monotonic function can describe the relationship between two variables, without making any assumption about the distribution of these variables. For more details, the reader is referred to *Lehmann and D'Abrera* (2006).

The results, illustrated in Table 1.2.4, show that all the coefficients are positive with strong correlation (>0.8) between the group of stations S1-S4 and the group S6-S8. Station S5 is a particular case; it has a higher value in relation with S4 and S7, but all its other values are

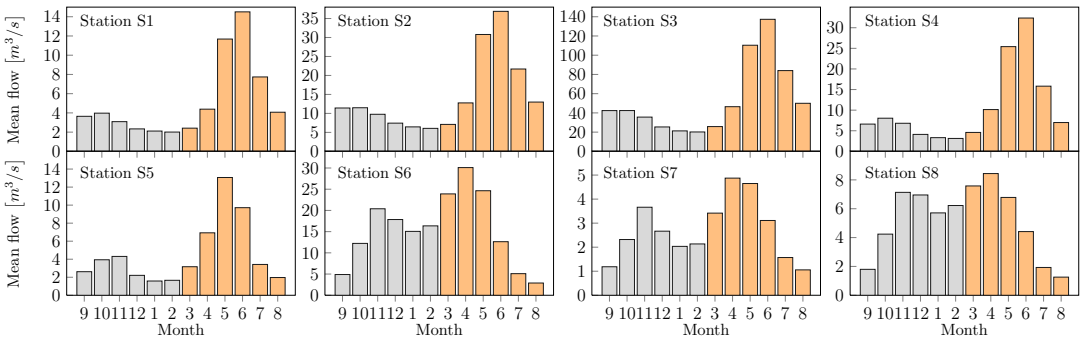

Figure 1.2.4: Hydrological regimes (monthly mean flow) for the eight stations of the Durance watershed (cold season in grey, warm season in orange). *Note:* the months are ordered from September to August for a clearer illustration of the two seasons, and it is usually called a hydrological year.

very close to each other. Assessment of the correlation for each of the two seasons (cold, warm) individually, show that for the cold season there is a decrease in dependence for both upper and middle Durance and S5 tends to be more similar to the middle Durance stations, while for the warm season the groups upper and middle Durance are better split, but station S5 still remains an "in-between" station.

Table 1.2.4: Spearman's rank correlation coefficients of the daily streamflow for: i) all seasons (on the top table), ii) cold season (bottom table, in blue), and iii) warm season (bottom table, in red)

	S2	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**
S1	0.92	0.93	0.87	0.66	0.08	0.28	0.05
S2		0.93	0.87	0.66	0.08	0.29	0.04
S3			0.90	0.68	0.09	0.31	0.04
S4				0.76	0.22	0.46	0.19
S5					0.57	0.73	0.53
S6						0.82	0.85
S7							0.85

	S1	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**
S1		**0.79**	**0.83**	**0.72**	**0.51**	**0.11**	**0.21**	**0.10**
S2	**0.96**		**0.83**	**0.74**	**0.56**	**0.15**	**0.28**	**0.13**
S3	**0.96**	**0.96**		**0.82**	**0.61**	**0.15**	**0.30**	**0.13**
S4	**0.91**	**0.91**	**0.92**		**0.67**	**0.26**	**0.44**	**0.27**
S5	**0.63**	**0.63**	**0.63**	**0.74**		**0.57**	**0.70**	**0.57**
S6	**0.04**	**0.04**	**0.03**	**0.18**	**0.62**		**0.80**	**0.84**
S7	**0.21**	**0.21**	**0.20**	**0.36**	**0.72**	**0.86**		**0.86**
S8	**0.02**	**0.02**	**0.00**	**0.15**	**0.56**	**0.85**	**0.87**	

2.2.3 Clustering analysis

In this part, we consider a clustering technique, called partitioning around medoids (PAM), to classify the stations based on their spatial/temporal characteristics. The idea of this approach is to divide the dataset into groups (clusters) so that the distance between them is minimized.

More specifically, PAM partitions the dataset of n objects into k clusters, where both the dataset and the number of clusters k are inputs of the algorithm. Each cluster is represented by a center called medoid. The algorithm works with a matrix of dissimilarities (distances), and it aims to minimize the overall distance between the medoids of each cluster and its members. It is very similar to the well-known k-means technique, but, in contrast to the k-means, PAM chooses data points as centers of the groups, and not a mean of data points like in the k-means case. Moreover, PAM is more robust to outliers when compared to k-means, because it minimizes a sum of pairwise dissimilarities instead of a sum of squared Euclidean distances. The detailed procedure of the technique can be found in *Kaufman and Rousseeuw* (1990).

To choose the relevant number of clusters and to determine if a station is well classified, we will use the silhouette coefficient, introduced by *Rousseeuw* (1987).

In our particular case, for a station i, the silhouette coefficient is defined as:

$$s_i = \frac{\delta_{-i} - d_{i,c(i)}}{\max\{d_{i,c(i)}, \delta_{-i}\}}, \tag{1.2.1}$$

where $d_{i,c(i)}$ represents the intracluster distance between medoid $c(i)$ and station i, and δ_{-i} corresponds to the smallest intercluster distance, *i.e.*, distance between station i and all the other stations except i.

The silhouette coefficient s_i can range from -1 to 1. That is, when we have:

- $s_i = 1$, it means that intracluster distances ($d_{i,c(i)}$) are much smaller than intercluster ones (δ_{-i}), so we are in the case of "well-classified".
- $s_i = 0$, it means that intracluster and intercluster distances are approximately equal, so we cannot be sure about the membership of station i.
- $s_i = -1$, it is contrary to the first case, we have a much larger intracluster distance compared to the intercluster one, so the classification is not good.

To compute PAM performance of a cluster with k components (stations), the average silhouette index is used: $s(k) = \frac{\sum_{i=1}^{n} s_i}{n}$, where n denotes the total number of stations.

We applied PAM classification on our daily flow data (S1-S8) by using two and three clusters, respectively. The results are illustrated in Figure 1.2.5.

If we look at the "all seasons" case with two clusters, the data are classified as Group 1 = {S1, S2, S3, S4} and Group 2 = {S5, S6, S7, S8}. This division is exactly the geographical split upper-middle Durance. When looking at the silhouette coefficients s_i of each station, we notice that S5 has a negative value, but close to zero (*i.e.,* $s_{S5} = -0.1086$), meaning that it may be not well-classified in Group 2. In the case with three clusters, the stations are classified as follows: Group 1 = {S1, S2, S3, S4}, Group 2 = {S5} and Group 3 = {S6, S7, S8}. In order to distinguish the best cluster dimension, we look at the average silhouette index, *i.e.*, $s(k)$. In our case, for PAM with two and three clusters, these indices are $s(2) = 0.51$ and $s(3) = 0.38$, respectively, meaning that by introducing another group, the clusters are less well defined. It is also interesting to notice that by introducing a third group (*i.e.,* S5), the silhouette index of Group 1 is decreasing from 0.67 to 0.48, while the one of Group 2 increases from 0.36 to 0.37. So, the inclusion of the third group has a larger influence on Group 1, but almost none on Group 2. These results are supported also by the hydrological regimes and the correlation matrix, S5 being an "in-between" station with a special behavior.

The application of PAM on the cold and warm season subsets yields more or less the same results with the same reasoning.

In conclusion, when trying to classify the eight stations, it is clear that the "hazy" behavior of station S5 makes the grouping a little bit uncertain, while the remaining stations preserve the geographical division of upper and middle Durance. These relationships will be used later in the choice of explanatory variables in our regression models.

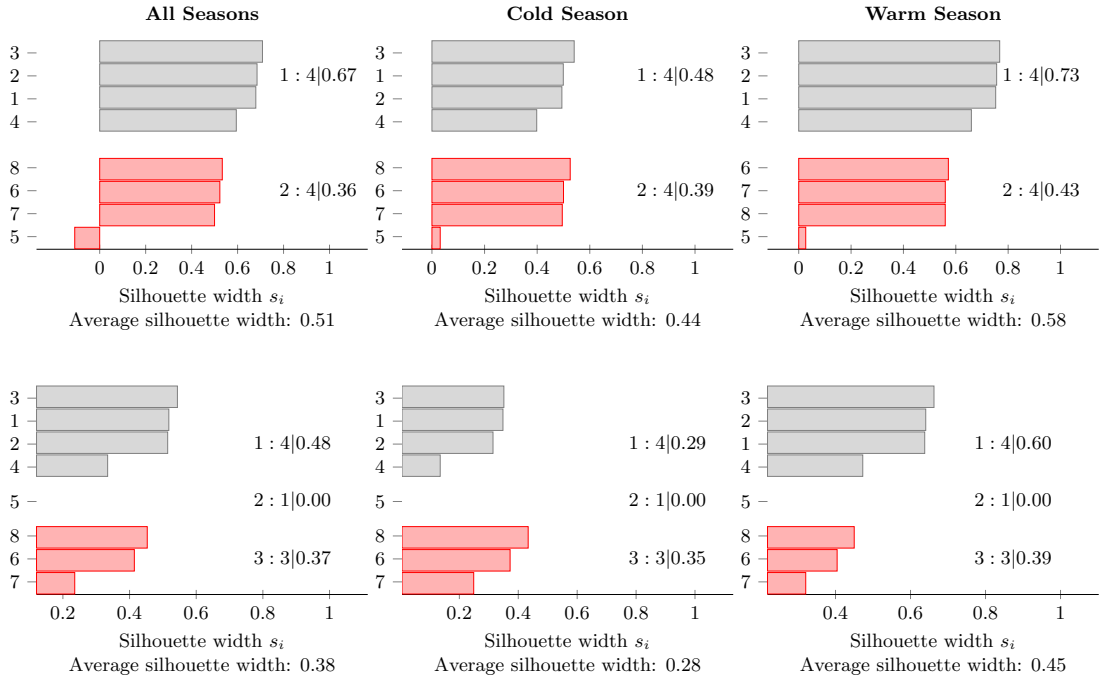

Figure 1.2.5: Results of the PAM classification with two and three clusters, when all, cold and warm seasons are considered, for the eight stations from the Durance watershed (top-plots show the results of PAM with two clusters, while the bottom-plots for PAM with three clusters)

Chapter 3

Statistical modeling with dynamic regression models

In general, the statistical modeling process involves several equally important steps, such as data collection and understanding, model choice and estimation, as well as testing and evaluation of these models. In this chapter we give a detailed methodology of what it means to work with dynamic regression models. It is important to understand the meaning and context of the applied model, so in Section 3.1 we introduce the theoretical background behind the dynamic regression models. Then, in Section 3.2 we give a step-by-step methodology for the model estimation and validation.

3.1 Theoretical background on dynamic regression models

Dynamic regression models, as referred in *Pankratz* (1991), or transfer function models according to *Box and Jenkins* (1976), are a class of statistical models that describe the relationship between a response variable and one or more explanatory variables using a dynamic form. These models can be regarded as extensions of the classical multiple linear regression (MLR) or the autoregressive moving average models (ARIMA). On the one hand, a MLR model includes the information from one or more explanatory variables, but it does not allow the incorporation of the time as a component. On the other hand, an ARIMA model considers the dynamics of time, but it does not include the relevant knowledge from other variables. Thus, by simply combining a MLR model with an ARIMA, it leads to the formulation of a richer class of models called dynamic regression models (DRM).

Remark 1.3.1. In order to avoid any confusion, for the remaining part of this chapter we use *"residual term"* or *"residuals"* to denote the difference between the observations and the estimates obtained by applying a MLR model, and *"error term"* or *"errors"* for the white noise process in the ARIMA model.

Remark 1.3.2. When examining a time series dynamically, it is more convenient to use some operators in order to simplify the notation. One of these operators is the *backshift operator* B, also known as the lag operator L. Some examples of its usage are given below:

- $B^i Y_t = Y_{t-i}$
- $Y_t - Y_{t-1} - Y_{t-2} = (1 - B - B^2)Y_t$
- $Y_t + \alpha Y_{t-1} + \alpha^2 Y_{t-2} = (1 + \alpha B + \alpha^2 B^2)Y_t$

Another useful operator in dynamics modeling is the *difference operator* ∇, and some examples are given in the following:

- $\nabla = (1 - B) \Leftrightarrow \nabla Y_t = (1 - B)Y_t = Y_t - Y_{t-1}$
- $\nabla_s^D = (1 - B^s)^D \Leftrightarrow \nabla_s^D Y_t = (1 - B^s)^D Y_t$

The general formulation of a DRM with l explanatory variables and an ARIMA(p, d, q) model for the residuals, can be expressed as:

$$Y_t = \beta_0 + \alpha_1(B)X_{t,1} + ... + \alpha_l(B)X_{t,l} + Z_t \tag{1.3.1a}$$
$$\phi(B)\nabla^d Z_t = \theta(B)e_t \tag{1.3.1b}$$

Therefore, a DRM has two components: the regression part given by (1.3.1a), and the ARIMA part given by (1.3.1b). In the following, we discuss individually the particularities of each component.

1. **Regression part (1.3.1a)**

   We consider that the response variable Y_t is a linear combination of l explanatory variables (or covariates) $X_{t,1}, ..., X_{t,l}$, and that the residuals of this model are collected in the term Z_t.

   In a DRM, the influence or contribution of each explanatory variable $X_{t,i}$ on the response Y_t has a dynamic structure. More specifically, instead of allowing just a simple coefficient with each covariate, we rather assign a polynomial $\alpha_i(B)$. Thus, this dynamic formulation enables the inclusion of time as a component of the model, *i.e.*, we can relate Y_t with its past values, but also with past values of $X_{t,i}$.

   Therefore, each polynomial $\alpha_i(B)$ (called so far *transfer functions*) has the form:

   $$\alpha_i(B) = \frac{\omega_i(B)}{\delta_i(B)} B^{b_i} \tag{1.3.2}$$

   where both the numerator $\omega_i(B)$ and denominator $\delta_i(B)$ are polynomials as well, denoted by $\omega_i(B) = \omega_{i,0} + \sum_{j=1}^{m_i} \omega_{i,j} B^j$ and $\delta_i(B) = 1 - \sum_{j=1}^{r_i} \delta_{i,j} B^j$. While the numerator gives the time-dynamic formulation for each explanatory variable, the denominator provides the same behavior for the response variable. Moreover, the exponent b_i from (1.3.2) referred as the *delay factor*, denotes the time elapsed until one explanatory variable affects the response.

   In conclusion, the regression component of the DRM requires setting three orders: m_i (order of the numerator's polynomial), r_i (order of the denominator's polynomial) and b_i. These orders are referred in the following as the orders of the transfer function and are noted as (b_i, m_i, r_i). We provide details about the choice of these orders in Section 3.2.1.

2. **ARIMA part (1.3.1b)**

   Apart from the relationships between observations at both present and past time modeled by the regression part, we can use ARIMA to model the correlation between the residuals Z_t. According to *Makridakis et al. (1998)*, an ARIMA model gives the linear combination of the present and past values of both the variable and its errors. More specifically, an ARIMA modeling of the residual Z_t allows the inclusion of both the present and past values of these residuals, but also a linear combination of the present and past values of the error term e_t.

Similar to the regression part, the ARIMA model consists of two polynomials: $\phi(B)$ and $\theta(B)$. On the one hand, the former polynomial, referred as the autoregressive (AR) component, is denoted by $\phi(B) = 1 - \sum_{i=1}^{p} \phi_i B^i$ and shows the linear combination of the past p values of the residual term Z_t. On the other hand, $\theta(B)$, the moving average (MA) part, is denoted by $\theta(B) = 1 - \sum_{i=1}^{q} \theta_i B^i$ and displays the linear combination of the past q values of the errors e_t.

ARIMA models require the time series under study to be stationary, see *Makridakis et al. (1998)*. However, they can also be applied to data that are non-stationary, but in this case an initial differencing or transformation step must be considered in order to remove the non-stationarity. The order of the differencing is accounted directly in the model formulation, and it is denoted by ∇^d, where d represents the differencing order.

Therefore, an ARIMA model also requires the setting of three orders: p (the order of the AR part), q (the order of the MA part) and d (the differencing order), and usually this model is denoted as ARIMA(p, d, q).

So far, we have focused on the non-seasonal time series data, and thus, on non-seasonal ARIMA models. However, ARIMA models can also be adapted for modeling seasonal data, by simply introducing additional seasonal terms for both the AR and MA components. More specifically, a seasonal ARIMA model, denoted as SARIMA$(p, d, q)(P, D, Q)_s$, has the following form:

$$\phi(B)\phi_s(B^s)\nabla^d\nabla_s^D Z_t = \theta(B)\theta_s(B^s)e_t \tag{1.3.3}$$

where $(P, D, Q)_s$ represent the orders of the seasonal components, *i.e.*, seasonal autoregressive (SAR) with order P, seasonal moving average (SMA) with order Q, the seasonal differencing D, and the number of time units (*e.g.*, a week, a month, etc.) per season s. The polynomials associated to the seasonal part, *i.e.*, $\phi_s(B^s)$, $\theta_s(B^s)$ are formulated similar to the non-seasonal one, detailed above.

We provide details about the choice of these orders in Section 3.2.1.

We have presented above a generic formulation of a DRM, but actually several particular cases of this model are frequently applied in time series modeling, known typically by other names. To simplify the notation, we consider the case of stationary time series and only one explanatory variable. This leads (1.3.1a)-(1.3.1b) to the following formulation

$$Y_t = \beta_0 + \frac{\omega(B)}{\delta(B)}X_{t-b} + \frac{\theta(B)}{\phi(B)}e_t \tag{1.3.4}$$

Considering (1.3.4), we can have different particular cases of the DRM:

1. Linear regression models

$$Y_t = \beta_0 + \omega_0 X_t + e_t \tag{1.3.5}$$

where all the orders are null $b = m = r = 0$ and $p = q = 0$, so that $\omega(B) = \omega_0$, $\delta(B) = \phi(B) = \theta(B) = 1$.

2. Distributed lagged regression models

$$Y_t = \beta_0 + \omega(B)X_{t-b} + e_t \tag{1.3.6}$$

where $r = 0$ and $p = q = 0$, so that $\delta(B) = \phi(B) = \theta(B) = 1$.

3. Regression with ARIMA residuals

$$Y_t = \beta_0 + \omega_0 X_t + \frac{\theta(B)}{\phi(B)} e_t \tag{1.3.7}$$

where $b = m = r = 0$, so that $\omega(B) = \omega_0$, $\delta(B) = 1$.

4. ARIMAX models

$$Y_t = \beta_0 + \frac{\omega_0}{\phi(B)} X_t + \frac{\theta(B)}{\phi(B)} e_t \tag{1.3.8}$$

where $b = m = 0$, so that $\omega(B) = \omega_0$ and $\delta(B) = \phi(B)$.

3.2 Dynamic regression model estimation and validation

3.2.1 Model estimation

We have seen in the previous section that a DRM is based on the dependences between the response and the explanatory variables, but it also contains an ARIMA component that models the residuals. Therefore, the estimation phase of a DRM comprises a unified process between these two parts.

An equally important aspect that needs to be addressed before any estimation methodology is the data preprocessing. More specifically, this step might include missing data detection and imputation, transformations, stationarity testing, variable selection, etc., and it must be carefully carried out because an incomplete analysis can lead to inaccurate results. Considering this, the DRM requires two main preprocessing elements, *i.e.*, the selection of the explanatory variables for each model, and stationarity testing, among others.

In order to choose the covariates of each model (*i.e.*, hydrometric station in our case), in Section 2.2 we have discussed and proposed different methods for detecting dependences between variables in a hydrological framework such as correlation or clustering analysis, as well as the study of the hydrological regimes.

After that, the stability of the dataset over time, *i.e.*, the behavior of the series' mean and variance over time, has been approached as well previously in Section 2.1.2.1, where we have pointed out some commonly applied methods in hydrology for testing possible shifts in time of the mean and variance, *i.e.*, standard normal homogeneity test, Buishand range test, Pettitt test, and von Neumann ratio test. Note that, throughout this study, we are interested in and assume a second-order stationarity (*i.e.*, first two moments of a series do not change over time), and not a strict stationary process. Besides the tests mentioned above, we also consider, as suggested by *Makridakis et al. (1998)*, some classical procedures applied in time series analysis to assess the stationarity, *i.e.*, the Augmented Dickey-Fuller (ADF) unit root test introduced by *Said and Dickey* (1984) and the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) stationarity test proposed by *Kwiatkowski et al.* (1992). Given a stationarity test decision, simple methods exist for transforming a non-stationary series into a stationary one. For instance, transformations such as logarithms can help in stabilizing the variance of a time series, while differencing can help maintain a constant mean.

After the preprocessing step is completed, the parameters of the DRM must be estimated, thus, in what follows we focus our attention on this aspect. More specifically, we review each step from this estimation methodology. The entire procedure is illustrated schematically in Figure 1.3.1.

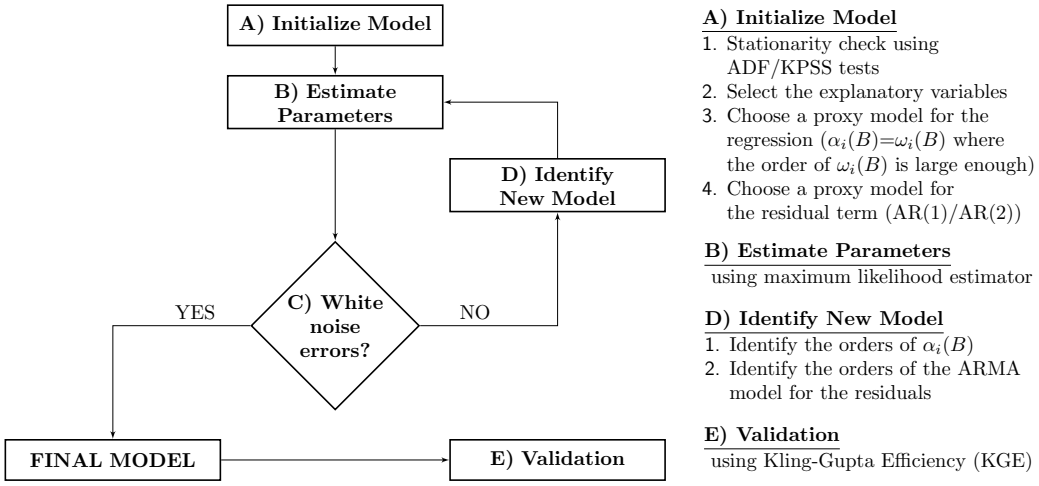

Figure 1.3.1: Schematic representation of the estimation methodology for a dynamic regression model

A. Model initialization

The first step is to choose a proxy model for both the regression and ARIMA parts. As suggested by *Pankratz* (1991), for the regression component it is recommended to consider in the proxy model only the polynomial $\omega(B)$ from (1.3.2), *i.e.*, at first to assume that the response variable is influenced only by the present and past values of its covariates. The order m of this polynomial is suggested (see *Pankratz* (1991)) to be not very large (*e.g.*, in common practice up to order 6-10). Additionally, for the ARIMA part, it is suggested to consider a low-order model, such as an AR(1) or AR(2).

B. Estimation procedure

The parameter estimation could be done by using the ordinary least squares (OLS) technique, if the moving average part of the ARIMA model is not introduced. In case the MA component is required, the problem becomes impossible to solve with OLS as the values for the past errors are unobservable. However, maximum likelihood estimation (MLE) could be applied in this case for the parameter estimation. The likelihood is computed via a state-space representation of the ARIMA process, and the errors and their variance are found by a Kalman filter. Readers can find a good discussion about this approach in *Ripley* (2002), as well as in the R Software documentation regarding the DRMs application, *i.e.*, package *forecast*.

C. Check errors

After each model fit, one should check if the residuals are uncorrelated, *i.e.*, if they are a white noise process (zero mean, finite variance and independent), *i.e.*, $\mathcal{WN}(0, \sigma^2)$. The Ljung-Box test (*Box and Jenkins* (1976), *Box and Pierce* (1970)), with the null hypothesis of independence, is applied to test the serial correlation.

D. Model identification

The initial proxy model might not yield a good fit, as for example the response variable might be influenced also by its past values, so we have to identify a new model. This procedure requires to find first the order of both the transfer function and ARIMA.

Table 1.3.1: Theoretical ACF and PACF

Process	ACF (ρ_k)	PACF (γ_k)
AR(p)	decays to zero	cuts off after lag p ($\gamma_k = 0$, for $k > p$)
MA(q)	cuts off after lag q ($\rho_k = 0$, for $k > q$)	decays to zero

Identification of the transfer function

This step implies the identification of the pairs (b_i, m_i, r_i), the orders of the α_i polynomials in (1.3.2). To find these orders, the pattern (plot) of the estimated coefficients associated with the polynomial $\omega_i(B)$ must be examined for each explanatory variable. There are some identification rules for each order, with reference to the theoretical functions, reported in *Pankratz* (1991), as follows:

(a) b_i (referred as *dead time*) represents the time elapsed until the covariate i affects the response variable. More specifically, it denotes the first non-null position in the sequence of estimated coefficient of the polynomial $\omega_i(B)$.

(b) The denominator factor $\delta_i(B)$ represents the *decay pattern,* and the order r_i of this polynomial is given by:

   i. $r_i = 0-$no decay in the pattern of the ω_i coefficients,

   ii. $r_i = 1-$exponential decay pattern of the ω_i coefficients,

   iii. $r_i = 2-$complex decay pattern of the ω_i coefficients, *e.g.,* alternating positive-negative spikes ($r_i > 2$ is very rare).

(c) The numerator factor $\omega_i(B)$ captures the *unpatterned spikes* (not part of the decay pattern) in the ω_i coefficients' pattern and the *decay start-up* value(s). The order of this polynomial is $m_i = u_i + r_i - 1$, where u_i represents the unpatterned coefficients.

   i. if $r_i > 0$, then u_i denotes the number of non-zero coefficients before the decay starts,

   ii. if $r_i = 0$, then all the non-zero parameters are considered unpatterned.

Identification of the ARIMA model

The order identification is done in this case by analyzing the sample autocorrelation function (ACF) and partial autocorrelation function (PACF) of the estimated coefficients, a popular approach of *Box and Jenkins* (1976).

Given a time series Y_t and the assumption that Y_t is a second order stationary process, the autocorrelation of lag k is defined as $\rho_k = \frac{Cov(Y_t, Y_{t-k})}{Var(Y_t)}$, and thus, the ACF represents the set of values $\{\rho_1, \rho_2, ..., \rho_k\}_{k\in\mathbb{N}}$. It is used in identifying the order of a MA(q) process, more exactly, ρ_k becomes not significantly different from 0 after q lags.

Given a time series Y_t and the assumption that Y_t is a second order stationary process, the partial autocorrelation of lag k is defined as $\gamma_k = \frac{Cov(Y_t, Y_{t-k}|Y_{t-1},...,Y_{t-k-1})}{Var(Y_t|Y_{t-1},...,Y_{t-k-1})}$, and the PACF denotes the set of values $\{\gamma_1, \gamma_2, ..., \gamma_k\}_{k\in\mathbb{N}}$. It is the autocorrelation between Y_t and Y_{t-k} conditional on the set of observation that are between these points, *i.e.*, $Y_{t-1}, ..., Y_{t-k-1}$. It helps identifying the order of an AR(p) process, such that γ_k becomes not significantly different from 0 after p lags.

Thus, we have to compute the sample ACF and PACF and compare them to the theoretical ones. The theoretical ACF and PACF, illustrated in *Box and Jenkins* (1976), are given here in Table 1.3.1.

The identification for the seasonal part of the process is similar with the non-seasonal one, but instead of looking at the lags $1, 2, ..., k$, we look this time at lags that are multiples of our seasonality span s, *i.e.,* $s, 2s, 3s..., ks$.

In Section 4.1.2 we show a step by step identification procedure of both the transfer function and ARIMA model, for our case study on streamflow data.

3.2.2 Model validation

Once the model is defined using the procedure above, one should test its performance and validate it by using a test dataset, different from the one used in the estimation. The performance of the model is computed by comparing the observed data from the test set with the values estimated by the model. Then, the model is validated by comparing it with several other models (*i.e.* simpler models, other category models, benchmark models, etc.).

In order to measure the efficiency we use the Kling-Gupta Efficiency (KGE). This criterion was introduced by *Gupta et al.* (2009) and it represents a decomposition of the Nash-Sutcliffe Efficiency (NSE), introduced by *Nash and Sutcliffe* (1970). Both measures are often used in hydrological modeling, but NSE is reported by many authors to overestimate the model performance during peak flows and underestimate it during low flow conditions.

KGE is defined in terms of three components: the correlation between the observed and estimated series, the global bias of the reconstruction, and the variability. The general formulation for the KGE is

$$KGE = 1 - \sqrt{(\rho - 1)^2 + (\alpha - 1)^2 + (\beta - 1)^2} \tag{1.3.9}$$

where $\rho = \frac{Cov(Y_t^{obs}, Y_t^{est})}{\sigma_{obs}\sigma_{est}}$, $\alpha = \frac{\hat{s}_{est}}{\hat{s}_{obs}}$, $\beta = \frac{\hat{x}_{est}}{\hat{x}_{obs}}$ ($\hat{x}$ and $\hat{s}$ represent the sample mean and standard deviation of a series and *est* and *obs* stand for estimation and observation set).

KGE ranges from $-\infty$ to 1, the closer to 1 the more accurate the model is.

Chapter 4

Case study: Durance watershed

In this chapter we apply the times series reconstruction methodology based on dynamic regression models on streamflow data from the Durance watershed. We consider a group of eight stations in the watershed, each containing daily measurements for 107 years (from 1904 to 2010). In Section 4.1, we are interested in the model identification and parameter estimation, and, thus, we review all the theoretical aspects presented previously but this time on our application study. Then in Section 4.2, we measure the performance of the estimated models, and finally compare it with the performance of a benchmark and hydrological model, respectively.

4.1 Model estimation

We apply a commonly used approach in time series analysis for the model estimation and validation, *i.e.*, we estimate the model on a training set, and then, assess the performance of the model on a test set, different than the training one. Thus, we consider for the training set the longest part of our dataset that has no missing values, namely the last years. Therefore, we use a sequence of 22 years (1980-2001).

4.1.1 Preprocessing steps before modeling

We have seen in the previous chapter that an important step before the model estimation is given by the data preprocessing. In the following, we address some possible pre-modeling issues that might need to be clarified in a dynamic regression modeling.

1. **Explanatory variable selection**

   We have seen in Section 2.2 that there is a very strong-correlated group of stations in the upper Durance, *i.e.*, S1-S4 (referred as Group 1) and one in the middle Durance with stations S6-S8 (referred as Group 2). Also, we have concluded that this correlation might change when different subsets are analyzed, subsets that we refer so far *all seasons* (all available data), *cold season* (only the data associated with autumn-winter), *warm season* (only the data associated with spring-summer).

   As a result, each station (response variable) is considered to have as explanatory variables in the regression all the other stations from the same groups. Particular attention is given to station S5, that has an unclear status. In this case, we look at the correlation (see Table 1.2.4 in Section 2.2.2) and select as explanatory variables only the stations that have a coefficient larger than 0.7. Consequently, S5 is used as an explanatory variable for the stations S4 and S7 in the all seasons and warm season cases, while for the cold season,

Table 1.4.1: Explanatory variables included in the regression part of the DRM (*i.e.*, full models) for each of the eight stations from the Durance watershed

	all data & warm season	**cold season**
S1	S2,S3,S4	S2,S3,S4
S2	S1,S3,S4	S1,S3,S4
S3	S1,S2,S4	S1,S2,S4
S4	S1,S2,S3,S5	S1,S2,S3
S5	S4,S7	S7
S6	S7,S8	S7,S8
S7	S5,S6,S8	S5,S6,S8
S8	S6,S7	S6,S7

Table 1.4.2: Explanatory variables included in the regression part of the DRM after applying the VIF multicolinearity test (*i.e.*, reduced models) for each of the eight stations from the Durance watershed

	all data & warm season	**cold season**
S1	S3	S3
S2	S1,S4	S1,S4
S3	S1,S4	S1,S4
S4	S1,S5	S3
S5	S4,S7	S7
S6	S7,S8	S7,S8
S7	S5,S6,S8	S5,S6,S8
S8	S6,S7	S6,S7

S5 explains only station S7. In return, S5 is modeled only by these stations, *i.e.,* S4 and S7, for all seasons and warm season, and S7 for cold season.

The relationships deduced from the above consideration provide the setting of the covariates in the regression part of the DRM, which we refer so far *full models*. An outline of these models, for each station and type of subset, is given in Table 1.4.1.

2. Multicolinearity

Due to the fact that in our exploratory analysis from Section 2.2 we have encountered very high correlated stations, we now want to examine if we are in the case of multicolinearity (almost perfect linear relationship among explanatory variables). In the presence of multicolinearity the standard errors of the coefficients tend to be large, thus producing wider confidence intervals, among others. More insight in the multicolinearity subject is clearly detailed in *Gujarati and Porter* (2008).

We assess the strength of the multicolinearity by computing the Variance Inflation Factor (VIF). This index measures how much the variance of estimated regression coefficients is increased when compared to having uncorrelated variables; see *Kutner et al.* (2004) and *Gujarati and Porter* (2008) for more details. When multicolinearity is found, the computed VIFs are larger than 5, as suggested in *Eng et al.* (2005), *Montgomery et al.* (2012). In this case, we remove the variable with the highest VIF (among the one with $VIF > 5$) and then reiterate the process until all remaining variables have $VIF \leq 5$.

In our case study, the results show that there is evidence of multicolinearity between the group of stations {S1, S2, S3, S4} from the upper Durance. This leads to a *reduced-form model*, as presented in Table 1.4.2. This reduced-form model is estimated and validated later in the study.

Table 1.4.3: Estimated parameters of the M.NS proxy model, with a transfer function $\omega(B)$ of order $m = 6$ and an AR(1), for station S1

Model parameters							
β_0	$\omega_{1,0}^{S3}$	$\omega_{1,1}^{S3}$	$\omega_{1,2}^{S3}$	$\omega_{1,3}^{S3}$	$\omega_{1,4}^{S3}$	$\omega_{1,5}^{S3}$	ϕ_1
-1.5681	0.7502	0.0417	0.0029	-0.0015	-0.0035	0.0053	0.9772

3. Stationarity

The second-order stationarity evaluation, *i.e.*, the stability of the mean and variance over time, is a very important step, because many modeling approaches start with the assumption of a stationary series. The case of dynamic regression models is no exception. Due to the extreme events and possible climate change, the variance of the Durance streamflow time series is not constant. Therefore, to reduce the effect of the extreme events, we transform the data, such that instead of modeling the raw time series, we use the log-transformed one.

Moreover, we apply two unit root tests, *i.e.*, ADF and KPSS tests, to verify the stability of the series mean. These tests are designed for determining whether differencing is required. For the Durance watershed, after applying these tests for the data from 1980 to 2001 (training dataset) and looking at their resulted p-value, it seems clear that all the stations are stationary at a 5% level. More specifically, all the p-values for the KPSS test (with the null hypothesis $\mathcal{H}_0$: stationarity) are greater than 0.05 and for the ADF test ($\mathcal{H}_0$: not stationary) less than 0.05, suggesting that the Durance streamflow data are stationary in mean.

4.1.2 Model identification and parameter estimation

Model initialization. We consider for the initial or proxy model, six lags $(t, t-1, ..., t-5)$ for each explanatory variable included in the regression according to the reduced-form model, and an AR(1) for the residuals. Moreover, as the models differ according to the type of season, we analyze both cases: i) no-season split, in which case the model is denoted by M.NS, and ii) 2-season split where the model is called M.2S.

Model identification. The errors of the proxy model mentioned above are checked using the Ljung-Box test and, as they are not a white noise process, a new model is needed.

A. Identification of the transfer function

We analyze the patterns of the estimated regression parameters from the proxy model for each explanatory variables (we consider six lags at each explanatory variable, as mentioned earlier). For illustration purposes, we discuss in details just the identification process for model M.NS of station S1. In Table 1.4.3 we show the estimated parameters of this proxy model, *i.e.*, the coefficients of the $\omega(B)$ polynomial. One should recall that the explanatory variable for S1 is station S3, as shown in Table 1.4.2.

For the identification of the transfer function we focus, thus, on the $\omega_{i,j}^{S3}$ coefficients illustrated in Table $\omega(B)$ polynomial. Considering the rules presented in Section 3.2.1, the orders (b_{S3}, m_{S3}, r_{S3}) are identified as follows:

(a) $b_{S3} = 0$, because there is no dead time, *i.e.*, we have no initial null coefficients, as the first coefficient is $\omega_{1,0}^{S3} = 0.7502$.

(b) $r_{S3} = 0$, meaning that there is no decay pattern in the $\omega_{i,j}^{S3}$ coefficients' values. According to the theoretical transfer function, $r_i > 0$ if an exponential, sinusoidal,

etc. pattern is encounter. However, in this case, the values of the $\omega_{i,j}^{S3}$ coefficients decrease, but then slightly increase at the last coefficient $\omega_{1,5}^{S3}$.

(c) $m_{S3} = 0$ or $m_{S3} = 1$; since all coefficients are unpatterned ($r_{S3} = 0$), and recalling that $m_i = u_i + r_i - 1$ with u_i denoting all the non-zero unpatterned coefficients, leads to $m_{S3} = u_{S3} - 1$. But out of the six $\omega_{i,j}^{S3}$ coefficients, we might say that only the first or the first two are significantly different from zero. Thus, this leads to the setting of $u_{S3} = 1$ or $u_{S3} = 2$, and finally to $m_{S3} = 0$ or $m_{S3} = 1$. As the identification process might be rather subjective, we consider several possible models and we validate all of them, letting the performance measure (*i.e.*, KGE) choose the best option.

The general conclusion is that we have no dead time for none of the explanatory variables, so $b_i = 0$ for all stations, and we have no decay pattern as well, so $r_i = 0$, for $i = \{1, ..., l\}$. Given that $r_i = 0$ (no pattern), it means that all the parameters are unpatterned, so $m_i = u_i - 1$, where u_i denotes non-zero unpatterned coefficients. In our case-study the behavior of the unpatterned coefficients is found to be as follows: the first coefficient (lag-0) is highly significant, the second one (lag-1) is close to zero but possibly significant, while the remaining four coefficients (lag-2,3,4,5) are non significantly different from zero. Therefore, we consider worth modeling both options: $u_i = 1/m_i = 0$ and $u_i = 2/m_i = 1$ (0-lag and 1-lag for each explanatory variable) and choose the best one in the validation section. The notation used for these models is, thus, M.NS.0lag or M.NS.1lag for the no-season split models, and M.2S.0lag or M.2S.1lag for the 2-season split ones.

B. Identification of SARIMA

To identify the orders of the SARIMA model, we employ, as indicated in the theoretical part from Section 3.2.1, the sample ACF and PACF plots. We show in Figure 1.4.1 just the plots for stations S1 and S2, because they have a different behavior, the remaining stations being similar to these ones.

For station S1, the proxy model AR(1) is not sufficient, and some correlation is still present in the residuals up to lag-5 in both ACF and PACF plots. This suggest that an ARIMA model of order up to $(5, 0, 5)$ should be applied for this case study. In the case of station S2, beside some correlation at the beginning up to lag-5 or -6, we have also some significant coefficients at lags multiple of 7, indicating a weekly seasonality, thus a SARIMA model.

As general conclusion in the SARIMA identification step, it was found that stations S2 and S3 have a weak weekly seasonality. Analyzing in more detail these time series, it was discovered that the periodicity starts in 1966. We found that in December 1965 a dam was installed upstream of station S2, called Pont-Baldy, and that the water is retained and released every week, causing the weekly seasonality. Therefore, we proposed for station S2 and S3 a $SARIMA(p, d, q)(P, D, Q)_7$, where $d = D = 0$ (due to stationary data), $p, q \leq 5$ and $P, Q \leq 1$, while for the remaining stations we chose an ARIMA(p, d, q), where $d = 0$, $p \leq 5$ and $q \leq 5$.

As several models have been estimated and we want to choose only one, the selection of the (S)ARIMA model is made by looking at the Akaike Information Criterion (AIC, *Akaike* (1974)) and the Bayesian Information Criterion (BIC, *Schwarz* (1978)). Consequently, Table 1.4.4 shows the best models chosen for each station (AIC and BIC yield the same result).

One last point we want to address regarding the SARIMA model identification is whether the models from stations S2 and S3 are multiplicative or additive, considering they also include a seasonal part. For illustration, we consider the SARIMA(1,0,2)(1,0,1)$_7$ model

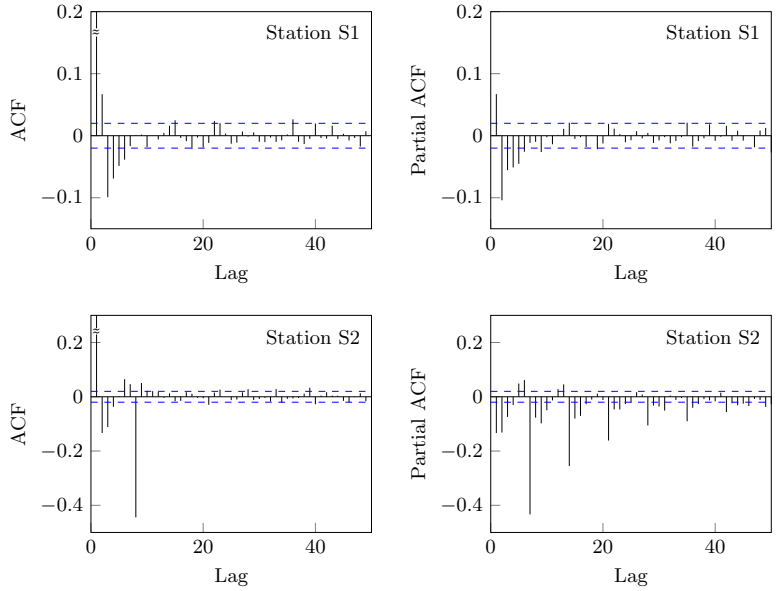

Figure 1.4.1: ACF and PACF plots of the residuals for stations S1 and S2 after using the initial proxy model AR(1)

Table 1.4.4: The selected models for the (S)ARIMA part of the dynamic regression model for the eight station of the Durance watershed

	Data	Best (S)ARIMA model
S1	all data	ARIMA(1,0,4)
	cold season	
	warm season	
S2	all data	SARIMA(1,0,2)(1,0,1)$_7$
	cold season	
	warm season	SARIMA(2,0,2)(1,0,1)$_7$
S3	all data	SARIMA(3,0,1)(1,0,1)$_7$
	cold season	SARIMA(2,0,2)(1,0,1)$_7$
	warm season	
S4-S8	all data	ARIMA(2,0,2)
	cold season	
	warm season	

of the cold season from S2. This model has the following mathematical formulation for the residuals Z_t:

$$\begin{aligned} Z_t =& \phi_1 Z_{t-1} + \phi_{s,1} Z_{t-7} - \phi_1 \phi_{s,1} Z_{t-8} \\ &+ e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \theta_{s,1} e_{t-7} + \theta_1 \theta_{s,1} e_{t-8} + \theta_2 \theta_{s,1} e_{t-9}, \end{aligned} \tag{1.4.1}$$

which is equivalent to an ARIMA(8,0,9) where the AR lags of order 2,3,4,5,6 ($\phi_2^{'}, ..., \phi_6^{'}$) and MA lags of order 3,4,5,6 ($\theta_3^{'}, ..., \theta_6^{'}$) are set to zero. The ARIMA(8,0,9) model formulation is displayed in (1.4.2).

$$\begin{aligned} Z_t =& \phi_1^{'} Z_{t-1} + \phi_7^{'} Z_{t-7} + \phi_8^{'} Z_{t-8} \\ &+ e_t - \theta_1^{'} e_{t-1} - \theta_2 e_{t-2} - \theta_7^{'} e_{t-7} - \theta_8^{'} e_{t-8} - \theta_9^{'} e_{t-9} \end{aligned} \tag{1.4.2}$$

As reported in *Suhartono* (2011), a multiplicative SARIMA assumes that the parameters related to the non-seasonal and seasonal combination (*i.e.,* parameters $\phi_8^{'}, \theta_8^{'}, \theta_9^{'}$ from (1.4.2)) are significant and that they are equal to the multiplication between the parameters of the non-seasonal and seasonal components (*i.e.*,$\phi_8^{'} = -\phi_1 \cdot \phi_{s,1}$,$\theta_8^{'} = -\theta_1 \cdot \theta_{s,1}$,$\theta_9^{'} = -\theta_2 \cdot \theta_{s,1}$ from (1.4.1)). Results show that all the estimated multiplicative parameters (*i.e.,* $\hat{\phi}_8^{'}, \hat{\theta}_8^{'}, \hat{\theta}_9^{'}$) are significant and, moreover, the multiplication of the non-seasonal and seasonal estimated coefficients ($-\hat{\phi}_1 \cdot \hat{\phi}_{s,1}$,$-\hat{\theta}_1 \cdot \hat{\theta}_{s,1}$,$-\hat{\theta}_2 \cdot \hat{\theta}_{s,1}$) prove to be each time inside the 95% confidence interval for the estimated $\phi_8^{'}, \theta_8^{'}, \theta_9^{'}$ (*e.g.,* $\hat{\phi}_8^{'} \pm 1.96 \cdot \hat{\sigma}_8$). We conclude, thus, that a multiplicative SARIMA is suited for both S2 and S3 stations.

Model checking. We analyze the errors of the selected models from the Identification step, and verify if they represent a white noise process (zero mean, finite variance and independent). Results show that all the models present independent errors, *i.e.,* all p-values of the Ljung-Box test are near 0.9, mean close to zero and finite variance. To continue our illustrative example considered earlier, in Figure 1.4.2 we display the errors of the final selected model for stations S1 and S2. It can be seen that the autocorrelation is not present anymore in the residuals, thus we can consider these models in the next step of the modeling procedure, *i.e.*, the validation step.

4.2 Model validation and performance evaluation

To validate and study the performance of our models, we used three different test sets each containing four years of daily flow data, that is: 1918-1921, 1931-1934 and 2002-2005. One should recall that the parameters were estimated by data from the 1980-2001 time span.

Regression problems become difficult when the explanatory variables are not available for the same time span as the response variable, and thus many researchers avoid this approach in the case of missing data imputation. However, we want to show that even in the case of missing explanatory variables, regression models can have stable estimates if calibrated accordingly. Therefore, we take into consideration these two situations:

1. The data for the explanatory variables in the regression are all present (complete-covariates model).

2. The data for the explanatory variables in the regression are partially or totally missing (missing-covariates model).

The performance of the models is then compared with a simpler, but common method of reconstructing missing meteorological data (*Bárdossy and Pegram* (2014), *Hirsch* (1979), *Wallis et al.*

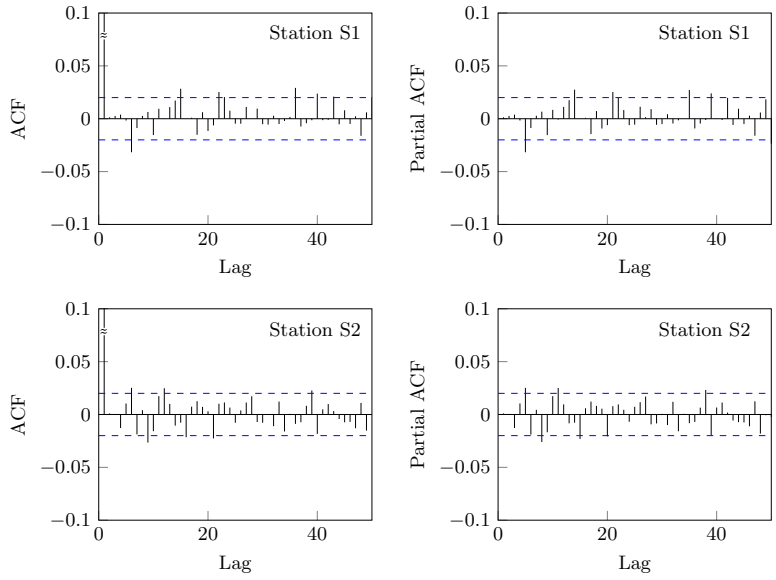

Figure 1.4.2: ACF and PACF plots of the residuals for stations S1 and S2 after using the selected (S)ARIMA models

(1991)), the nearest-neighbors technique (NN). This method allows the infilling of missing data for a station by taking information from neighbor stations (transferred directly or weighted). For this study, we use as neighbors the explanatory variables initially selected for the regression part of our full model with all data (see Table 1.4.1). The missing values of the target station are obtained by weighting the neighbor station(s) with the ratio of daily mean flow of the target station over the daily mean flow of each neighbor.

Beside the NN, we also use also for comparison continuous streamflow time series over the 1904-2010 period, obtained from meteorological data reconstruction (ANATEM method, see *Kuentz* (2013), *Kuentz et al.* (2013))) and rainfall-runoff (RR) modeling. More specifically, this reconstruction approach is based on a combination of large scale climatological variables (atmospheric pressure fields) and regional scale measurements (observed precipitation or air temperature data). Then, the reconstructed climatological time series are transferred in a rainfall-runoff model, allowing the computation of the streamflow simulations.

To demonstrate the reliability of our estimations, we end this section with the performance results of the estimated models on simulated data.

4.2.1 Validation of complete-covariates model

The accuracy of the models is investigated through the KGE criteria described in Section 3.2.2. The results are illustrated in Table 1.4.5.

One important aspect that must be emphasized with respect to the KGE criteria results, is that they are rather consistent in the model choice over the three test sets. That is, although in the parameters estimation step we used only 22 years, the models behave the same according to KGE at the beginning or middle part of the 107 years time span. One exception has to be noticed at station S2, where for the period 2002-2005 the M.2S.0lag model is selected by KGE, while for the other two periods, the M.2S.1lag model resulted to be the best, but the two KGE values are close to each other, i.e., KGE = 0.946 for the M.2S.0lag and KGE = 0.933 for the M.2S.1lag.

Table 1.4.5: KGE results for the validation of the test-period 1918-1921, 1931-1934 and 2002-2005 for the complete-covariates models and the two alternative methods of infilling, NN and ANATEM-RR

	S1	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**
				Period 1918-1921				
M.NS.0lag	0.778	0.763	0.879	0.627	**0.857**	0.685	0.845	0.787
M.2S.0lag	0.863	0.873	0.935	0.764	0.835	0.713	0.844	**0.794**
M.NS.1lag	0.827	0.795	0.886	0.712	0.812	0.744	**0.902**	0.751
M.2S.1lag	**0.866**	**0.890**	**0.941**	**0.811**	0.786	**0.776**	0.893	0.784
NN	***0.926***	***0.904***	*0.656*	*0.652*	*0.702*	*0.724*	*0.815*	*0.522*
ANATEM-RR	*0.751*	*0.627*	*0.871*	*0.593*	*0.770*	*0.730*	*0.561*	*0.220*
				Period 1931-1934				
M.NS.0lag	0.839	0.826	0.761	0.664	**0.722**	0.651	0.749	0.722
M.2S.0lag	0.910	0.914	0.823	0.833	0.635	0.671	0.740	**0.741**
M.NS.1lag	0.888	0.856	0.767	0.733	0.662	0.705	**0.796**	0.691
M.2S.1lag	**0.912**	**0.923**	**0.838**	**0.871**	0.591	**0.729**	0.777	0.734
NN	*0.897*	*0.769*	***0.907***	*0.743*	*0.543*	*0.519*	*0.793*	*0.410*
ANATEM-RR	*0.861*	*0.831*	*0.707*	*0.731*	*0.516*	***0.773***	*0.495*	*0.261*
				Period 2002-2005				
M.NS.0lag	0.718	0.898	0.853	0.769	**0.780**	0.636	0.809	0.850
M.2S.0lag	0.804	**0.946**	0.919	0.905	0.724	0.673	0.810	**0.890**
M.NS.1lag	0.769	0.930	0.858	0.859	0.709	0.694	**0.870**	0.809
M.2S.1lag	**0.812**	0.933	**0.931**	**0.936**	0.672	**0.743**	0.867	0.876
NN	*0.774*	*0.694*	*0.885*	*0.839*	*0.587*	*0.591*	*0.769*	*0.805*
ANATEM-RR	***0.823***	*0.873*	*0.880*	*0.889*	*0.755*	***0.829***	*0.804*	*0.706*

[1] **Legend:**
bold-red = best model out of the four estimated models;
bold-square = cases when NN or ANATEM-RR performs better compared to our best-case model for each station;

[2] **Notation:**
M.NS.0lag / M.NS.1lag = no-season split model with 0-/1-lag for each explanatory variable
M.2S.0lag / M.2S.1lag = 2-season split model with 0-/1-lag for each explanatory variable

First, the results show that six stations, *i.e.,* all stations from upper Durance (S1-S4) and two stations from middle Durance (S6, S7), are better fitted by a model that includes 1-lag for the explanatory variables, while only two stations from middle Durance (S5, S8) work better with models that have 0-lag terms. The results are coherent when looking at the hydrological regimes and the characteristics of the stations. These stations (S1-S4, S6, S7) are situated at a high altitude and have mainly a regime influenced by snowmelt and low temperature, so it is probable that some delay may appear for the flow. Considering station S5 and S8, the absence of lags is due to the fact that the watershed has a small drainage area (S5) or is characterized by a lowland basin (S8).

Second, for some stations (S1-S4, S6, S8) the models with 2-season split are selected, while for the others (S5, S7) the ones with no-season split. There is no clear hydrological explanation for this behavior. It is to be noted that we only look at some characteristics, like hydrological regimes, watershed surface and altitude, however, we can have other influential factors that may drive these two stations.

In Table 1.4.6, one can find a summary of the selected models for each station, along with their estimated parameters.

The superiority of our approach is emphasized when comparing the KGE results of the proposed models with the ones from the two approaches mentioned earlier, NN and ANATEM-RR. The results reveal that, except for some isolated cases (three for NN and three for ANATEM-RR), our approach performs better. An important aspect that must be highlighted is that with DRM the efficiency of the models measure with KGE is never lower than 0.72, while NN and ANATEM-RR, due to lack of robustness, reduce up to a level of 0.41 and 0.22, respectively, as seen in Table 1.4.5.

An illustration of the reconstructed series (considering the best DRM for each station) versus the observed one for the period 2002-2005 is shown in Figure 1.4.3, along with the 95% confidence intervals. One can notice that the reconstructions of stations S1, S4 and S5 do not reproduce entirely the peak flows, but the recessions are good. Stations S2 and S3 catch very well the peaks, but the weekly fluctuations (stronger at S2, see zoomed areas (a) and (c) in Figure 1.4.3) decrease in estimation performance for the long term reconstructions (see zoomed sectors (a) and (b)). Regarding the other stations, S6 and S7 have mainly well-modeled reconstructions, while station S8 shows some overestimated peaks. These aspects should be further studied and addressed in a future research. Same conclusion can be drawn for the other two periods.

4.2.2 Validation of missing-covariates model

We discussed in Section 4.2.1, that there are cases when the complete-covariates model from the previous section cannot be applied as the data for the explanatory variables are missing. The purpose of this section is to test how the proposed models behave in this case. Therefore, in order to be able to apply the estimated (complete-covariates) models, we have to "temporary replace" the missing values of the explanatory variables with some "temporary estimates". Thus, we use for these "temporary estimates" the weighted values from the correlated-neighbor stations (*i.e.*, same procedure as in the case of NN estimation, presented at the beginning of Section 4.2). When all the covariates are missing, so we have no correlated-neighbor stations, we use the daily mean (mean of the non-missing values for a certain day for that station).

In order to validate this procedure we take the best models selected for the complete-covariates model study (see Table 1.4.6). Then, for each station at a time, we overlay on each test set (*i.e.*, 1918-1921, 1931-1934, 2002-2005) the pattern of missing values from two periods, 1904-1907 (denoted Scenario 1) and 1951-1954 (denoted Scenario 2). The advantage of this procedure is that we created two scenarios with missing input-variables, but we have also the observations in order to test the accuracy. For having the best possible output, we proceed first with the stations with the most complete set of explanatory variables, finishing

Table 1.4.6: Summary of the selected models for each of the eight stations from the Durance watershed

	Selected models[(1)]		**Model parameters**										
S1	**M.2S.1lag**	**Cold**	β_0	$\omega_{1,0}^{S3}$	$\omega_{1,1}^{S3}$	ϕ_1	θ_1	θ_2	θ_3	θ_4			
			-1.37	0.69	0.05	0.99	0.02	-0.10	-0.06	-0.03			
		Warm	β_0	$\omega_{1,0}^{S3}$	$\omega_{1,1}^{S3}$	ϕ_1	θ_1	θ_2	θ_3	θ_4			
			-1.32	0.78	-0.03	0.97	0.15	-0.06	-0.07	-0.04			
S2	**M.2S.1lag**	**Cold**	β_0	$\omega_{1,0}^{S1}$	$\omega_{1,1}^{S1}$	$\omega_{2,0}^{S4}$	$\omega_{2,1}^{S4}$	ϕ_1	θ_1	θ_2	$\phi_{s,1}$	$\theta_{s,1}$	
			1.29	0.47	-0.03	0.16	0.06	0.89	-0.25	-0.16	0.94	-0.48	
		Warm	β_0	$\omega_{1,0}^{S1}$	$\omega_{1,1}^{S1}$	$\omega_{2,0}^{S4}$	$\omega_{2,1}^{S4}$	ϕ_1	ϕ_2	θ_1	θ_2	$\phi_{s,1}$	$\theta_{s,1}$
			1.08	0.68	0.06	0.12	0.02	0.10	0.62	0.59	-0.17	0.83	-0.36
S3	**M.2S.1lag**	**Cold**	β_0	$\omega_{1,0}^{S1}$	$\omega_{1,1}^{S1}$	$\omega_{2,0}^{S4}$	$\omega_{2,1}^{S4}$	ϕ_1	ϕ_2	θ_1	θ_2	$\phi_{s,1}$	$\theta_{s,1}$
			2.18	0.51	-0.03	0.35	0.04	1.45	-0.46	-0.56	-0.11	0.87	-0.81
		Warm	β_0	$\omega_{1,0}^{S1}$	$\omega_{1,1}^{S1}$	$\omega_{2,0}^{S4}$	$\omega_{2,1}^{S4}$	ϕ_1	ϕ_2	θ_1	θ_2	$\phi_{s,1}$	$\theta_{s,1}$
			2.16	0.54	0.07	0.32	0.07	0.47	-0.49	-0.51	-0.10	0.90	-0.87
S4	**M.2S.1lag**	**Cold**	β_0	$\omega_{1,0}^{S3}$	$\omega_{1,1}^{S3}$	ϕ_1	ϕ_2	θ_1	θ_2				
			-1.61	0.98	-0.03	1.50	-0.52	-0.58	-0.14				
		Warm	β_0	$\omega_{1,0}^{S1}$	$\omega_{1,1}^{S1}$	$\omega_{2,0}^{S5}$	$\omega_{2,1}^{S5}$	ϕ_1	ϕ_2	θ_1	θ_2		
			0.58	0.56	0.14	0.27	0.10	1.65	-0.66	-0.61	-0.16		
S5	**M.NS.0lag**		β_0	$\omega_{1,0}^{S4}$	$\omega_{2,0}^{S7}$	ϕ_1	ϕ_2	θ_1	θ_2				
			-0.55	0.62	0.56	1.53	-0.54	-0.73	-0.04				
S6	**M.2S.1lag**	**Cold**	β_0	$\omega_{1,0}^{S7}$	$\omega_{1,1}^{S7}$	$\omega_{2,0}^{S8}$	$\omega_{2,1}^{S8}$	ϕ_1	ϕ_2	θ_1	θ_2		
			1.04	0.51	0.03	0.51	0.12	1.37	-0.40	-0.45	-0.15		
		Warm	β_0	$\omega_{1,0}^{S7}$	$\omega_{1,1}^{S7}$	$\omega_{2,0}^{S8}$	$\omega_{2,1}^{S8}$	ϕ_1	ϕ_2	θ_1	θ_2		
			1.07	0.58	0.07	0.39	0.07	1.62	-0.63	-0.59	-0.14		
S7	**M.NS.1lag**		β_0	$\omega_{1,0}^{S5}$	$\omega_{1,1}^{S5}$	$\omega_{2,0}^{S6}$	$\omega_{2,1}^{S6}$	$\omega_{3,0}^{S8}$	$\omega_{3,1}^{S8}$	ϕ_1	ϕ_2	θ_1	θ_2
			-0.17	0.29	0.03	0.10	-0.01	0.31	0.07	1.37	-0.37	-0.53	-0.11
S8	**M.2S.0lag**	**Cold**	β_0	$\omega_{1,0}^{S6}$	$\omega_{2,0}^{S7}$	ϕ_1	ϕ_2	θ_1	θ_2				
			-0.20	0.37	0.74	1.49	-0.51	-0.73	-0.01				
		Warm	β_0	$\omega_{1,0}^{S6}$	$\omega_{2,0}^{S7}$	ϕ_1	ϕ_2	θ_1	θ_2				
			-0.29	0.31	0.80	1.55	-0.57	-0.66	-0.10				

[(1)] Corresponding mathematical formulations of the selected model (illustration of S1, similar for the other stations):

$$\text{S1:}\begin{cases} Y_t^{S1} = \beta_0 + \omega_{1,0}^{S3} X_t^{S3} + \omega_{1,1}^{S3} X_{t-1}^{S3} + Z_t \\ Z_t = \phi_1 Z_{t-1} + e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \theta_3 e_{t-3} - \theta_4 e_{t-4} \end{cases}$$

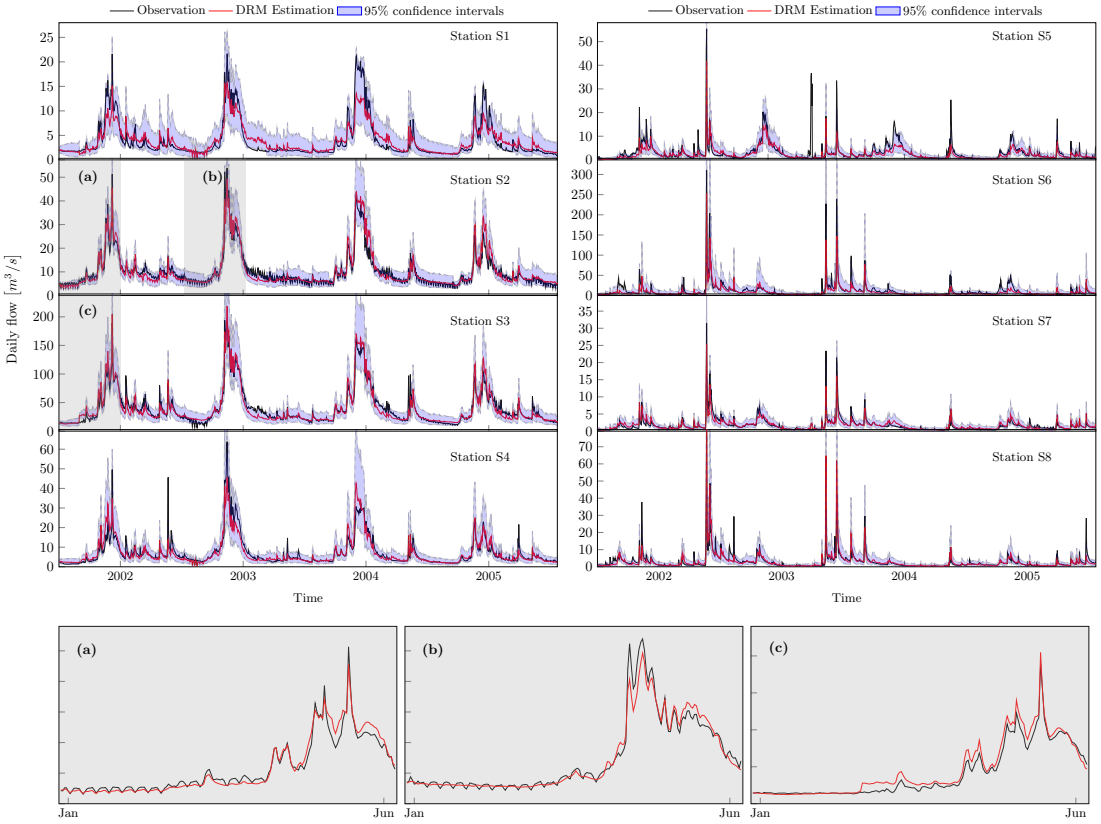

Figure 1.4.3: Daily streamflow estimations versus observations for period 2002-2005 in the case of a complete-covariate model, for the eight stations in the Durance watershed. We show also three zoomed areas (a), (b) and (c).

the reconstruction with the station with the least complete model.

As it was seen in Section 4.2.1, the proposed technique of reconstructing streamflow data yields very good results when all input variables are available (complete-covariates model), surpassing the performance of more complex models like ANATEM-RR. However, the KGE criteria results for the missing-covariates models in Table 1.4.7 show that by replacing the missing values in the input-variables with the weighted values from the correlated neighbors or with the daily mean, we slightly decrease in performance, but, overall, the KGE is still above 0.5.

To better understand the results, we take one case and examine it in more details, *i.e*, we focus on the estimation period 2002-2005, Scenario 2 (*i.e.*, results from the last two rows in Table 1.4.7). According to our reconstruction procedure for the missing-covariates models, we start with the stations with the most complete set of explanatory variables, so in this case we have the order {S4, S2, S3, S1, S5, S7, S6, S8}. We give below a short explanation for each station:

- for stations S2 and S3 (which have as explanatory variables stations S1 and S4), except the only one point where S1 is missing, the models are estimated with a complete-covariates model. This explains why the KGE results are the same in the complete- and missing-covariates models.

- station S5 (which has as explanatory variables stations S4 and S7) is not affected in

Table 1.4.7: KGE results for the validation of the test-period 1918-1921, 1931-1934 and 2002-2005 for the missing-covariates models (results shown only for the best-case complete-covariates models). The number of missing values for each scenario and each station is indicated in the lines $\#NAs$.

	S1	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**
				Period 1918-1921				
Scenario 1	0.866	0.890	0.941	0.811	0.857	0.679	0.753	0.674
#NAs	*1461*	*0*	*0*	*0*	*0*	*731*	*1311*	*1461*
Scenario 2	0.866	0.890	0.941	0.811	0.726	0.702	0.799	0.688
#NAs	*1*	*731*	*1461*	*0*	*730*	*1096*	*1461*	*517*
				Period 1931-1934				
Scenario 1	0.912	0.923	0.838	0.871	0.722	0.654	0.553	0.542
#NAs	*1461*	*0*	*0*	*0*	*0*	*731*	*1311*	*1461*
Scenario 2	0.912	0.923	0.838	0.871	0.688	0.595	0.650	0.570
#NAs	*1*	*731*	*1461*	*0*	*730*	*1096*	*1461*	*517*
				Period 2002-2005				
Scenario 1	0.812	0.933	0.931	0.936	0.780	0.670	0.715	0.673
#NAs	*1461*	*0*	*0*	*0*	*0*	*731*	*1311*	*1461*
Scenario 2	0.811	0.933	0.930	0.936	0.780	0.696	0.717	0.788
#NAs	*1*	*731*	*1461*	*0*	*730*	*1096*	*1461*	*517*

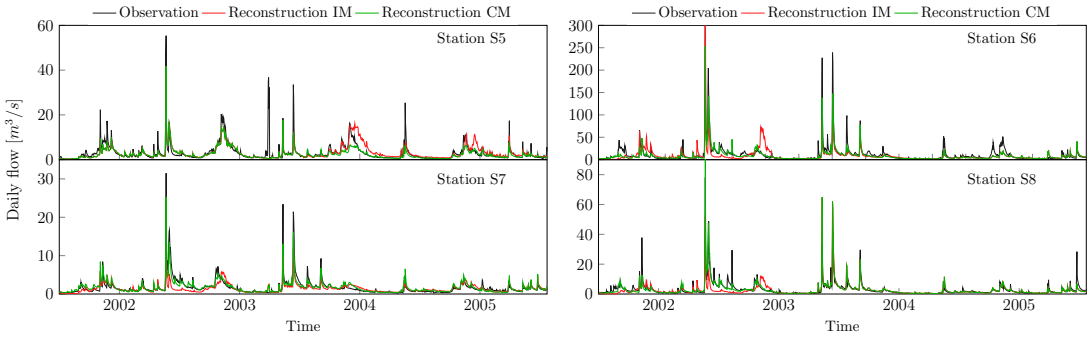

Figure 1.4.4: Daily streamflow estimations versus observations for period 2002-2005 in the case of a missing-covariate model, when Scenario 2 (overlay period 1951-1954) is taken into consideration

performance by the fact that its second explanatory variable (*i.e.*, S7) was temporary replaced by the weighted NN.

- stations S6, S7, S8 have explanatory variables that either were reconstructed earlier above, or that are missing. This explains the decrease in KGE criteria compared to the complete-covariates model.

In Figure 1.4.4 we illustrate the daily streamflow observations versus estimations for stations S5 to S8 when considering the model with missing-covariates. For a more clear overview of the different behavior of the complete- and missing-covariates models, we include also in these plots the estimates from the complete-covariates model. We do not present the plots for S1 to S4 because in Scenario 2 they are estimated with the complete-covariates model, so we already have these outputs in Figure 1.4.3.

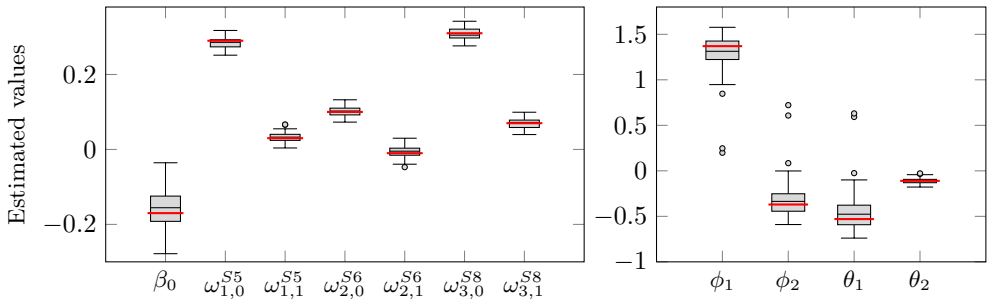

Figure 1.4.5: Boxplots of the estimated parameters on simulated data for station S7, based on 50 replicates. The true values of each parameter are represented by horizontal red lines.

4.2.3 Validation on simulated data

In Section 4.2.1 and Section 4.2.2, we validate our DRMs using a deterministic procedure, thus providing a unique KGE value for each model and station. However, as the used infilling models are stochastic by nature, a single run of the model might not provide enough information about the KGE. Therefore, it is recommended to run the model several times and treat the KGE as a random variable.

In this case, we simulate daily streamflow data for the eight stations for the period 2002-2005. For each station, we start by randomly generating $nsim$ ($nsim = 50$) different white noise sequences for the error terms (e_t in (1.3.1b)) and used them, along with the already estimated (S)ARIMA parameters (see Section 4.1), to create $nsim$ residuals series (Z_t in (1.3.1b)). Then, using the explanatory variables (from the observed daily streamflow series, period 2002-2005) and the previously estimated regression parameters, we performed $nsim$ daily streamflow simulations (denoted sim_i, $i = 1, ..., nsim$).

Afterwards, considering sim_i, we estimate the parameters of each simulated series following the same modeling procedure as in the case of observed streamflow. Then, we compute the fitted values of each series (denoted fit_i, $i = 1, ..., nsim$). The performance of each estimation is computed with KGE, between fit_i and sim_i data. We discuss and illustrate just the results for the model M.NS.1lag, the other models having similar interpretations.

In Figure 1.4.5, we display the boxplots of the estimated parameters over the $nsim = 50$ simulations for station S7 (similar behavior is noticed for the other stations), and we show also the true value of these parameters. The general conclusion that can be drawn is that the parameters of the ARIMA model have a higher variability compared to the ones from the regression model. However, both ARIMA and regression models seem to provide good parameter estimates over the 50 simulated series.

The validation methodology for the simulated data is similar to the one used in Section 4.2.1 and Section 4.2.2 when we worked with observed daily streamflow data. On the top plot of Figure 1.4.6, the validation of the complete-covariates model on simulated data shows that we have a very good performance in the upper Durance (average KGE above 0.99 and variability smaller than 0.002) and a slightly smaller one in middle Durance (average KGE above 0.96 and variability smaller than 0.003), behavior that, in fact, reinforces the statements from the validation on observed data. On the bottom plot of Figure 1.4.6, the validation for the missing-covariates model is illustrated for the two scenarios mentioned in Section 4.2.2. Attention must be paid when analyzing these plots because the two scales are very different. Once again, it is shown that we decrease in performance when we replace the missing input variables with the weighted values of the correlated neighbors, but the average KGE remains, mainly, above 0.5.

Finally, the reconstructed series for the eight stations are illustrated in Figure 1.4.7. The

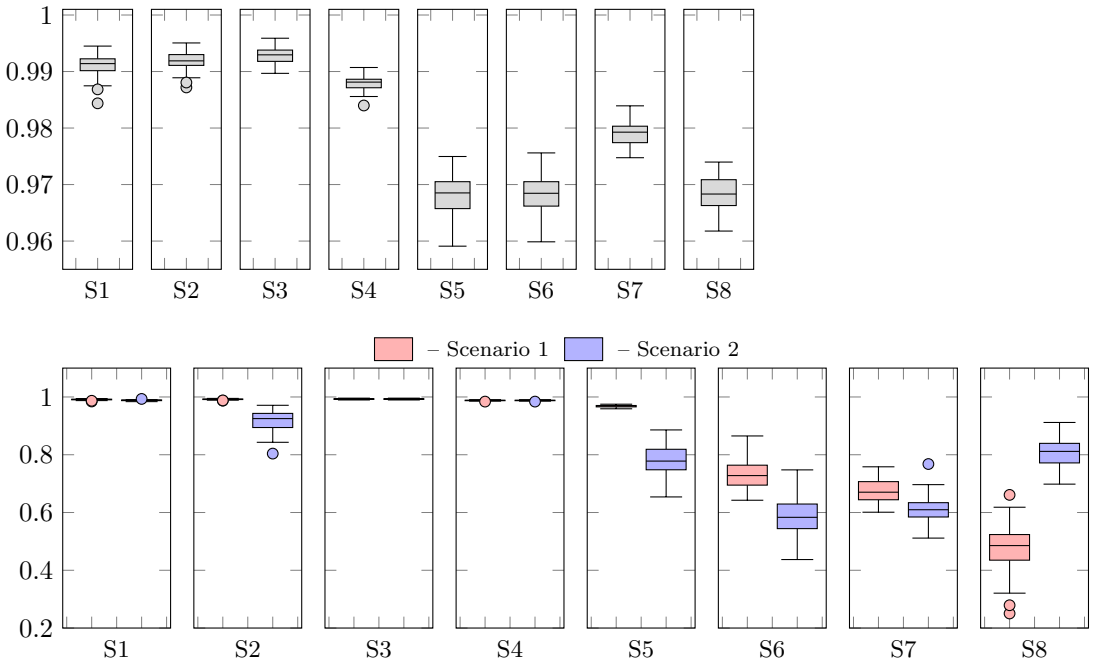

Figure 1.4.6: Boxplots of the KGE for the 50 simulations: top-plot illustrates the results for the complete-covariates model; bottom-plot illustrates the results for the missing-covariates model for both scenarios (Scenario 1=overlay periods 1904-1907 and Scenario 2=overlay period 1951-1954)

Table 1.4.8: Summary regarding the use of complete-covariates model in the Durance daily streamflow reconstructions

		S1	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**
Total NAs	#	9217	4900	5903	1207	3340	5473	9711	7067
Reconstructions	#	9188	2937	5604	378	904	973	758	1345
using C-CM	%	99.69	59.94	94.93	31.32	27.07	17.78	7.81	19.03

reconstructions show once more that in case of an infilling using the complete-covariates model (all covariates are present) the estimations are very good (see stations S1, S2, S3) and they slightly decrease in performance when we deal with missing explanatory variables in the model, see the case of the stations from the middle Durance (S5, S6, S7, S8).

To have a more clear overview, we provide in Table 1.4.8, for each station, a summary regarding the total number of missing values (NAs) and how many reconstructions (values) were estimated using the complete-covariates model (that we will denote here C-CM). We can see that while station S1 was almost 100% reconstructed with a complete-covariates model, station S7 used for more than 90% of the cases the missing-covariates model.

Remark 1.4.1. The computations were performed with the R Software, using the packages: *stats* (general-main computations), *iki.dataclim* (homogeneity tests), *cluster* (PAM exploratory analysis), *tseries* (stationarity analysis) and *forecast* (DRM fit and prediction).

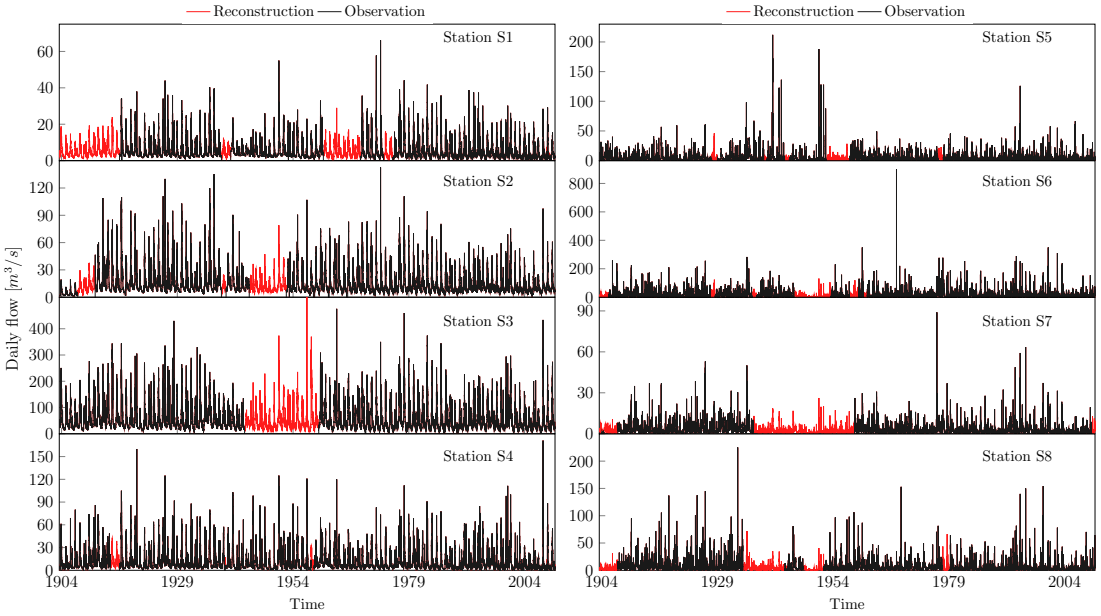

Figure 1.4.7: Daily streamflow reconstructed series of the eight stations of the Durance watershed

Chapter 5

Discussions, conclusions and perspectives

5.1 Discussions and conclusions

Complete records of flow data are very important and critical to a sustainable management of water resources. During the past decades researchers have developed techniques to reconstruct these series using a variety of methods such as linear and nonlinear models, parametric and non-parametric approaches, etc.

In this study we reconstruct daily streamflow data by using dynamic regression models. This method uses the linear relationship between the correlated stations at different lags and it adjusts the residuals by fitting a (S)ARIMA model. Unlike previous studies that address either the inclusion of multiple explanatory variables or the modeling of the residuals from a simple linear regression, the use of DRMs allows to take into account both aspects, and thus improves the performance of the model. Moreover, by applying this technique we managed to reconstruct the data without making use of additional variables, such as precipitation, like other models require.

We employ this approach for reconstructing data from the Durance watershed, one of the most important river in southern France. This watershed is characterized by a variety of hydro-climatic processes and it is defined by its many water-related uses, especially for EDF, the administrator of the Durance-Verdon hydroelectric facilities. The Durance watershed offers many purposes like hydropower generation, irrigations, water supply for cities like Marseille and Aix-en-Provence, streamflow regulation, or tourism near the lakes.

Our case study involves the eight hydrometric stations of the Durance watershed that has the longest data availability (1904-2010). The conclusion shows that dynamic regression models outperform two other modeling approaches, nearest neighbor technique and a more complex meteorological model (ANATEM-RR). When measuring the accuracy of the estimates, it has been proven that the choice of the model is highly dependent on the station's characteristic and hydrological regimes and no generalization can be made for all stations. In other words, we have that for all the stations from upper Durance, the dynamic regression model with 1-lag explanatory variables and 2-season performs best, while for the middle Durance, as the stations have different hydrological regimes, we have selected models with or without past lags included and with or without seasonal split.

An extended model validation study has showed that, even if there are missing explanatory variables (partially or completely) in the regression, the selected models can perform well. More specifically, in this case the missing parts of each explanatory variable are replaced with the weighted measurements from the correlated neighbor stations or with the daily mean (mean of

the non-missing values for a day for that station) when all explanatory variable are missing.

In conclusion, we introduced in this study a method for reconstructing hydrological data that is very general, flexible and requires only streamflow data. Moreover, this model can reconstruct data with missing intervals of various length (from several to thousands of observations) in a short run-time period. Apart from this, our study has been conducted on a large watershed characterized by several hydrological regimes and various data quality issues, providing thus a complex and reliable imputation technique.

5.2 Perspectives

While we managed to attain an accurate reconstruction of the Durance daily streamflow data, a lot of research and development is still required in this area. One important aspect that requires improvement is the case when the explanatory variables from the regression are missing. We have seen in the case study that a temporary replacement of these missing values with the weighted nearest neighbors observations or with the daily mean, produce less variability in those parts of the time series. This issue has a great impact in the analysis of extreme events. Therefore, a more robust method should be used in this case, so that the variability of the time series is preserved.

Also, a supplementary analysis of this work, could comprise a comparative study of different streamflow imputation methods. Since we have taken into consideration just two benchmark models, the DRM's efficiency in reconstructing streamflow data could be emphasized more clearly by examining other imputation models reported in the literatures. Moreover, the comparison should be based not only on the overall model performance (*e.g.*, KGE measure), but also on particular aspects such as, the performance of the models in the case of different lengths of missing interval, or the computational time required for the calibration and the reconstruction.

In addition to this, based on the results obtained for the Durance watershed, it is interesting to consider reconstructing the data from other types of watersheds, *e.g.*, from other regions with new hydrological regimes. Also, the flexibility offered by the dynamic regression models makes them a valid candidate in applications related to the reconstruction of other variables, such as precipitation or temperature.

Finally, this reconstruction technique has been applied to a dataset composed of eight hydrometric stations. However, for larger datasets, such as tens or even hundreds of stations, it becomes impractical to individually calibrate each station's DRM. Therefore, it would be interesting to define some standard parameterization on subsets of homogenous time series.

Part II

Flexible semiparametric approaches to model the full-range of precipitation amounts

Chapter 1

Introduction

Modeling heavy-tailed distributions is a difficult task and very often the values in the bulk of the distribution are disregarded in order to better fit the tail. Improvements have been made in this sense and, recently, a new class of distributions suited for rainfall data called extended generalized Pareto distribution (EGPD) was introduced by Naveau et al. (2016). This approach is very appealing as it models the entire range of data and does not require the selection of a threshold, unlike many extreme mixture models. Our aim in Part II of this manuscript is to add more flexibility to the EGDP class in order to improve their behavior for the bulk, while maintaining the EVT behavior of the tails. We start by giving a background on heavy-tailed distributions and precipitation modeling in Section 1.1 and Section 1.2, respectively. Section 1.3 is dedicated to a brief overview of Extreme Value Theory and to the theoretical aspects regarding the EGPD class.

1.1 Heavy-tailed distributions

Heavy-tailed distributions have been used to model phenomena in various fields, such as insurance, finance, environmental sciences, computer and communications networks, among others. Several application examples can be found in *Embrechts et al.* (1997), *Coles* (2001) and *Castillo et al.* (2005).

The traditional theory is more focused on the body or bulk of a distribution, the tails being often disregarded. Ignoring these observations can only lead to a biased estimation. Thus, special methods are required in order to accurately analyze the heavy-tailed distributions. The study of these distributions is different from a classical distribution analysis, mainly due to some special characteristics like sparse observations in the tail, or a decay to zero slower than an exponential.

Generally, two main methods are reported in the literature for modeling extreme tail behavior, *i.e.*, the probability of exceeding a high value: i) *block maxima* (maximum values selected from sequences of data), and ii) *peak over threshold* (values larger than a threshold). The block maxima are modeled by a generalized extreme value distribution (GEV), while the peak over threshold by a generalized Pareto distribution (GPD). More details regarding these distributions are provided in Section 1.3.1.

However, sometimes it is necessary to evaluate the heavy-tailed distributions as a whole, so we are interested not just in the tails or the bulk of the distribution, but on both. Experience of modeling shows that models that estimate well the bulk, do not provide a good fit for the tails, and *vice-versa*. Moreover, the estimation becomes even more complicated if the distributions are multimodal.

Several studies reported in the literature deal with the issue of modeling the entire range of heavy-tailed data. In the following, we offer a review of the methods by classifying them in two main groups: mixture models and transformed-retransformed models.

1. **Mixture models**

   This type of models have been intensively studied lately and they are frequently referred to as *extreme value mixture*s. The main idea behind these models is to join the distribution that fits well the bulk (parametric, semiparametric or nonparametric) with the one that fits well the tail (usually a GPD).

   The work presented in *Frigessi et al.* (2002) suggests a dynamic weighted model by mixing a Weibull distribution for the bulk with a GPD for the tail. The threshold, *i.e.*, the joint point of the two distributions, is defined dynamically by means of a weight function. This model has the advantages of no threshold selection, and offers a smooth transition at the junction point of the Weibull and the GPD. However, the inference of the model is not that straightforward and problems may appear especially for the estimation of the GPD shape parameter.

   In *Behrens et al.* (2004) another mixture model that does not require threshold selection is introduced. More specifically, the threshold is considered a parameter that has to be estimated. The bulk of the mixture could be any distribution, such as Weibull, gamma, normal, etc., and the tail is a GPD. Unfortunately, this model has a discontinuity point at the threshold. Similar approaches can also be found in *Mendes and Lopes* (2004), *Zhao et al.* (2010).

   The problem of density discontinuity was solved later in the work proposed by *Carreau and Bengio* (2008). This mixture model, called hybrid Pareto, stitches a normal distribution with a GPD, while enforcing the continuity of the density and of its first derivative. More exactly, the threshold is computed implicitly as a function of the mixture parameters. Nevertheless, the hybrid Pareto model did not perform satisfactory in practice. This aspect was addressed by an improved version presented in *Carreau and Bengio* (2009), where a mixture with hybrid Pareto components was proposed. This new approach is a conditional density estimator and it can handle all types of distributions: asymmetric, heavy-tailed or multimodal.

   We have pointed earlier that often the distribution under study is multimodal, so a single distribution, such as normal, gamma, etc, is not enough to model the body of the distribution. Studies like *do Nascimento et al.* (2011), *Lee et al.* (2012) take into consideration a k-component mixture model for the main body of the distribution. Problems might appear in this case in the inference, as the number of parameters is increasing.

   The need to effectively describe a complex-multimodal data structure, but at the same time avoid the estimation of a high number of parameters, resulted in the appearance of nonparametric density estimators. Unlike parametric models, nonparametric estimators do not make any prior assumption about the distribution under study and take into consideration just the information from the data.

   Among the most known nonparametric estimators (*e.g.*, histogram, projection, kernel estimator, as pointed in *Markovich* (2007)), the most popular is kernel because it provides attractive statistical properties, *i.e.*,. continuity, or the fact that it is a proper density function. Kernel density estimator can be viewed as an extreme end or special case of finite mixture models, in which the n components in the mixture have equal proportion $1/n$, where n is the size of the data. Unfortunately, even though this model adds more flexibility in the estimation, issues might appear at the estimation of heavy-tailed distributions because a unique smoothing parameter is not sufficient to also catch the sparse observations from the tails.

There are studies reported in the literature, *e.g., Devroye and Györfi* (1985), *Silverman* (1986), regarding an improved kernel estimator for long tail distributions, namely the variable bandwidth kernel estimator, *i.e.*, it includes multiple degrees of smoothing. Nevertheless, even though different bandwidths are considered for the body and the tails, the fact that the kernel function has a bounded support will provide a more biased estimation in the unbounded tail domain. Another interesting approach is provided by *MacDonald* (2011), *MacDonald et al.* (2011a,b), where the advantages of a parametric tail model and the flexibility of nonparametric estimators for the body are combined. The proposed mixture model is based on combinations between kernel density estimator and GPD (for one or both tails). This model is attractive, flexible, and has less parameters, but it has an increased boundary bias and it is complex from a computational point of view.

More recently, researchers focus on a similar approach to the variable kernel estimator, but based on sparse linear combinations of density basis functions. The general idea of this approach is that, in a high dimensional mixture model (high number of components), it is believed that this representation is sparse, *i.e.*, just few mixture proportions are different from zero. There are not many studies on this topic, however we want to mention the work of *Bunea et al.* (2010), where the authors consider a sparse mixture of normal densities with pre-set location and scale parameters. Under some identifiability conditions, this model becomes very flexible and requires only the estimation of the mixture proportions. However, this study does not contain any information about the behavior of the estimator in the case of heavy-tailed data. We provide more details about the topic of sparse mixture models and how the sparsity is recovered in Chapter 3.

2. **Transformed-retransformed models**

   Transformed-retransformed models provide a different way of modeling heavy-tailed distributions. This approach consists of, firstly, transforming the data via a continuous, monotone and increasing function and, then, the transformed data are approximated by a nonparametric estimator. The idea of this methodology is to provide different smoothing for different parts of the distribution, but at the same time to keep a bounded support for the nonparametric estimator.

   Therefore, besides the choice of the nonparametric estimator, a very important point is the choice of the transformation function. As mentioned in *Markovich* (2007), there are two main groups of transformation functions:

   (a) fixed transformations

   This type of transformations do not require any information about the distribution under study. A very simple and commonly used transformation function is the logarithm. Other examples include the use of a trigonometric function, such as the one displayed in (2.1.1) and used in the work of *Markovich* (2007), or a parametric one as shown in (2.1.2) and applied by *Markovitch and Krieger* (2000), *Wand et al.* (1991), *Yang and Marron* (1999).

$$T(x) = \frac{2}{\pi} \arctan(x) \tag{2.1.1}$$

$$T(x) = \begin{cases} x^{\lambda}\mathrm{sgn}(\lambda), & \lambda \neq 0 \\ \ln(x) & \lambda = 0 \end{cases} \tag{2.1.2}$$

   Here λ denotes the parameter of the function and $\mathrm{sgn}(\cdot)$ represents the sign function, *i.e.*, it extracts the sign of a real number.

   Generally, the heavy tail of the distribution cannot be accurately estimated if no prior assumption is made about it during the transformation.

(b) adaptive transformations

A more flexible estimation is given by the adaptive transformations. If any information about the behavior of the distribution is known, this knowledge should be included in the transformation process, and thus, further improve the nonparametric estimation and the overall performance. There are several studies, with application in finance, that take a cumulative distribution function (cdf) as an adaptive transformation, *i.e.*, *Buch-Larsen et al.* (2005), *Buch-Kromann et al.* (2007), *Bolancé et al.* (2008) or *Alemany et al.* (2013). These studies mainly consider transformations based on Champernowne (see *Champernowne* (1953)) or log-normal distributions.

1.2 Rainfall modeling

The stochastic modeling process of rainfall data has two components: a discrete distribution of rainfall occurrences (sequence of dry and wet days) and a continuous one of rainfall amounts (amount of water on each wet day). We are interested for this work just in the statistical modeling of wet days, that is the nonzero rainfall data.

Since rainfall amounts are heavily skewed to the right, distributions like Weibull (*Zucchini and Adamson* (1984)), exponential or mixture of exponentials (*Garavaglia et al.* (2010), *Richardson* (1981), *Wilks* (1999), *Woolhiser et al.* (1979)), lognormal (*Apipattanavis et al.* (2007)), gamma (*Katz* (1977), *Stern and Coe* (1984), *Wilks* (1989, 1999)), or a power transformation to normality (*Katz and Parlange* (1995)) have been used to model precipitation. According to *Vrac et al.* (2007) and *Wilks* (2011) the most common choice, though, is a gamma distribution. However, as pointed out by *Katz et al.* (2002), the tail of a gamma distribution can be too light for rainfall amounts and underestimation can occur.

To solve this issue and to improve the fit of the right tail, many researchers become interested only in the largest rainfall amounts and thus, they use the popular framework of Extreme Value Theory (EVT), more exactly the GPD to model the tail. Numerous promising studies can be found in this area, for example the work of *Katz et al.* (2002), *Nadarajah* (2005), *Cooley et al.* (2007). One immediate drawback of this approach is the need of a threshold selection (the limit of large rainfall) which is still a delicate task in the field of EVT. These models take into consideration only extreme precipitation and provide no indication of the behavior of the bulk of the data. Therefore, a different distribution must be found for the remaining observations (under the threshold) and stitched with the GPD distribution of the tail in order to have a unified model of the entire range of precipitation amounts.

In the previous section several such extreme mixture models were developed and applied in many fields involving extreme events. One interesting extreme mixture model reported in the field of precipitation modeling is the work of *Furrer and Katz* (2008), where a hybrid model based on a mixture of gamma and GPD distributions is presented. An important advantage of this model is that it accounts for the continuity at the jointure point (*i.e.*, threshold) by adjusting the scale parameter of the GPD.

Also, *Vrac and Naveau* (2007) applied the dynamic weighted model suggested by *Frigessi et al.* (2002), but with a gamma distribution instead of a Weibull. Unlike the previous extreme mixture models, this model does not require a threshold selection, but instead the threshold is defined dynamically by means of a weight function. This model is designed to weight more the gamma distribution for low-intensity values and the GPD for high values. The results show that the inference is not that straightforward, especially regarding the GPD shape parameter. Moreover, the number of parameters is high.

A more recent approach has been introduced by *Naveau et al.* (2016). The aim of this method is to model the entire precipitation range, without using parametric mixture models that can increase the number of parameters and also to bypass the threshold selection. The

proposed solution is to use an extended version of GPD. This approach provides a smooth transition between the tails and the bulk of the distribution. A more detailed description of this model is given in Section 1.3.2.

1.3 EGPD in the EVT framework

EGPD is a statistical model in which EVT is applied in both left and right tails, allowing a smooth transition between the two ends. We start by providing in Section 1.3.1 some key points regarding EVT and then, in Section 1.3.2, we illustrate the theoretical background behind the EGPD class.

1.3.1 Extreme value theory

As presented in *Coles* (2001), two different approaches exist for extreme value analysis:

- block maximum

  It involves modeling the maximum values of a sequence, *e.g.,* a year, and it is modeled with a Generalized Extreme Value (GEV) distribution.

- peaks over threshold

  It relies on the modeling of the values above a threshold, and it is modeled by a GPD.

Let $X_1, X_2, ..., X_n$ be a sequence of n i.i.d. variables with common marginal distribution function F. The distribution of maximum values M_n, *i.e.*, $M_n = \max\{X_1, X_2, ..., X_n\}$, is given by (2.1.3).

$$\begin{aligned}\mathbb{P}(M_n \leq z) =& \mathbb{P}(X_1 \leq z, X_2 \leq z, ..., X_n \leq z)\\ =& \mathbb{P}(X_1 \leq z)...\mathbb{P}(X_n \leq z)\\ =& F^n(z)\end{aligned} \tag{2.1.3}$$

In practice, the distribution function F is unknown, so (2.1.3) cannot be directly applied to model M_n, but the Fisher–Tippett–Gnedenko Theorem (*Fisher and Tippett* (1928), *Gnedenko* (1943)), stated below in Theorem 2.1.1, provides an asymptotic result, as $n \to \infty$.

Theorem 2.1.1. *Let $X_1, X_2, ..., X_n$ be a sequence of n i.i.d. variables with a common marginal distribution function F, and $M_n = \max\{X_1, X_2, ..., X_n\}$. If there exists a sequence of real numbers (a_n, b_n) such that $a_n > 0$, then*

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{M_n - a_n}{b_n} \leq z\right) = \lim_{n\to\infty} F^n(a_n z + b_n) = G(z),$$

where G is a continuous distribution function, belonging to either the Gumbel, the Fréchet or the Weibull family.

These three families can be grouped into a single distribution called the generalized extreme value (GEV) distribution, which has the cumulative distribution function

$$G(z) = \exp\left\{-\left(1 + \xi' \frac{z - \mu'}{\sigma'}\right)_+^{-\frac{1}{\xi'}}\right\}. \tag{2.1.4}$$

Here, $y_+ = \max(0, y)$, $\mu' \in \mathbb{R}$ is the location parameter, $\sigma' > 0$ the scale parameter and $\xi' \in \mathbb{R}$ the shape parameter.

As discussed in *Coles* (2001), a block maxima approach, can be wasteful. As the number of extreme values is very small by definition, by selecting just the maximum of a block we restrict even more the information given by the data. It is natural to take as extreme values the observations X_i that exceed some high threshold u. Considering this, an extreme event can be characterized by the conditional probability presented in (2.1.5).

$$\mathbb{P}(X > u + y | X > u) = \frac{1-F(u+y)}{1-F(u)}, \quad y > 0 \tag{2.1.5}$$

Going further, it was shown (*Coles* (2001)) that, for a large u, the distribution of the variable $X-u$, conditional on $X > u$ can be approximated by a family of distributions called Generalized Pareto distributions, recalled in Theorem 2.1.2.

Theorem 2.1.2. *Let $X_1, X_2, ..., X_n$ be a sequence of n i.i.d. observations of a random variable X, the excesses $X-u$ of an enough high threshold u can be well approximated by a Generalized Pareto distribution, defined as:*

$$H_\xi(x) = \begin{cases} 1 - \left(1 + \xi \frac{x-u}{\sigma}\right)_+^{-\frac{1}{\xi}}, & \xi \neq 0 \\ 1 - \exp(-\frac{x-u}{\sigma}), & \xi = 0, \end{cases} \tag{2.1.6}$$

for $x > 0$ and $y_+ = \max(0, y)$. Here $\sigma = \sigma' + \xi(u - \mu')$ and $\xi = \xi' \in \mathbb{R}$, are the scale and shape parameters, respectively, and μ', σ' and ξ' are given in (2.1.4).

The shape parameter ξ describes the GPD tail behavior:

- if $\xi < 0$: bounded upper tail,
- if $\xi = 0$: light-tailed (exponential distribution),
- if $\xi > 0$: heavy-tailed.

1.3.2 Extended generalized Pareto distribution

The excess over a threshold u (large values) of a random variable X can be approximated by a GPD, that is:

$$\mathbb{P}(X > x | X > u) = 1 - H_\xi(x) = \bar{H}_\xi(x), \tag{2.1.7}$$

where H_ξ is defined in (2.1.6).

As we are interested in modeling rainfall amounts, it is assumed that non-zero rainfall data have either an exponential tail ($\xi = 0$), or a heavy tail ($\xi > 0$); assumption that according to *Katz et al.* (2002), appears to be satisfied in practice by most rainfall datasets.

On the other hand, the left side of the distribution, that is the lower tail, can be regarded as well as an extreme value distribution, but this time by considering the minimum values. So, instead of looking at the random variable X, we look now at $Y = -X$. The largest negative values can be approximated by a GPD with a negative shape parameter, say $-1/\kappa$ for some $\kappa \geq 0$ and scale parameter υ. This translates to

$$\mathbb{P}(Y > -x | Y > -u) = 1 - H_{-1/\kappa}(x) = \bar{H}_{-1/\kappa}(x), \tag{2.1.8}$$

where $\bar{H}_{-1/\kappa}(x) = \left(1 - \frac{1}{\kappa}\frac{-x+u}{\upsilon}\right)^\kappa$.

As the upper limit of Y is zero, the threshold u has to be chosen so that $\bar{H}_{-1/\kappa}(0) = 0$, leading to the constraint $u = \kappa\upsilon$. By replacing u in (2.1.8) with this constraint, we have

$$\mathbb{P}(Y > -x | Y > -u) = (\tfrac{x}{\kappa\upsilon})^\kappa = (\kappa\upsilon)^{-\kappa} \cdot x^\kappa = cx^\kappa, \quad \text{for any small } x \geq 0. \tag{2.1.9}$$

Combining (2.1.7) and (2.1.9), we have the following tails behavior

$$\mathbb{P}(X \leq x) = \begin{cases} H_\xi(x), & \text{for any large } x \\ cx^\kappa, & \text{for any small } x \geq 0. \end{cases} \tag{2.1.10}$$

We can notice from (2.1.10) that the gamma density, defined in (2.1.11) and commonly used in rainfall modeling, is suitable for the small values, but it does not represent correctly the large ones, and the contrary happens for GPD.

$$g(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} \tag{2.1.11}$$

In order to build a full model such as (2.1.10), *Naveau et al.* (2016) related this problem to the idea introduced by *Papastathopoulos and Tawn* (2013) regarding generalized Pareto extensions. More specifically, this latter work proposed an extension to the GPD by incorporating into the model an additional parameter that allows the modeling of the bulk of the distribution. The idea comes from the fact that, any continuous random variable X can be generated by applying its inverse cdf to some uniform random draws U. That is, a generalized Pareto (GP) random variable is generated via

$$X = \sigma H_\xi^{-1}(U), \tag{2.1.12}$$

where U is a uniform random variable on the unit interval, and H_ξ^{-1} is the inverse cdf of GPD.

As shown in *Papastathopoulos and Tawn* (2013), a simple way of introducing flexibility into this simulation is to replace the draws U from (2.1.12) with $V = G^{-1}(U)$, where G is a continuous cdf on the interval $[0, 1]$. This leads to

$$X = \sigma H_\xi^{-1}\left\{G^{-1}(U)\right\}. \tag{2.1.13}$$

Therefore, (2.1.13) leads to the following formulation of the extended generalized Pareto cdf and pdf

$$\begin{aligned} F(x) &= G\left\{H_\xi(x)\right\}, \\ f(x) &= g\left\{H_\xi(x)\right\} h_\xi(x). \end{aligned} \tag{2.1.14}$$

Here, $h_\xi(x)$ and $g(\cdot)$ denote the densities of GPD and G, respectively.

Naveau et al. (2016) adapted the EGPD class of models from (2.1.14) for the entire range of rainfall amounts. In fact, this adaptation is defined by finding a suitable G function, such that the behavior of the tails from (2.1.10) is maintained, that is $\lim_{x\to\infty} \frac{\bar{F}(x)}{\bar{H}_\xi(x)}$ and $\lim_{x\to 0} \frac{F(x)}{x^\kappa}$ should be finite and positive. As shown in *Naveau et al.* (2016), three constraints must be imposed on the function G:

$$\text{C1). } \lim_{v\to 0} \frac{\bar{G}(v)}{v} = c_1, \qquad \text{C2). } \lim_{v\to 0} \frac{G\{v\omega(v)\}}{G(v)} = c_2, \qquad \text{C3). } \lim_{v\to 0} \frac{G(v)}{v^\kappa} = c_3, \tag{2.1.15}$$

for some finite and positive numbers c_1, c_2, c_3.

Remark 2.1.1. Constraint C2) is just a supplementary constraint resulted from the computation $\lim_{v\to 0} \frac{F(v)}{v^\kappa} = \lim_{v\to 0} \frac{G\{v\omega(v)\}}{G(v)} \cdot \lim_{v\to 0} \frac{G(v)}{v^\kappa}$. This constraint is always constant and non-null (more exactly, equal to 1) as long as $\omega(v) = \frac{H_\xi(v)}{v}$, due to the fact that $\omega(v) = 1 + o(v)$, as $v \to 0$ (Taylor expansion).

Naveau et al. (2016) focused on four parametric families for G that comply with the three constraints from (2.1.15), based on the power function and beta cdf. More specifically, they

considered:

$$1)\ G(u) = u^{\kappa}, \kappa > 0, \tag{2.1.16a}$$
$$2)\ G(u) = pu^{\kappa_1} + (1-p)u^{\kappa_2}, \kappa_1, \kappa_2 > 0 \text{ and } p \in [0,1], \tag{2.1.16b}$$
$$3)\ G(u) = 1 - Q_{\delta}\left\{(1-u)^{\delta}\right\}, \delta > 0, \tag{2.1.16c}$$
$$4)\ G(u) = \left[1 - Q_{\delta}\left\{(1-u)^{\delta}\right\}\right]^{\kappa/2}, \kappa, \delta > 0, \tag{2.1.16d}$$

where Q_{δ} is the cdf of a two-parameter $(1/\delta, 2)$ beta random variable, *i.e.*, Q_{δ} becomes in this case $Q_{\delta}(u) = \frac{1+\delta}{\delta} u^{1/\delta}\left(1 - \frac{u}{1+\delta}\right)$.

A comparison of the four parametric EGPD models showed, in case studies on simulated samples and rainfall data, that EGPD model based on (2.1.16a) has the best performance. One drawback of these parametric EGPD models is that they do not offer enough flexibility in modeling of the bulk of the distribution.

In the next two chapters we introduce two new approaches that add more flexibility for the estimation of the bulk of the distribution. More specifically, in Chapter 2 we introduce the transformed kernel density estimator with an EGPD transformation, while in Chapter 3 we propose a new EGPD model based on a sparse mixture of beta densities.

Chapter 2

EGPD and kernel density estimation

Various methods have been proposed in Chapter 1 for estimating the entire distribution, simultaneously capturing the bulk and the extreme tails. Some of them use two distributions to model the data above and below a certain threshold. This threshold is considered in some studies as a parameter to be estimated, while others bypass this choice by using a smooth transition function between the bulk and the tail, thus overpassing the uncertainty in the threshold estimation. In this chapter we are interested in the nonparametric framework, thus we focus on the retransformed nonparametric kernel estimators with an EGPD transformation. Section 2.1 introduces a brief background on kernel density estimators, then Section 2.2 describes the proposed model. In Section 2.3 and Section 2.4 the performance of the proposed model is illustrated in case studies on simulated samples and rainfall data.

2.1 Classical univariate kernel density estimator

Early works regarding nonparametric density estimation were published by *Rosenblatt* (1956) and *Parzen* (1962). Since then, the interest in the nonparametric framework has increased intensively resulting in an extended literature. The reader is referred to *Silverman* (1986) and *Wand and Jones* (1995) for classical books on this subject.

For a sample $X_1, X_2, ..., X_n$ of n i.i.d. observations from a random variable X, the univariate kernel estimator of the unknown density $f(x)$ is defined as

$$f_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right), \tag{2.2.1}$$

where $K(\cdot)$ is a positive weighting function called kernel function, and $h > 0$ is the bandwidth or smoothing parameter. Various kernel functions are proposed in the literature, most common ones being the uniform, Epanechnikov or gaussian kernel.

While the kernel function is not very important in the estimation process, as stated by *Silverman* (1986), the choice of the smoothing parameter h is crucial. If it is too small, the density is overestimated and the variance of the estimations is increased, while if it is too large, the density is underestimated and the bias of the estimations is increased. Thus, a trade-off between the variance and the bias of the estimator must be found.

In the following, first, we recall the asymptotic bias and variance for a kernel density estimator in Section 2.1.1, and then, in Section 2.1.2 we review some common methods for the estimation of the bandwidth parameter.

2.1.1 Asymptotic theory for kernel density estimator

Assumption 2.2.1. Assumptions about the unknown density *f:*

(1). Let X be a random variable with unknown density f.

(2). Let f be a bounded and twice continuously differentiable function with bounded derivatives.

Assumption 2.2.2. Assumptions about the kernel function:

(1). $K(u) = K(-u)$,

(2). $\mu_0(K) = \int_{\mathbb{R}} K(u)du = 1$,

(3). $\mu_1(K) = \int_{\mathbb{R}} uK(u)du = 0$,

(4). $\mu_2(K) = \int_{\mathbb{R}} u^2K(u)du \neq 0$.

Lemma 2.2.1. *Under Assumption 2.2.1.(1) and 2.2.2.(2), it was stated that*

(a) $\mathbb{B}\text{ias}\{f_h(x)\} = \int_{\mathbb{R}} K(u)\{f(x+uh) - f(x)\}du$,

(b) $\mathbb{V}\text{ar}\{f_h(x)\} = \frac{1}{nh}\int_{\mathbb{R}} K^2(u) f(x+uh)du - \frac{1}{n}\left\{\int_{\mathbb{R}} K(u) f(x+uh)du\right\}^2$.

Considering Lemma 2.2.1, in Theorem 2.2.1 we recall the bias, variance and the mean integrated square error (MISE) for the kernel density estimator, as derived by *Silverman* (1986).

Theorem 2.2.1 (*Silverman* (1986))**.** *Suppose that $h \to 0$ and $nh \to \infty$ as $n \to \infty$. Then, under Assumption 2.2.1 and 2.2.2,*

(a) $\mathbb{B}\text{ias}\{f_h(x)\} = \frac{h^2}{2}f''(x)\mu_2(K) + o(h^2)$,

(b) $\mathbb{V}\text{ar}\{f_h(x)\} = \frac{1}{nh}f(x)\int_{\mathbb{R}} K^2(u)du + o(\frac{1}{nh})$,

(c) $\text{MISE}\{f_h(x)\} = \frac{1}{nh}\int_{\mathbb{R}} K^2(u)du + \frac{h^4}{4}\mu_2^2(K)\int_{\mathbb{R}} f''(x)^2dx + o(\frac{1}{nh} + h^4)$.

Thus, a small bandwidth h will decrease the second term of the MISE, but it will increase the first one because it is proportional to $(nh)^{-1}$. Therefore, as $n \to \infty$, h must change in such a way that both components in MISE are small.

2.1.2 Smoothing parameter estimation

There are several methods reported in the literature for the estimation of the bandwidth, see for example *Silverman* (1986) and *Wand and Jones* (1995).

One possible approach is to minimize an error function such as MISE, leading to an optimal bandwidth of the form

$$h_{opt}^{\text{MISE}} = \left\{\frac{\|K\|_2^2}{n\mu_2^2(K)\|f''(x)\|_2^2}\right\}^{\frac{1}{5}}, \tag{2.2.2}$$

where $||\cdot||_2$ denotes the l_2-norm.

The computation of h_{opt}^{MISE} is not that straightforward as f is an unknown function. However, some conclusions can be drawn from (2.2.2) with respect to the optimal bandwidth. First, it can be noticed that h converges to zero as the sample size n increases. Also, considering

that $\|f^{''}(x)\|_2^2$ measures, in a sense, how quick the density is changing the curvature, then an accelerated change leads to a smaller bandwidth h.

A natural method, or "quick and simple" as referred in *Wand and Jones* (1995), is to compute h_{opt}^{MISE} with reference to a standard family of densities. For example, if K is a Gaussian kernel function and the reference distribution is a standard normal density with variance σ^2, then the optimal bandwidth from (2.2.2) becomes

$$h_{opt} = \left(\frac{4\sigma^5}{3n}\right)^{\frac{1}{5}}, \tag{2.2.3}$$

As reported in *Silverman* (1986), an easy way to compute h_{opt} from (2.2.3) is to replace σ with the sample standard deviation. This estimation approach is also known as *"Silverman's rule-of-thumb"*. As pointed out in this reference, while this choice of bandwidth is working well if the true distribution is really normally distributed, it might lead to over-smoothing if the true distribution is multimodal.

Other methods for the estimation of h involve least-squares cross-validation initially proposed by *Rudemo* (1982) and *Bowman et al.* (1984), maximum likelihood estimator (referred in the literature as cross-validated likelihood) introduced by *Duin* (1976) or plug-in approaches. Even though there are several methods for selecting h, there is no general accepted one that works well in both theory and practice.

2.1.3 Boundary effects and correction methods

The classical kernel density estimator, originally introduced by *Rosenblatt* (1956), has the purpose of estimating densities with unbounded support. But rainfall or runoff need a compact support, such as $[0, \infty)$. In this case, the bias and variance of the kernel density estimator, recalled in Theorem 2.2.1, can be expressed as

$$\begin{aligned}\mathbb{B}\text{ias}\left\{f_h(x)\right\} =& -f(x)\int_0^{\infty} K(u)du - hf^{'}(x)\int_{-\infty}^{0} uK(u)du \\ &+ \frac{h^2}{2}f^{''}(x)\int_{-\infty}^{0} u^2K(u)du + o(h^2),\end{aligned} \tag{2.2.4}$$

$$\mathbb{V}\text{ar}\left\{f_h(x)\right\} = \frac{1}{nh}f(x)\int_{-\infty}^{0} K^2(u)du + o\left(\frac{1}{nh}\right).$$

For a symmetric and non-negative kernel function, $\int_0^{\infty} K(u)du \neq 1$ and $\int_{-\infty}^{0} uK(u)du \neq 0$, resulting in a non-consistent estimator at the boundary points, *i.e.*, near zero in this case.

An immediate "naive" modification that can be applied to the kernel density estimator in order to make it consistent for a support $[0, \infty)$, is to normalize it with $\int_0^{\infty} K(u)du$. However, this adjustment does not account for the bias order at the boundary, which it still of order $o(h)$ rather than $o(h^2)$.

A variety of methods have been proposed in the literature for correcting the boundary bias. Among these, we recall the reflection method introduced by *Schuster* (1985), which is a consistent estimator at the boundary, but it still maintains the $o(h)$ bias order unless the assumption $f^{'}(x) = 0$ holds. The generalized jack-knifing of *Jones* (1993) overcomes this problem, but the new estimator no longer integrates to 1. More recent methods, such as the works introduced by *Chen* (1999, 2000) or *Jones and Henderson* (2007), deal with boundary correction for a compact support $[0, \infty)$ or $[0, 1]$, respectively, by using gamma-, beta-, or copula-based kernels functions.

2.2 Kernel density estimator with EGPD transformation

We have recalled in Section 2.1 the formulation and properties of the kernel density estimator. As our objective is modeling the full range of rainfall amounts, in this section we address the problem of kernel density estimation when the distribution under study is heavy tailed. We have reviewed in Chapter 1 several methods that can be applied for modeling heavy tailed distributions while preserving the nonparametric framework, but the most promising one appeared to be the transformed-retransformed model with an adaptive transformation. Briefly, we remind that this approach consists of, first, transforming the data with a continuous cdf, and then, estimating the density of the transformed data with a nonparametric density estimator.

The transformed-retransformed model has been previously used in some applications in the field of finance, as we have seen in Chapter 1. However, to the best of our knowledge, it has never been applied in hydrology for rainfall modeling, until now. The lack of research in this direction is probably due to the fact that, until recently, there was no suitable distribution for modeling the full range of rainfall amounts, *i.e.*, the cdf required by the transformation step. The work of *Naveau et al.* (2016) has opened new possibilities in this direction, as a new class of distributions was introduced under the name of Extended Generalized Pareto Distribution (EGPD). This model is in compliance with the EVT for both lower and upper tails, and at the same time, it allows a smooth transition between the two ends. However, the estimation of the bulk of the distribution was limited to a few parametric cases. The novelty of our study is to adopt a kernel density estimator within the EGPD class in order to add flexibility.

2.2.1 Transformed kernel: definition and properties

Let $X_1, X_2, ..., X_n$ be n i.i.d. observations from a random variable X with unknown density f. Suppose that f cannot be accurately estimated by the classical kernel density estimator from (2.2.1) due to the scarce observations from the upper tail. One possible solution in this case is to apply a transformation to the data in order to obtain a new variable $Y_1, Y_2, ..., Y_n$ which has a density g_t that can be more easily estimated by a kernel density estimator. Then, by back-transforming the estimate of g_t, one would obtain the estimate of f. According to *Wand and Jones* (1995), the resulted estimator is referred to as *transformation kernel density estimator* or *transformed kernel density estimator*.

Suppose now that this transformation is $Y_i = T(X_i)$, where T is a monotone increasing and differentiable function on the support of X. Then, by using standard properties of the statistical distribution theory (see *Bertsekas and Tsitsiklis* (2008)), *i.e.*, $X_i = T^{-1}(Y_i)$ and $F(x) = \mathbb{P}(X \leq x) = \mathbb{P}\left\{T^{-1}(Y) \leq x\right\} = \mathbb{P}\left\{Y \leq T(x)\right\}$, the cdf and pdf of the variable X based on the transformation T were derived as

$$\begin{aligned} F(x) &= G_t\left\{T(x)\right\}, \\ f(x) &= g_t\left\{T(x)\right\} T^{'}(x). \end{aligned} \tag{2.2.5}$$

Here, g_t and G_t are the unknown pdf and cdf of Y, while T^{-1} and $T^{'}$ represent the inverse and the derivative of the transformation function T. By replacing g_t from (2.2.5) with the classical kernel density estimator from (2.2.1), it leads to the following explicit formulation of the transformed kernel density estimator based on the transformation function T, as shown in *Wand and Jones* (1995),

$$f_{\mathrm{TK},h}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left\{\frac{T(x) - T(X_i)}{h}\right\} T^{'}(x), \tag{2.2.6}$$

where $K(\cdot)$ is a kernel function and $h > 0$ is the bandwidth parameter.

Considering Lemma 2.2.1, the bias and variance of a transformed kernel estimator can be easily derived, as shown in *Buch-Larsen et al.* (2005) and restated here in Theorem 2.2.2.

Theorem 2.2.2 (*Buch-Larsen et al.* (2005))**.** *Suppose that* $h \to 0$ *and* $nh \to \infty$*, as* $n \to \infty$*. Then, under Assumptions 2.2.1 and 2.2.2 we have,*

(a) $\mathbb{B}\text{ias}\left\{f_{TK,h}(x)\right\} = \frac{h^2}{2}\mu_2(K)\left[\left\{\frac{f(x)}{T'(x)}\right\}' \frac{1}{T'(x)}\right]' + o(h^2)$,

(b) $\mathbb{V}\text{ar}\left\{f_{TK,h}(x)\right\} = \frac{1}{nh} f(x) T'(x) \int_{\mathbb{R}} K^2(u)du + o\left(\frac{1}{nh}\right)$.

In the following, we focus on the case where the transformation function T is the cdf F_{EGPD} of EGPD, previously introduced in (2.1.14). Recall that $F_{\text{EGPD}}(x) = G\left\{H_\xi(x)\right\}$, where G represents a monotone increasing and differentiable function on $[0,1]$ and H_ξ denotes the cdf of GPD. It can be noticed that, in fact, EGPD is itself a class of distributions such as the ones displayed in (2.2.5), where G is the cdf of the transformed variable $Z = H_\xi(X)$. Thus, the cdf and pdf of the variable X based on the EGPD transformation becomes

$$\begin{aligned} F(x) &= G_t\left\{F_{\text{EGPD}}(x)\right\}, \\ f(x) &= g_t\left\{F_{\text{EGPD}}(x)\right\} f_{\text{EGPD}}(x). \end{aligned} \tag{2.2.7}$$

Here f_{EGPD} denotes the derivative of F_{EGPD}, *i.e.*, the pdf of the EGPD. If g_t is replaced with the classical kernel density estimator from (2.2.1), the transformed kernel density estimator based on the transformation function F_{EGPD} is

$$f_{\text{TK},h}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left\{\frac{F_{\text{EGPD}}(x) - F_{\text{EGPD}}(X_i)}{h}\right\} f_{\text{EGPD}}(x). \tag{2.2.8}$$

The use of a cdf as transformation function implies a change in the definition domain of the g_t function from (2.2.7) or the K function from (2.2.8). More specifically, while for a general transformation T the domain is $\mathbb{R}$, *i.e.*, $g_t : \mathbb{R} \to \mathbb{R}$, for a transformation F_{EGPD} the domain becomes $[0,1]$, *i.e.*, $g_t : [0,1] \to \mathbb{R}$. Thus, a boundary correction must be considered for g_t or K in order to ensure the consistency of the transformed kernel estimator, as already explained in Section 2.1.3.

2.2.1.1 Boundary corrected kernel function on $[0,1]$: definition and properties

For this work, we are interested in a boundary corrected kernel function on the unit interval, and thus, we focus on the work of *Jones and Henderson* (2007) that introduced a copula-based kernel function. This kernel is simply defined as the conditional density of a symmetric copula, such as a gaussian copula.

Let (U,V) be a bivariate standard normal random variable with correlation ρ, then the conditional gaussian copula density of U conditional on $V = v$ is

$$\begin{aligned} c_{u|v}(u;v,\rho) =& \frac{1}{\sqrt{1-\rho^2}} \exp\left[-\frac{\rho^2\left\{\Phi^{-1}(u)\right\}^2 - 2\rho\Phi^{-1}(u)\Phi^{-1}(v) + \rho^2\left\{\Phi^{-1}(v)\right\}^2}{2(1-\rho^2)}\right] \\ =& \frac{1}{\sqrt{1-\rho^2}\,\phi\left\{\Phi^{-1}(u)\right\}} \phi\left\{\frac{\Phi^{-1}(u) - \rho\Phi^{-1}(v)}{\sqrt{1-\rho^2}}\right\}, \end{aligned} \tag{2.2.9}$$

where ϕ and Φ^{-1} are the standard normal density function and the inverse of its cdf, respectively.

Assumption 2.2.3.

(1). Let $U_1, U_2, ..., U_n$ *be* n *i.i.d. observations from a random variable* U *with support* $[0,1]$ *and unknown density* g*.*

(2). Let g be a bounded and twice continuously differentiable function with bounded derivatives.

Under Assumption 2.2.3, we restate in (2.2.10) the gaussian copula-based kernel density estimator of g, as showed by *Jones and Henderson* (2007).

$$g_{\mathrm{GC},h}(u) = \frac{1}{n}\sum_{i=1}^{n} c_{u|U_i}(u; U_i, 1-h^2) \tag{2.2.10}$$

Here $h = \sqrt{1-\rho}$ is the bandwidth of the kernel function.

Since $c_{u|v}(u; v, \rho)$ is a proper density function, *i.e.*, $\int_0^1 c_{u|v}(u; v, \rho)du = 1$, the gaussian copula-based kernel density estimator $g_{\mathrm{GC},h}(u)$ integrates to one as well, and thus, it does not require normalization as in other boundary corrected kernel cases, such as *Jones* (1993) or *Chen* (1999).

Theorem 2.2.3 provides the bias and variance for the above estimator, as stated in *Jones and Henderson* (2007).

Theorem 2.2.3 (*Jones and Henderson* (2007))**.** *Suppose that $h \to 0$ and $nh \to \infty$, as $n \to \infty$. Under Assumption 2.2.3, we have that*

1. $\mathbb{B}\mathrm{ias}\left\{g_{GC,h}(u)\right\} = -2h^2\Phi^{-1}(u)\phi\left\{\Phi^{-1}(u)\right\}g^{'}(u) + h^2\phi^2\left\{\Phi^{-1}(u)\right\}g^{''}(u) + o(h^2)$,

2. $\mathbb{V}\mathrm{ar}\left\{g_{GC,h}(u)\right\} = \begin{cases} \frac{g(u)}{2\sqrt{\pi}\phi\{\Phi^{-1}(u)\}nh} + o\left(\frac{1}{nh}\right), & \text{if } \frac{u}{h^m} \text{ and } \frac{1-u}{h^m} \to \infty, \\ \frac{g(u)}{\sqrt{2}\eta^2 nh^{2m+1}} + o\left(\frac{1}{nh^{2m+1}}\right), & \text{if } \frac{u}{h^m} \text{ or } \frac{1-u}{h^m} \to \eta, \end{cases}$

where $m, \eta > 0$.

The asymptotic bias is of order $o(h^2)$ for all u, meaning that copula-based kernel estimator is free of boundary effects. But unfortunately, this bias is achieved with the cost of an increased boundary variance near the endpoints 0 or 1 (see 2$^{\text{nd}}$ term in the variance formulation from Theorem 2.2.3).

In this work we use the *"rule-of-thumb"* bandwidth selector proposed by *Jones and Henderson* (2007) as

$$h_{\mathrm{GC}}^{\mathrm{ROT}} = \hat{\sigma}\left\{2\hat{\mu}^2\hat{\sigma}^2 + 3(1-\hat{\sigma}^2)^2\right\}^{-\frac{1}{5}} n^{-\frac{1}{5}}, \tag{2.2.11}$$

where $\hat{\mu}$ and $\hat{\sigma}$ are the sample mean and variance. We refer also to Section 2.1.2 for some other general methods applied for computing the smoothing parameter of a kernel function.

2.2.2 Boundary-corrected transformed kernel (BCTK)

2.2.2.1 Model estimation

Our proposed model is the transformed kernel estimator based on EGPD transformation when the boundary-corrected kernel is the copula-based kernel. Therefore, by replacing g_t in (2.2.7) with the copula-based kernel $g_{\mathrm{GC},h}$, the pdf and the cdf of our proposed estimator can be expressed as

$$\begin{aligned} f_{\mathrm{BCTK},h}(x) &= g_{\mathrm{GC},h}\left\{F_{\mathrm{EGPD}}(x)\right\} f_{\mathrm{EGPD}}(x), \\ F_{\mathrm{BCTK},h}(x) &= G_{\mathrm{GC},h}\left\{F_{\mathrm{EGPD}}(x)\right\}. \end{aligned} \tag{2.2.12}$$

Here, $G_{\mathrm{GC},h}$ is the cdf of the gaussian copula-based kernel density estimator, *i.e.*, $G_{\mathrm{GC},h}(u) = \int_0^u g_{\mathrm{GC},h}(v)dv$.

The estimation of a transformed kernel model is a procedure including several steps. First, the parameters of the transformation function have to be estimated, and then, the bandwidth of the kernel function must be computed. The steps that one should follow to find the approximation $f_{\mathrm{BCTK},h}$ of the unknown density f, are:

Step 1: Estimate the parameters of the EGPD model considering the random variable $X = \{X_1, ..., X_n\}$.

There are several inference methods that can be used in this case, including the maximum likelihood estimator and the probability weighted moments, presented in the work of *Naveau et al.* (2016).

Step 2: Transform X with the EGPD cdf estimated in the previous step and find the transformed data or the pseudo-observations set $U = \{U_1, ..., U_n\}$, computed as:

$$U_i = \hat{F}_{\text{EGPD}}(X_i), \text{ for } i = 1, ..., n.$$

Thus, this transformation converts the data into pseudo-observations in the interval $[0, 1]$.

Step 3: Apply the boundary-corrected kernel on the transformed data U.

1. Compute the kernel bandwidth $h_{\text{GC}}^{\text{ROT}}$ defined in (2.2.11).
2. Compute the gaussian copula-based kernel $g_{\text{GC},h_{\text{GC}}^{\text{ROT}}}(u)$.

Step 4: Compute the estimator of the original data X, as

$$f_{\text{BCTK},h_{\text{GC}}^{\text{ROT}}}(x) = g_{\text{GC},h_{\text{GC}}^{\text{ROT}}}\left\{\hat{F}_{\text{EGPD}}(x)\right\}\hat{f}_{\text{EGPD}}(x).$$

2.2.2.2 Lower and upper tail equivalence

Naveau et al. (2016) showed that both the lower and upper tails of rainfall amounts distribution comply with EVT. More exactly, the lower tail (small values) behaves like a power function, while the upper tail (large values) like a GPD. In this subsection, we check if this behavior is satisfied by our proposed model. Therefore, we compute the two limits: $\lim_{x\to 0}\frac{F_{\text{BCTK},h}(x)}{x^k}$ and $\lim_{x\to\infty}\frac{\bar{F}_{\text{BCTK},h}(x)}{\bar{H}_\xi(x)}$. In theory, the tails are equivalent if these two limits are finite and positive. Unfortunately, we prove below that these limiting constraints are null and, thus, they are not satisfied for $F_{\text{BCTK},h}$.

Upper tail: $\lim_{x\to\infty}\frac{\bar{F}_{\mathbf{BCTK},h}(x)}{\bar{H}_\xi(x)} = 0$

Proof.

$$\lim_{x\to\infty}\frac{\bar{F}_{\text{BCTK},h}(x)}{\bar{H}_\xi(x)} = \lim_{x\to\infty}\frac{1-G_{\text{GC},h}\{F_{\text{EGPD}}(x)\}}{\bar{H}_\xi(x)} = \lim_{x\to\infty}\frac{1-G_{\text{GC},h}[G\{H_\xi(x)\}]}{\bar{H}_\xi(x)}$$

By replacing $v = \bar{H}_\xi(x)$ and applying l'Hopital rule, we have:

$$\lim_{v\to 0}\frac{1-G_{\text{GC},h}\{G(1-v)\}}{v} = \lim_{v\to 0} g_{\text{GC},h}\{G(1-v)\}\cdot g(1-v) = 0, \text{ because } g_{\text{GC},h}(1) = 0.$$

■

Lower tail: $\lim_{x\to 0}\frac{F_{\text{BCTK},h}(x)}{x^k} = 0$

Proof.

$$\lim_{x\to 0}\frac{F_{\text{BCTK},h}(x)}{x^k} = \lim_{x\to 0}\frac{G_{\text{GC},h}\{F_{\text{EGPD}}(x)\}}{F_{\text{EGPD}}(x)}\cdot\frac{F_{\text{EGPD}}(x)}{x^k}$$

We already know from *Naveau et al.* (2016) that the second term in the above limit is finite and positive (see (2.1.15) in Section 1.3.2), so we must focus now on the first part of the limit.

Thus, by replacing $v = F_{\text{EGPD}}(x)$ and applying l'Hopital rule, we have:

$$\lim_{v\to 0} \frac{G_{\text{GC},h}(v)}{v} = \lim_{v\to 0} g_{\text{GC},h}(v) = 0$$

■

In conclusion, the transformed kernel estimator with EGPD transformation does not satisfy the tail equivalence suggested by *Naveau et al.* (2016), neither for the lower, nor for the upper tail, since both limits are null. However, we will see in the case study section that, even without complying with these conditions, the model $F_{\text{BCTK},h}$ provides accurate and improved estimates, if the sample size is large enough.

2.3 Case study: simulated data

In this section, simulations from various distributions with different degree of heaviness and multimodality, as detailed in Section 2.3.1, are used to assess the performance of the proposed kernel density estimator with EGPD transformation.

2.3.1 Simulated samples

We draw data from three distributions: a mixture of two gamma and GP densities (Mix2GaGPD), a mixture of three gamma and GP densities (Mix3GaGPD), and a mixture of two Singh-Maddala densities (Mix2SM). The setting for each corresponding mixture scenario is displayed in Table 2.2.1.

Table 2.2.1: The expression of the densities and the corresponding true parameters for the three scenarios used in the simulation study, namely Mix2GaGPD, Mix3GaGPD, and Mix2SM

Scenario	Density expression		True parameters
Mix2GaGPD	$f(x) = \begin{cases} f_{2Ga}(x;\alpha,\beta,p), & x \le u \\ \{1 - F_{2Ga}(u;\alpha,\beta,p)\}\, h_\xi(\frac{x}{\sigma}), & x > u \end{cases}$		$\alpha = (1,4)$, $\beta = (1,2)$ $(p,u,\xi) = (0.5, q_{90}, 0.3)$
Mix3GaGPD	$f(x) = \begin{cases} f_{3Ga}(x;\alpha,\beta,p), & x \le u \\ \{1 - F_{3Ga}(u;\alpha,\beta,p)\}\, h_\xi(\frac{x}{\sigma}), & x > u \end{cases}$		$\alpha = (2,6,10)$, $\beta = (1,2,1)$ $p = (0.5, 0.3, 0.2)$ $(u,\xi) = (q_{90}, 0.3)$
Mix2SM	$f(x) = \sum_{i=1}^{2} p_i \cdot f_{SM}(x; a_i, b_i, c_i)$		$a = (2,3)$, $b = (0.25, 0.6)$ $c = (4, 1.7)$, $p = (0.4, 0.6)$

where:

q_{90} denotes the 90%-quantile

$f_{kGa}(x;\alpha,\beta,p) = \sum_{i=1}^{k} p_i \cdot f_{Ga}(x;\alpha_i,\beta_i)$ and f_{Ga} is the pdf of Gamma distribution.

$f_{SM}(x) = aq\frac{1}{b^a}x^{a-1}\left\{1 + \left(\frac{x}{b}\right)^a\right\}^{-1-q}, x > 0.$

The continuity of $f(x)$ at u, *i.e.*, the jointure point of the mixture, is preserved by considering the scale parameter of GPD being automatically estimated with respect to the other parameters. More specifically, we consider $\sigma = \frac{1-F_{2Ga}(u;\alpha,\beta,p)}{f_{2Ga}(u;\alpha,\beta,p)}$ for Mix2GaGPD, and $\sigma = \frac{1-F_{3Ga}(u;\alpha,\beta,p)}{f_{3Ga}(u;\alpha,\beta,p)}$ for Mix3GaGPD, respectively.

Figure 2.2.1 displays the three pdfs considered in Table 2.2.1.

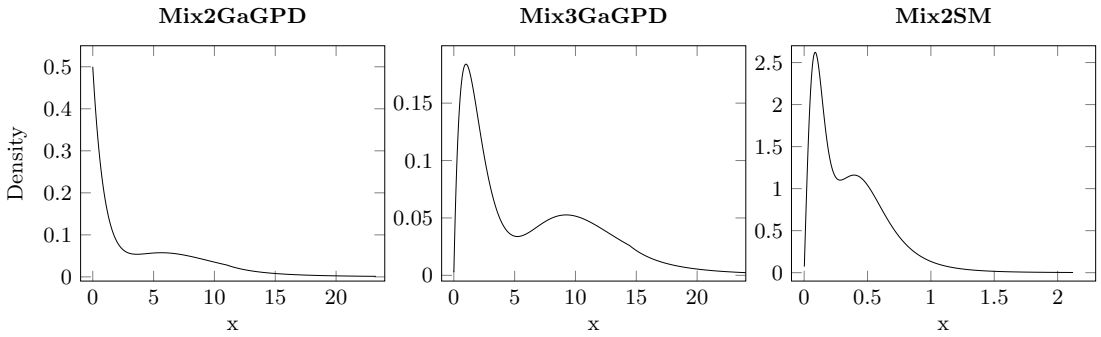

Figure 2.2.1: The shape of the three mixture densities used in the simulation case study, namely Mix2GaGPD, Mix3GaGPD, and Mix2SM, with the parameters setting summarized in Table 2.2.1

2.3.2 Measures of error

The performance of the fitted models is evaluated with respect to the estimated quantiles and estimated densities. The error of a given estimated quantile is measured by the following Root Mean Squared Error (RMSE)

$$\text{RMSE}(q_p) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{q}_{p,i} - q_{p,i})^2}, \tag{2.2.13}$$

where q_p and $\hat{q}_p$ denotes respectively, the true and estimated quantile of probability p, considering that $q_p = F^{-1}(p)$.

Then, two models, *e.g.,* M1 and M2, are compared by means of the Ratio Root Mean Squared Error (R_{RMSE}) as

$$\text{R}_{\text{RMSE}}(q_p) = \frac{\text{RMSE}_{\text{M1}}(q_p)}{\text{RMSE}_{\text{M2}}(q_p)}. \tag{2.2.14}$$

The density approximation error is evaluated by computing the Integrated Absolute Error, formulated in (2.2.15).

$$\text{IAE} = \int_{-\infty}^{+\infty} \left|\hat{f}(x) - f(x)\right| dx \tag{2.2.15}$$

2.3.3 Performance of the estimated models

We focus on the case where the EGPD transformation is the EGPD model from (2.1.16a). Our choice is based on the fact that this model performed the best in the numerical study from *Naveau et al.* (2016), where four models were evaluated (see (2.1.16a)-(2.1.16d)). From now on, we refer to this EGPD model as EGPD_1, and the transformed kernel estimator based on this transformation as TK_1.

The TK_1 model is fitted according to the procedure presented in Section 2.2.2.1. The performance of this model is evaluated by analyzing the estimated quantiles and the density approximation error on the simulated samples, but also by comparing these estimations with the ones yielded by EGPD_1. For each simulated mixture, we consider two sample sizes $n = (300, 1000)$, and each scenario is replicated 1000 times. We estimate 99 quantiles, from 0.01 to 0.99, equally spaced.

First, we focus on the estimation accuracy of the quantiles. Figure 2.2.2 displays the boxplots of the RMSEs of the 99 estimated quantiles over the 1000 replicates, for both TK_1 and $EGPD_1$, for each simulation setting. It can be noticed from this figure that while the performance of $EGPD_1$ is not greatly influenced by the increase in sample size, TK_1 is achieving an improved estimation error when the sample size is $n = 1000$, for all three mixtures. Also, TK_1 yields less variable and smaller RMSEs of the estimated quantiles compared to $EGPD_1$, especially for the case when the sample size is larger.

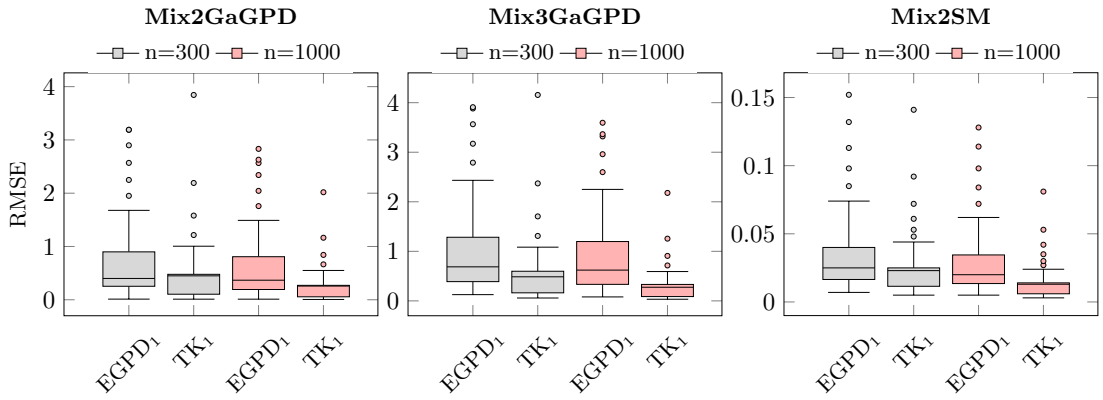

Figure 2.2.2: Boxplots of the RMSEs of the 99 estimated quantiles of both TK_1 and $EGPD_1$ models, over 1000 replicates from three simulated mixtures with sample sizes $n = 300$ and $n = 1000$

These boxplots, though, do no show exactly which and how many quantiles are better estimated by $EGPD_1$. We display, in this sense, in Table 2.2.2 the percentage of times the ratio between $RMSE_{EGPD_1}$ and $RMSE_{TK_1}$ of each quantile is larger than 1, that is, we account for the times when $EGPD_1$ has larger quantile estimation error. So, out of the 99 estimated quantiles, about 90% are better estimated by the TK_1 model, but this percentage increases as the sample size becomes larger. An extended analysis of the RMSEs and the ratios (not displayed here), shows that, generally, the quantiles around the mean, for both $n = 300$ and $n = 1000$, are better estimated by $EGPD_1$.

Table 2.2.2: Percentage of time the ratio between $RMSE_{EGPD_1}(q_p)$ and $RMSE_{TK_1}(q_p)$ of each estimated quantile is larger than 1, over 1000 replicates from three simulated mixture with sample sizes $n = 300$ and $n = 1000$

	n=300	**n=1000**
Mix2GaGPD	88.89%	93.94%
Mix3GaGPD	90.91%	94.95%
Mix2SM	87.88%	92.93%

We have seen in Section 2.2.2.2, that from a theoretical point of view, TK_1 does not have a GPD tail. Hence, we are interested now in the efficiency of the TK_1 model in estimating the extreme quantiles, such as 90%-, 95%-, 99%-quantile (q_{90}, q_{95}, q_{99}). The boxplots in Figure 2.2.3 represent the estimated values of the three extreme quantiles over the 1000 replicates, for each simulation scenario and the two sample sizes, *i.e.*, $n = 300$ and $n = 1000$. It can be noticed that, for all three simulated mixtures, TK_1 has a good assessment of the true quantiles (see horizontal red line), and the estimates have a smaller variability when the sample size is larger.

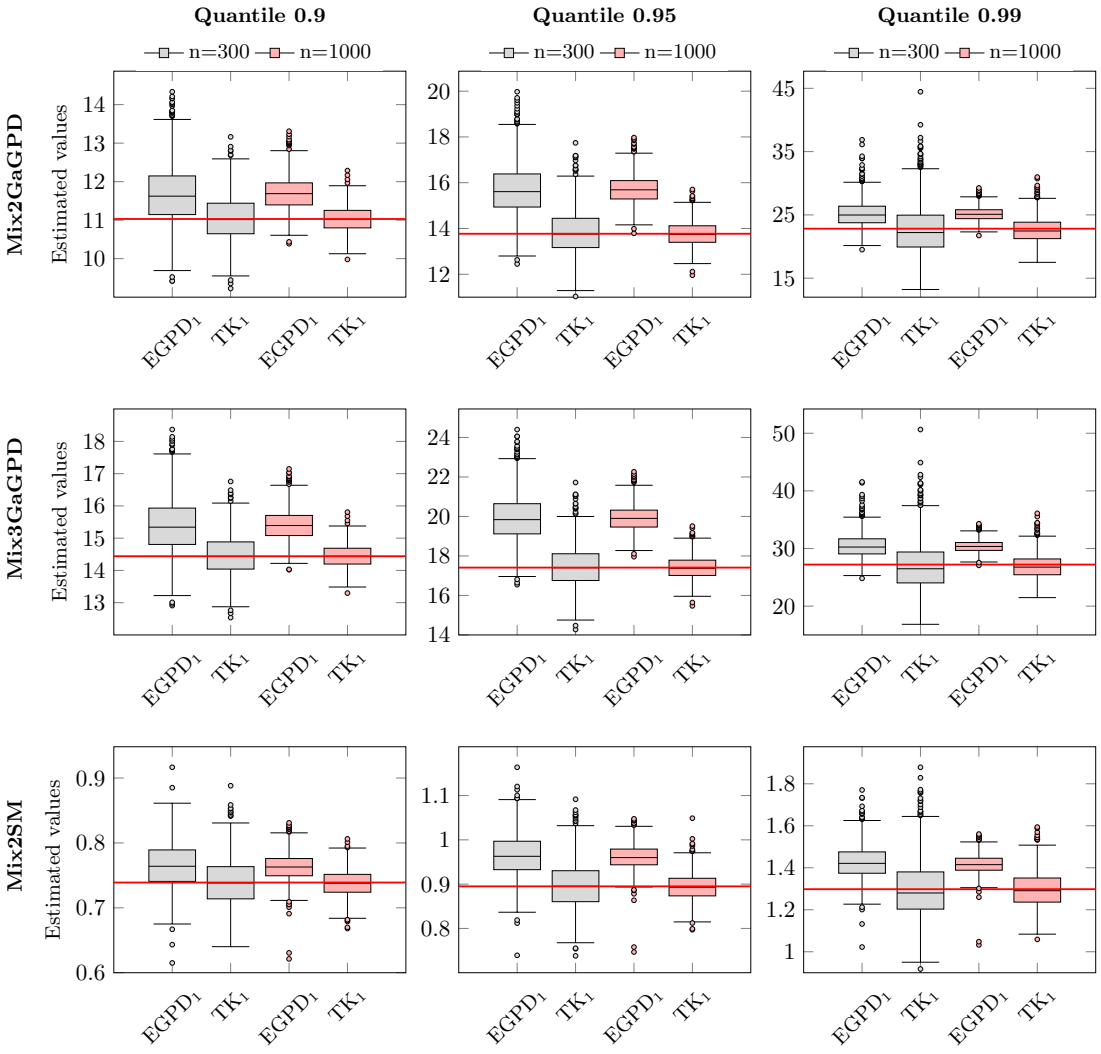

Figure 2.2.3: Boxplots of the estimated q_{90}, q_{95}, q_{99} quantiles of both TK_1 and EGPD_1 models, over 1000 replicates from three simulated mixtures with sample size $n = 300$ and $n = 1000$ (the red horizontal line indicates the true value of the quantiles)

To compare the performance of TK_1 with EGPD_1 more deeply, Table 2.2.3 displays the ratio between $\text{RMSE}_{\text{EGPD}_1}$ and $\text{RMSE}_{\text{TK}_1}$ for the three quantiles, *i.e.*, q_{90}, q_{95}, q_{99}. Recall that a ratio larger than 1 implies that TK_1 has a better performance in estimating that quantile than EGPD_1. Therefore, the ratios presented in this table show that the EGPD_1 model provides better estimates only for q_{99} when the sample size is small (red color cells), but these estimates are improving as the sample size increases to $n = 1000$. This behavior is also observed in Figure 2.2.3, where even though EGPD_1 seems to overestimate the true value of q_{99}, TK_1 provides considerably more variable estimates.

We have seen so far that TK_1 provides a good extreme quantile estimation, but also that it might be outperformed by EGPD_1 for some other particular quantiles, *e.g.*, q_{50}. We are interested now in the overall efficiency in quantile estimation, *i.e.*, we compute the RMSE of all quantiles for each replicate out of the 1000 runs. Table 2.2.4 displays the percentage of time the ratio between the RMSE of EGPD_1 and TK_1 (*i.e.*, $\text{R}_{\text{RMSE}}(sim_i) = \frac{\text{RMSE}_{\text{EGPD}_1}(sim_i)}{\text{RMSE}_{\text{TK}_1}(sim_i)}$) is

Table 2.2.3: Ratio between the RMSEs of EGPD_1 and TK_1 for $q_{0.9}, q_{0.95}, q_{0.99}$ quantiles, over 1000 replicates from three simulated mixtures with sample sizes $n = 300$ and $n = 1000$ (in red the cases where EGPD_1 performs better than TK_1)

	n=300			**n=1000**		
	$q_{0.9}$	$q_{0.95}$	$q_{0.99}$	$q_{0.9}$	$q_{0.95}$	$q_{0.99}$
Mix2GaGPD	1.708	2.238	0.830	2.424	3.716	1.272
Mix3GaGPD	2.000	2.574	0.934	3.040	4.389	1.544
Mix2SM	1.216	1.604	1.078	1.600	2.400	1.580

larger than 1. It can be seen that out of the 1000 runs, TK_1 has a better overall quantile fit compared to EGPD_1 in more than 90% of the simulated sets, and this percentage increases as the sample size is larger.

Table 2.2.4: Percentage of time the ratio between $\text{RMSE}_{\text{EGPD}_1}(sim_i)$ and $\text{RMSE}_{\text{TK}_1}(sim_i)$ is larger than 1, for $i = 1, 2, ..., 1000$ replicates from three simulated mixtures with sample sizes $n = 300$ and $n = 1000$

	n=300	**n=1000**
Mix2GaGPD	90.40%	99.60%
Mix3GaGPD	98.40%	100%
Mix2SM	90.80%	99.60%

We focus now on the last point analyzed in this simulation study, *i.e.*, the approximation error of the density. Table 2.2.5 shows the MIAE, for both TK_1 and EGPD_1 and each simulation scenario. A smaller MIAE is reported for TK_1 each time, and moreover, TK_1's performance in approximating the density is highly related to the sample size, yielding around a three times smaller MIAE compared to EGPD_1 when the sample size is $n = 1000$. This is even more clearly illustrated in Figure 2.2.4, where we display the fitted densities, for each model and simulated mixture. While EGPD_1 does not fit accurately the bulk of the distributions, TK_1 is more flexible in this sense.

Table 2.2.5: MIAE of the TK_1 and EGPD_1 models, over 1000 replicates from three simulated mixtures with sample sizes $n = 300$ and $n = 1000$ (the best approximation error between TK_1 and EGPD_1, on each sample size case, are indicated in red)

	n=300		**n=1000**	
	EGPD_1	**TK_1**	**EGPD_1**	**TK_1**
Mix2GaGPD	0.274	0.146	0.269	0.091
Mix3GaGPD	0.357	0.147	0.354	0.092
Mix2SM	0.213	0.149	0.209	0.092

2.4 Case study: rainfall data

We now apply our proposed density estimator on two rainfall datasets. The first dataset consists of hourly precipitation from 1996 to 2011 recorded at the Lyon station, France. The second one is composed of daily mean areal precipitation data over the 1948-2010 period for the Durance watershed, located in south-east France. More specifically, this aggregated time series is taken from the SPAZM meteorological analysis introduced by *Gottardi et al.* (2012), where a geostatistical approach based on weather pattern classification is applied to compute

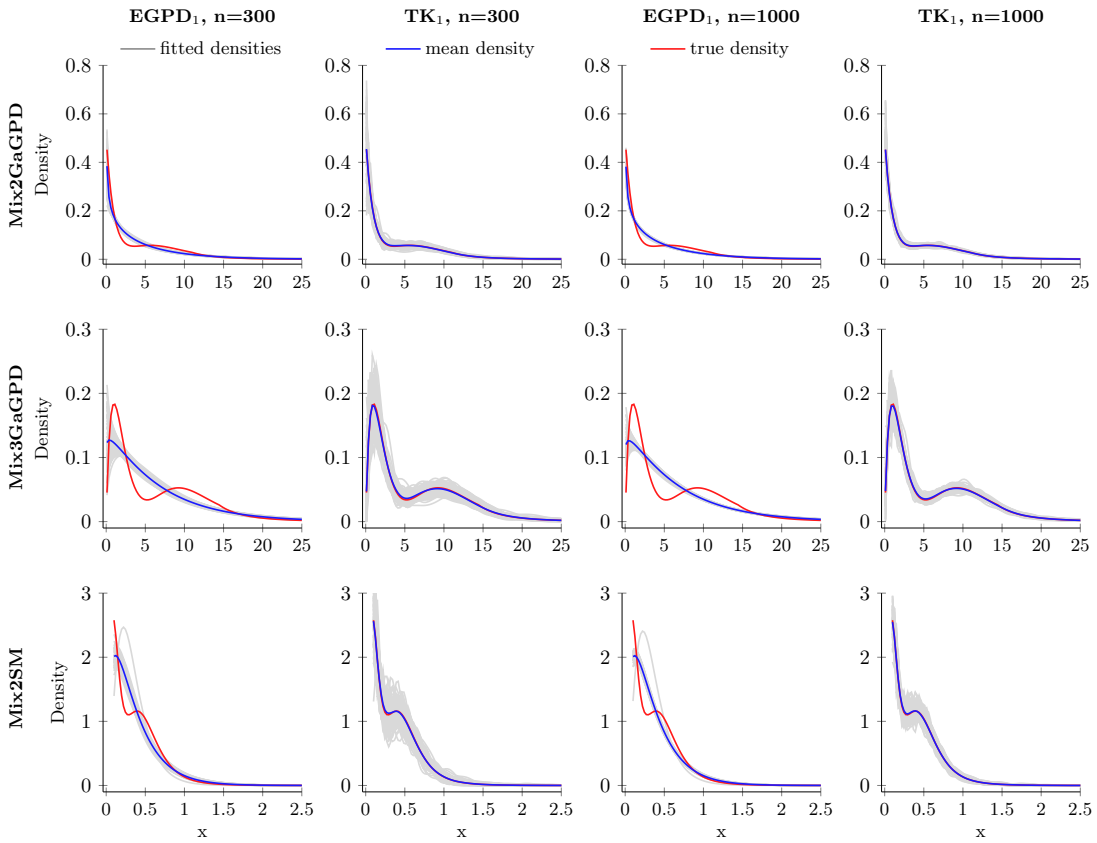

Figure 2.2.4: Fitted densities of the TK_1 and $EGPD_1$ models, over 1000 replicates from three simulated mixtures with sample sizes $n = 300$ and $n = 1000$ (red = the true density, gray = the density fit of each replicate, blue = the mean of the 1000 fitted densities)

the areal rainfall in the mountains. In the following, in order to simplify the presentation, we refer to the SPAZM series as "observations" and the location of the measurements as Durance station, even though they are not direct recordings at a specific point.

A short exploratory analysis of both time series showed that there is a clear seasonal pattern, so we model separately the four seasons: Spring (March-April-May), Summer (June-July-August), Fall (September-October-November), and Winter (December-January-February).

As there was some short-time temporal dependence between the observations, we only retained each third observation. Also as part of the preprocessing step, considering that our interest is in modeling just the amounts and not the occurrences, we removed the dry events (*i.e.*, zero precipitation values). After these steps, the sample sizes of our datasets are:

Lyon: Spring-282, Summer-251, Fall-336, Winter-184.

Durance: Spring-726, Summer-755, Fall-600, Winter-590.

The analysis of rainfall data is based on a comparison between the TK_1 and $EGPD_1$ models.

2.4.1 Rainfall at the Lyon station

Table 2.2.6 displays the estimated parameters and the associated 95% confidence intervals, for the TK_1 model, by applying the probability weighted moments (PWMs) approach for the

EGPD_1 fit, and the rule-of-thumb bandwidth for the kernel function. While the bandwidth is similar for all four seasons, the EGPD_1 parameters are different. This points out that an adaptive transformation, *i.e.*, based on a cdf, and not a fixed transformation (*e.g.*, logarithm) for the TK_1 is a good choice.

Table 2.2.6: Estimated parameters and the associated 95% confidence intervals for the TK_1 model at Lyon station, for each season

	$\hat{k}$	$\hat{\sigma}$	$\hat{\xi}$	$\hat{h}$
Spring	$1.055_{[0.94,1.37]}$	$1.484_{[1.14,1.73]}$	$0.035_{[0.00,0.18]}$	$0.071_{[0.069,0.073]}$
Summer	$0.864_{[0.74,1.07]}$	$2.365_{[1.85,2.92]}$	$0.128_{[0.00,0.28]}$	$0.075_{[0.074,0.076]}$
Fall	$0.883_{[0.76,1.03]}$	$1.686_{[1.30,2.00]}$	$0.193_{[0.10,0.31]}$	$0.068_{[0.067,0.070]}$
Winter	$1.008_{[0.79,1.31]}$	$0.834_{[0.55,1.09]}$	$0.312_{[0.14,0.49]}$	$0.070_{[0.066,0.073]}$

Figure 2.2.5 displays the histograms of the data with the estimated densities and the quantile-quantile plots (QQ-plots) with the associated 95% confidence intervals. The QQ-plots indicate a satisfactory fit, especially for seasons like Spring and Fall. TK_1 yields very close estimates to EGPD_1, but it improves considerably the EGPD_1 fit for the middle part of the distribution in Summer.

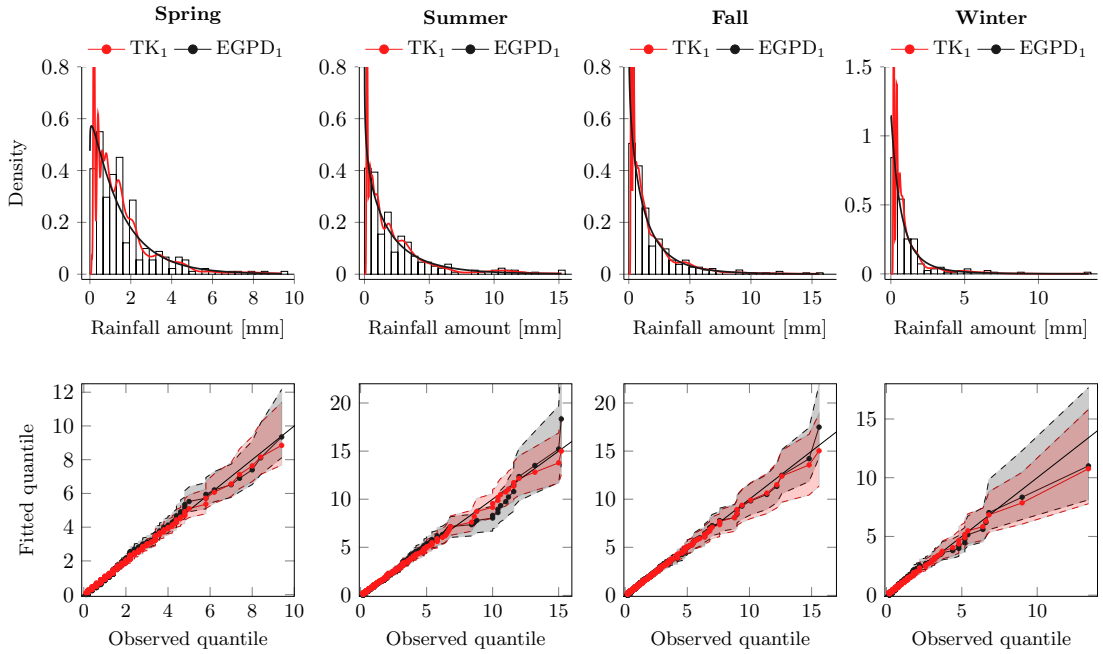

Figure 2.2.5: Histograms with fitted densities and QQ-plots with the associated 95% confidence intervals for the TK_1 and EGPD_1 models at Lyon Station, for each season

It is important also to notice that an increased variability is present in the TK_1 density estimation for small precipitation values. This effect might be caused by the estimated bandwidth of the copula-based kernel, which in this case might be too small, or by the discrete nature of these small value observations generated by the instrumental rounding at 0.1 [mm].

As part of the TK_1 inference, the density approximation on the transformed data is shown in Figure 2.2.6 for each season and both TK_1 and EGPD_1 models. Recall that for a random variable X, the transformed data for the TK_1 and EGPD_1 models are, respectively, $U = F_{\text{EGPD}_1}(X)$

and $V = H_\xi(X)$, where F_{EGPD_1} and H_ξ are the cdf of EGPD_1 and GPD, respectively. Thus, we display the histograms of the transformed data U and V with the fit of the copula-based kernel function and the power function (*i.e.*, $g(u) = ku^{k-1}$) of, respectively, TK_1 and EGPD_1 model. We can see that EGPD_1 is very close to a straight line (due to the fact that the estimated k parameters are near 1, thus exact GPD model), while TK_1 captures well the peaks and valleys of the density, thus adds more flexibility in modeling the middle part of the distribution.

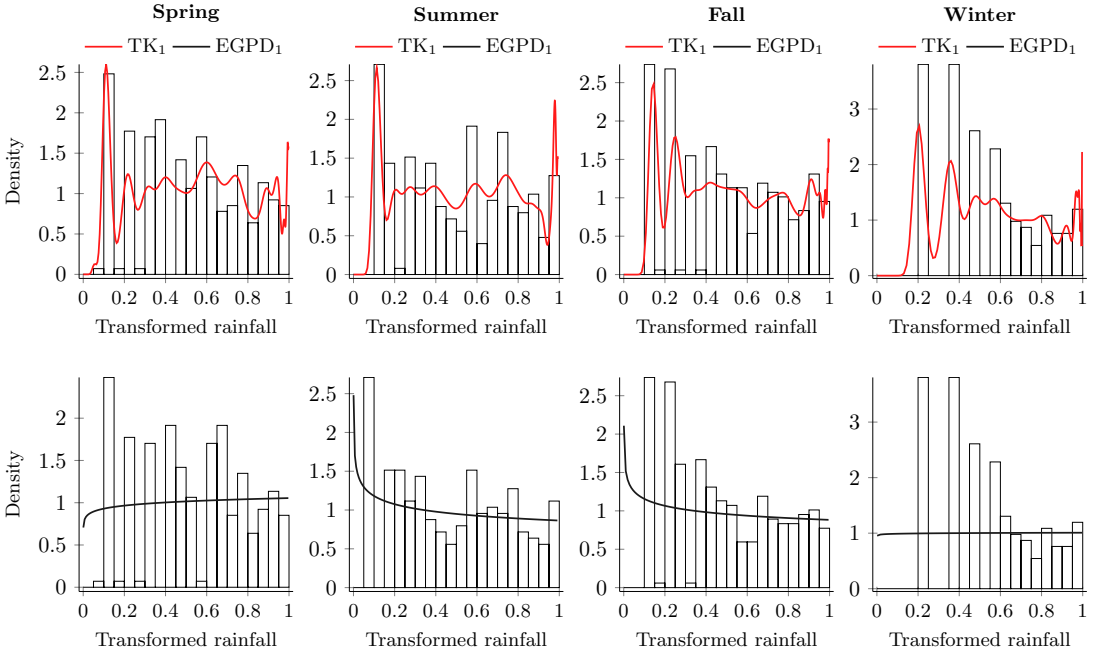

Figure 2.2.6: Histograms with fitted densities of the transformed data $U = F_{\text{EGPD}_1}(X)$ and $V = H_\xi(X)$, for, respectively, TK_1 and EGPD_1 models at Lyon station, for each season

We have seen that the estimation of the upper tail is, more or less, the same for both EGPD_1 and TK_1. Figure 2.2.7, showing a zoom on the small values, illustrates that the same conclusion can be drawn for the small values, with slightly better TK_1 estimates.

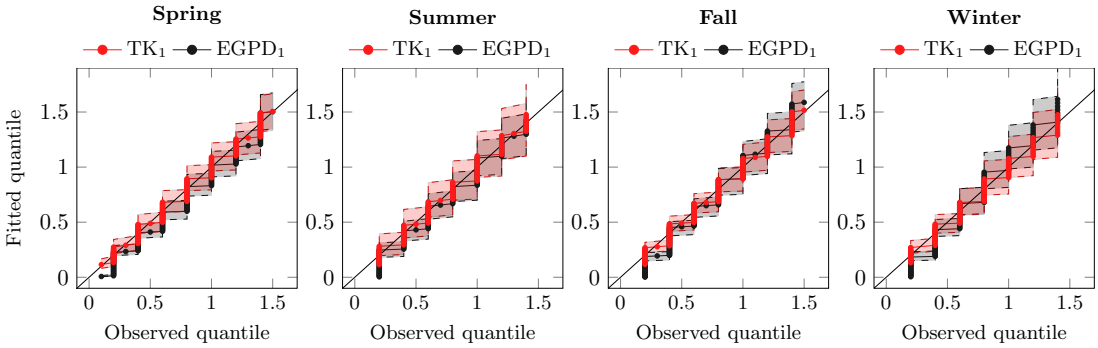

Figure 2.2.7: Zoom on the small values of the QQ-plots for the TK_1 and EGPD_1 models at Lyon station, for each season

2.4.2 Rainfall at the Durance station

In this section we analyze the performance of TK_1 on a daily precipitation dataset from the Durance station. The estimated parameters are shown in Table 2.2.7, along with the corresponding 95% confidence intervals. The bandwidth estimates are invariant with respect to the season, like in the previous rainfall case study, but, while for the Lyon dataset the parameter k was very close to 1, in this case study it is smaller, except for Summer.

Table 2.2.7: Estimated parameters and the associated 95% confidence intervals for the TK_1 model at Durance station, for each season

	$\hat{k}$	$\hat{\sigma}$	$\hat{\xi}$	$\hat{h}$
Spring	$0.719_{[0.64,0.83]}$	$7.690_{[6.32,9.66]}$	$0.137_{[0.03,0.23]}$	$0.063_{[0.062,0.063]}$
Summer	$0.996_{[0.87,1.16]}$	$4.368_{[3.65,5.26]}$	$0.276_{[0.17,0.35]}$	$0.063_{[0.062,0.063]}$
Fall	$0.688_{[0.60,0.78]}$	$12.242_{[9.69,15.50]}$	$0.107_{[0.02,0.20]}$	$0.066_{[0.065,0.066]}$
Winter	$0.683_{[0.63,0.80]}$	$10.418_{[8.27,12.04]}$	$0.018_{[0.00,0.18]}$	$0.066_{[0.066,0.068]}$

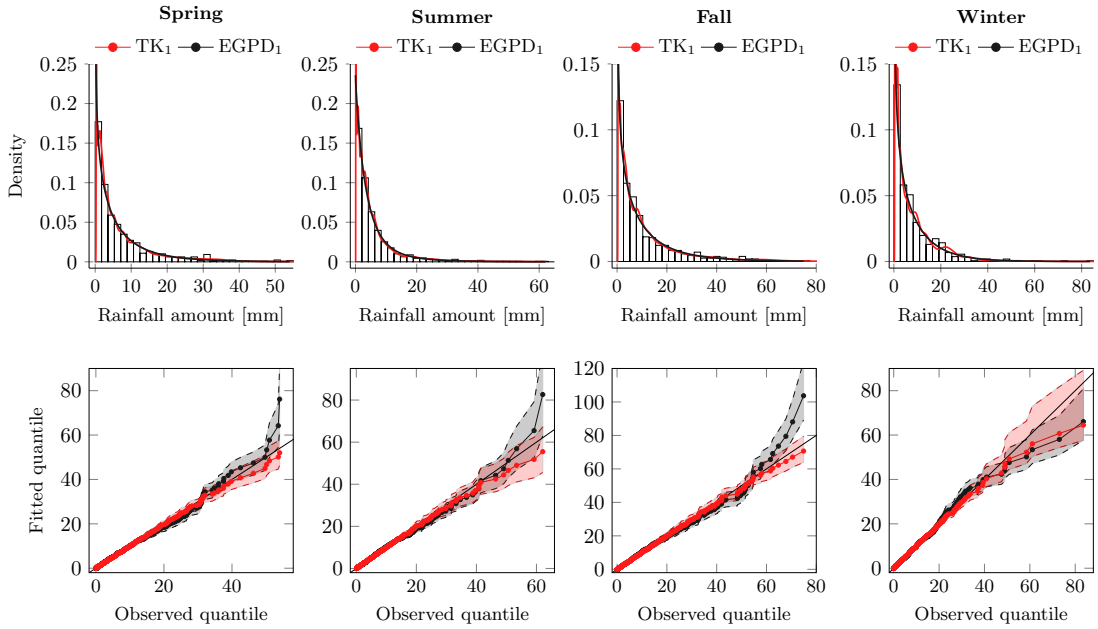

Figure 2.2.8: Histograms with fitted densities and QQ-plots with the associated 95% confidence intervals for the TK_1 and $EGPD_1$ models at Durance station, for each season

The QQ-plots illustrated in Figure 2.2.8 indicate that the right tail is poorly estimated by $EGPD_1$ for all seasons, while TK_1 corrects this behavior, especially during Fall.

To visualize the fit of the small values, Figure 2.2.9 shows a zoom on the QQ-plots for the four seasons. Contrary to the large values, the small ones are underestimated in the case of $EGPD_1$ and, thus, TK_1 outperforms $EGPD_1$, especially for the Spring, Fall and Winter seasons.

Figure 2.2.10 illustrates the fitted densities on the transform data on the interval $[0, 1]$, for both TK_1 and $EGPD_1$. TK_1 captures considerably better the behavior of these data, thus, offers more flexibility for the bulk of the rainfall amount distribution.

In summary, TK_1 estimator yields good estimates for both Lyon and Durance rainfall datasets. The use of a nonparametric family to model the transformed data adds flexibility compared to the power function choice from the $EGPD_1$ model.

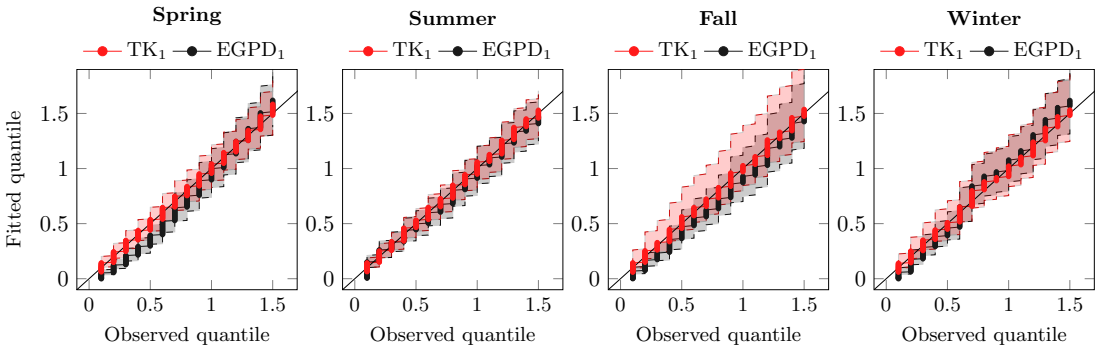

Figure 2.2.9: Zoom on the small values of the QQ-plots for the TK_1 and $EGPD_1$ models at Durance station, for each season

Remark 2.2.1. The analysis presented in the case studies of this chapter was performed in the R Software, more specifically we used for the fit of the $EGPD_1$ model the packages *mev*, particularly the *egp2* and *egp2.fit* functions.

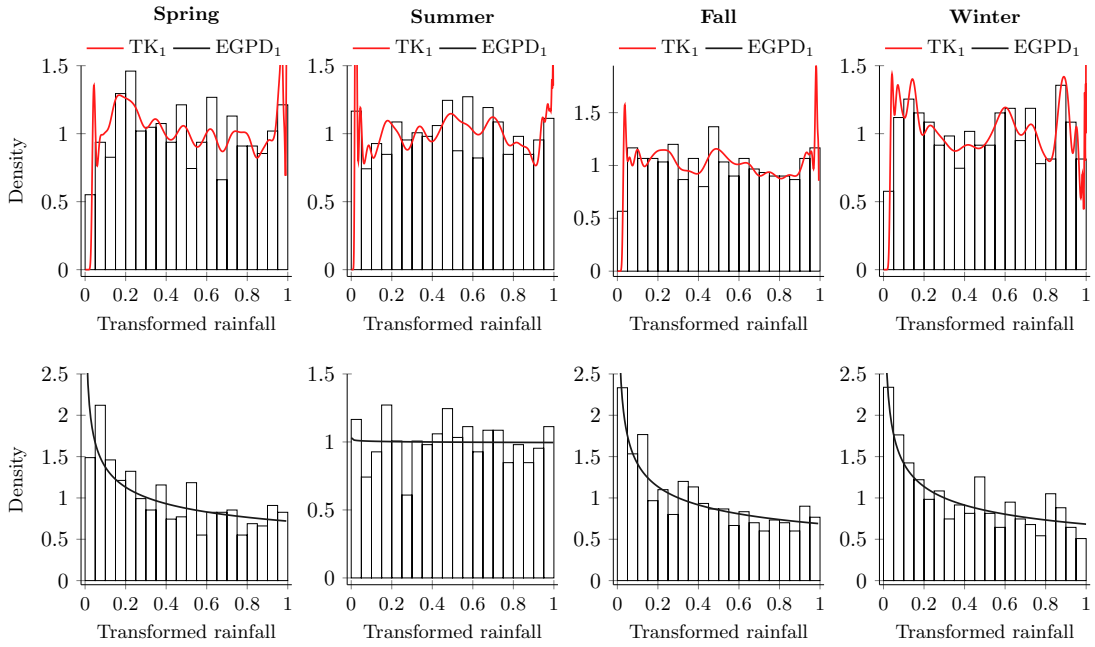

Figure 2.2.10: Histograms with fitted densities of the transformed data $U = F_{\mathrm{EGPD}_1}(X)$ and $V = H_\xi(X)$, for, respectively, TK_1 and EGPD_1 models at Durance station, for each season

Chapter 3

EGPD and sparse mixture models

This chapter presents a method for improving the fit of EGPD for the bulk of the distribution. We propose a semiparametric model in which flexibility is achieved by using the Bernstein polynomials. More specifically, we use a sparse mixture of beta densities. Section 3.1 gives a short background on Bernstein polynomials and its modeling assumptions. In Section 3.2 we show that the proposed model adds flexibility to the EGPD while preserving both lower and upper tails behavior. We conclude this section by presenting the estimation procedure for the parameters. Section 3.3 and Section 3.4 are dedicated to case studies on simulated data and on two rainfall datasets, respectively.

3.1 Bernstein polynomials and density estimation

Naveau et al. (2016) explained that the key component of the EGPD class is the continuous function G. This function is, in fact, the link between the two tails and also models the bulk of the distribution. As mentioned in their work, one single parametric form for G might be to stringent for the multimodal distributions, thus a more flexible family of models is needed.

Parametric density mixtures are a flexible and powerful tool for approximating smooth functions. While, when working with the whole $\mathbb{R}$ domain, researchers often use a mixture of gaussian densities, when a bounded interval, like $[0, 1]$ is needed, these gaussian mixtures yield a rather poor estimation due to the boundary effects at the end points of this interval, as pointed out by *Carreau and Bengio* (2008), for example. In this case, mixtures of beta densities, flexible 2-parameter models, can handle these boundary constraints, as shown by *Ji et al.* (2005).

Parametric mixture models can become problematic when the number of components is increasing, thus more parameters need to be estimated. For example, a mixture of beta densities with five components leads to the estimation of 14 parameters. It is well known that a large number of components increases the flexibility in the estimation, but can lead to over-fitting. Considering this, nonparametric models, such as the kernel density estimator, polynomials approximation, projections, can be preferred. They offer the same flexibility as a mixture model, but with no inconvenience regarding the parameters.

We want to consider in this study an in-between model, that is a semiparametric approach. In what follows, we explore how and why a polynomial approximation can be used to model G, thus leading to a new family of EGPD models. We also recall in the following that the class of Bernstein polynomials for density estimation is in fact a subclass of beta mixtures.

3.1.1 Background on Bernstein polynomials

Bernstein polynomials were introduced by *Bernstein* (1912) as a proof for Weierstrass Approximation Theorem. The author showed that any continuous function G on the interval $[0,1]$ can be approximated by Bernstein polynomials up to some degree of accuracy. For any $t \in [0,1]$, a Bernstein polynomial of degree m is defined by

$$P_m(t,G) = \sum_{k=0}^{m} G\left(\frac{k}{m}\right) b_{k,m}(t). \tag{2.3.1}$$

Here, $b_{k,m}$ is the Bernstein basis defined as

$$b_{k,m}(t) = \binom{m}{k} t^k (1-t)^{m-k}. \tag{2.3.2}$$

These bases have many attractive properties, such as: non-negativity, partition of unity, symmetry, etc. For a complete and detailed reference of these properties we refer the reader to *Farouki* (2012).

The many remarkable properties of the Bernstein basis can be viewed as compensation for the slow rate of convergence of $P_m(t,G)$ to G. As stated in Theorem 2.3.1, Bernstein polynomials $P_m(t,G)$ converges uniformly to G and the convergence rate is m^{-1}.

Theorem 2.3.1 (Bernstein polynomials convergence rate, *Voronovskaja* (1932))**.** *Given any continuous and differentiable function $G(t)$ on the interval $[0,1]$ with bounded second derivative, then for $m \to \infty$, there exists a polynomial $P_m(t,G)$, such that its approximation error is*

$$|G(t) - P_m(t,G)| = \frac{1}{2m}\left\{(1-2t)G^{'}(t) + t(1-t)G^{''}(t)\right\} + o\left(\frac{1}{m}\right).$$

3.1.2 From Bernstein polynomials to sparse mixture of beta densities

Despite its attractive basis properties, Bernstein polynomial approximation has remained for many years a theoretical approximation rather than a practical one, mainly due to its slow convergence rate. It started to gain notoriety after the work of *Vitale* (1975) was published. The author proposed an approach for density estimation based on the derivative of the Bernstein polynomials introduced in (2.3.1). More exactly, he considered that a density function $g(t)$ with $t \in [0,1]$, can be approximated by a Bernstein polynomial as displayed in (2.3.3).

$$\begin{aligned} p_{m-1,}(t,g) =& \frac{d}{dt} P_m(t,G) \\ =& \frac{d}{dt} P_m(t,G_n) \\ =& m \sum_{k=0}^{m-1} \left\{ G_n\left(\frac{k+1}{m}\right) - G_n\left(\frac{k}{m}\right) \right\} b_{k,m-1}(t) \end{aligned} \tag{2.3.3}$$

Here, the coefficients of each basis are based on the observations. More specifically, they are set as a function of the empirical cumulative distribution function (ecdf) $G_n(t)$. We recall that, for an i.i.d. random variable $(T_1, T_2, ..., T_n)$ on $[0,1]$ with a cdf $G(t)$, the ecdf is defined as $G_n(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{\{T_i \leq t\}}$, where $\mathbb{I}$ denotes the indicator function. By making use of the Bernstein basis properties and the ones of the ecdf (*i.e.*, $G_n(0) = 0$ and $G_n(1) = 1$), it can be easily checked that $p_{m-1}(t,g)$ is a probability density function, that is

i. $p_{m-1}(t,g) \geq 0$, for any $t \in [0,1]$,

ii. $\int_0^1 p_{m-1}(t,g)dt = 1$.

From now on, we call this estimator the *Bernstein density estimator*.

This estimator has been studied by other researchers as well. For example, *Babu et al.* (2002) investigate the asymptotic properties of the Bernstein density estimator to approximate the density and distribution functions. It was shown in this study that, for consistent convergence results, the degree m of the Bernstein polynomial should be included in the interval $2 \leq m \leq n/\log(n)$, where n is the sample size. *Leblanc* (2010, 2012a,b) studied the boundary properties of this estimator for both density and distribution functions. Generalizations of this estimator, including the Bayesian and multivariate extensions, were presented in *Bouezmarni and Rolin* (2007), *Ghosal* (2001), *Kakizawa* (2004), *Petrone* (1999).

Furthermore, as *Vitale* (1975) shortly mentioned in his work, the Bernstein density estimator $p_{m-1}(t,g)$, can be represented also as a linear combination of beta density bases. By simply changing the summation from (2.3.3) into $k = 1$ to m, it leads to $p_{m-1}(t,g) = m\sum_{k=1}^{m} \omega_{k,m} b_{k-1,m-1}(t)$, where $mb_{k-1,m-1}(t)$ is no more then a beta density basis. We call this estimator the *Bernstein-beta density estimator,* defined as

$$g_{BB,m}(t) = \sum_{k=1}^{m} \omega_{k,m}\beta_{k,m-k+1}(t). \tag{2.3.4}$$

Here $\omega_{k,m} = G_n\left(\frac{k}{m}\right) - G_n\left(\frac{k-1}{m}\right)$ and $\beta_{a,b}(t)$ denotes the beta density function with parameters a and b defined as

$$\beta_{a,b}(t) = \begin{cases} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} t^{a-1}(1-t)^{b-1}, & t \in [0,1] \\ 0 & \text{otherwise,} \end{cases} \tag{2.3.5}$$

where $\Gamma(\cdot)$ denotes the gamma function.

The estimator $g_{BB,m}(t)$ is a proper probability density, with the cumulative distribution function defined as

$$G_{BB,m}(t) = \sum_{k=1}^{m} \omega_{k,m} B_{k,m-k+1}(t), \tag{2.3.6}$$

where $B_{a,b}(x)$ is the beta distribution function with parameters a and b.

More generally, the Bernstein-beta density estimator belongs to the subclass of beta densities mixture, where both the weights and beta density parameters are pre-set, *i.e.*, considered as known. Thus, our motivation for choosing the Bernstein-beta density estimator over the mixture of beta comes especially from the significant decrease in the number of parameters, *i.e.*, the only parameter left to be estimated is the degree m.

This degree can be seen as a smoothing parameter and it is often compared with a kernel bandwidth problem, so similar estimation methods can be applied in this case, for example methods based on minimization of mean integrated squared error (MISE), as shown in *Silverman* (1986) or *Wand and Jones* (1995).

We direct our attention to the weights $\omega_{k,m}$ and explain why we believe this vector can be regarded as sparse. Considering the approximation given in (2.3.4), for a large degree m, most weights, *i.e.*, $\omega_{k,m} = G_n\left(\frac{k}{m}\right) - G_n\left(\frac{k-1}{m}\right)$, are almost null because there are no data points within the narrow intervals, and thus the ecdf is close to zero.

So, even though the approximation of the weights with the ecdf is an easy and appealing approach, having in mind its possible sparse representation for a large degree m, we want to consider the case when these weights are to be estimated and penalized to zero if they become too small.

To summarize, in what follows we apply the Bernstein-beta density estimator from (2.3.4), the weights are to be first estimated and then penalized to zero if they become too small. In order to maintain the properties of a probability density function, in the estimation step, we must recall the constraints $\omega_{k,m} \geq 0$, for any $k = 1, 2..., m$, and $\sum_{k=1}^{m} \omega_{k,m} = 1$.

3.1.3 Properties of the Bernstein density estimator

3.1.3.1 Asymptotic results

The asymptotic results of the Bernstein density estimator were first detailed in *Vitale* (1975), and later restated by *Babu et al.* (2002), *Leblanc* (2010) and *Leblanc* (2012b).

Assumption 2.3.1.

1. *Let U be a random variable with support $[0,1]$ and unknown density g.*
2. *Let g be a bounded and twice continuously differentiable function with bounded derivatives on $[0,1]$.*

In Theorem 2.3.2 we recall the bias, variance and the mean integrated square error (MISE) of the Bernstein density estimator, as derived by *Vitale* (1975).

Theorem 2.3.2 (*Vitale* (1975))**.** *Under Assumption (2.3.1) we have, for $t \in [0,1]$ and $m \to \infty$,*

a) $\mathbb{B}\text{ias}\left\{g_{BB,m}(t)\right\} = \frac{\Delta(t)}{m} + o\left(\frac{1}{m}\right)$

b) $\mathbb{V}\text{ar}\left\{g_{BB,m}(t)\right\} = \begin{cases} \frac{\sqrt{m}}{n} g(t)\psi(t) + o\left(\frac{\sqrt{m}}{n}\right), & \text{for } t \in (0,1) \\ \frac{m}{n} g(t) + o\left(\frac{m}{n}\right), & \text{for } t = 0 \text{ or } t = 1 \end{cases}$

c) $\text{MISE}\left\{g_{BB,m}(t)\right\} = \frac{\sqrt{m}C_1}{n} + \frac{C_2}{m^2} + o\left(\frac{\sqrt{m}}{n}\right) + o\left(\frac{1}{m^2}\right)$

where $\Delta(t) = \frac{1}{2}\left\{(1-2t)g'(t) + t(1-t)g''(t)\right\}$, $\psi(t) = \{4\pi t(1-t)\}^{-1/2}$, $C_1 = \int_0^1 g(t)\psi(t)dt$, *and* $C_2 = \int_0^1 \Delta^2(t)dt$.

It can be seen that, while the estimator has a uniform bias order of $o\left(\frac{1}{m}\right)$ over the whole unit interval, the variance behaves differently at the boundary points, *i.e.*, it has a greater order of magnitude, which in this case is $o\left(\frac{m}{n}\right)$ compared to the case of interval $(0,1)$ where the order is $o\left(\frac{\sqrt{m}}{n}\right)$.

Therefore, the asymptotically optimal choice for m, *i.e.*, given by the minimization of the MISE, is $m_{opt} = \left(\frac{4C_2}{C_1}\right)^{2/5} n^{2/5}$, such that

$$\text{MISE}\left\{g_{BB,m_{opt}}(t)\right\} = \frac{5}{4}\left(\frac{4C_1^4 C_2}{n^4}\right)^{1/5} + o\left(\frac{1}{n^{4/5}}\right).$$

We have mentioned in Section 3.1 that the degree of the Bernstein density estimator is often compared with the bandwidth of the kernel density estimator, in the way that, a higher degree (smaller bandwidth) gives more flexibility in smoothing a function. Going further with this analogy and setting $h = \frac{1}{m}$, we can see from Theorem 2.3.2 that the bias of $g_{BB,m}(t)$ is $o(h)$ as opposed to being $o\left(h^2\right)$ when using a kernel estimator (we refer for comparison to the boundary corrected kernel, see Section 2.2.1.1). On the other hand, looking at the variance in Theorem 2.3.2, it can be seen that, compared to the kernel estimator which has a asymptotic variance of $o\left(\frac{1}{nh}\right)$ and $o\left(\frac{1}{nh^2}\right)$ for the interior and boundary points, respectively, the variance of $g_{BB,m}(t)$ is smaller for the entire unit interval. Finally, the optimal bandwidth for $g_{BB,m}(t)$ is $h_{opt} \propto n^{-2/5}$ instead of $h_{opt} \propto n^{-1/5}$ in the kernel case.

3.1.3.2 Model identifiability

As mentioned in *MacLachlan and Peel* (2000), parameter identifiability is an important property in the statistical inference of a mixture model. The idea is that different parameters values in a model must correspond to different distributions. The identifiability condition becomes even more important when working with a maximum likelihood estimator, as it represents one of the constrains imposed by a consistent estimator.

We want to establish that our Bernstein-beta density estimator, *i.e.*, the mixture of beta densities, is identifiable. Thus, we relate our problem to the work of *Passow* (1977), where there is a discussion about the identifiability of polynomials with positive coefficients, which is exactly our case. In fact, it was mentioned shortly in the mentioned study that these polynomials are generalizations of Bernstein polynomials. The idea behind Passow's proof is that any polynomial with positive coefficients has always a polynomial of best approximation and this one is unique.

In Theorem 2.3.3 we recall this statement. Let $\Pi_m = \{p_m : p_m(t) = \sum_{i=0}^{m} c_i \cdot b_i(t), c_i \geq 0 \text{ for } i = 1, 2, ..., m\}$ be a set of polynomials with positive coefficients of degree m.

Theorem 2.3.3 (Existence and Uniqueness, *Passow* (1977))**.** *Let $g(t)$ be a continuous and positive function on $[0, 1]$. Then, for any fixed $m \in \mathbb{N}$, there exist $p_m^*(t) \in \Pi_m$ that minimizes $\|g(t) - p_m^*(t)\|_\infty$ among all $p_m(t) \in \Pi_m$, and it follows that the polynomial of best approximation $p_m^*(t) \in \Pi_m$ is unique.*

The proof of the theorem can be found either in the work of *Passow* (1977) or references regarding polynomials approximation, such as *Phillips* (2003).

In Theorem 2.3.4 we show the exact definition of identifiability in the case of the Bernstein-beta density estimator.

Theorem 2.3.4. *Let $\Pi_m = \{g_{BB,m} : g_{BB,m}(t; \omega) = \sum_{k=1}^{m} \omega_{k,m} \beta_{k,m-k+1}(t), \omega_{k,m} \geq 0 \text{ for } k = 1, 2, ..., m\}$ be a class of Bernstein-beta density estimators with degree m and weights vector ω. Given two families from Π_m,*

i) $g_{BB,m}(t; \omega) = \sum_{k=1}^{m} \omega_{k,m} \beta_{k,m-k+1}(t)$,

ii) $g_{BB,m}(t; \pi) = \sum_{k=1}^{m} \pi_{k,m} \beta_{k,m-k+1}(t)$,

for any $m \in \mathbb{N}$, $g_{BB,m}(t; \omega) = g_{BB,m}(t; \pi)$, if and only if $\omega = \pi$.

Proof. Suppose that $g_{BB,m}(t; \omega) = g_{BB,m}(t; \pi)$.

By simple algebra, we obtain

$$\sum_{k=1}^{m} (\omega_{k,m} - \pi_{k,m}) \beta_{k,m-k+1}(t) = 0 \Leftrightarrow \sum_{k=1}^{m} c_k \beta_{k,m-k+1}(t) = 0 \quad \text{(a)} \quad .$$

Next, we use the property that any Bernstein basis can be transformed to a power (monomial) basis (see *Farouki* (2012), Section 5), *i.e.*, $\beta_{k,m-k+1}(t) = k \sum_{i=k}^{m} (-1)^{i-k} \binom{m}{i} \binom{i}{k} t^{i-1}$.

If we further replace this relationship in (a), then we have

$$\sum_{k=1}^{m} c_k \beta_{k,m-k+1}(t) = 0 \Leftrightarrow \sum_{k=1}^{m} c_k k \sum_{i=k}^{m} (-1)^{i-k} \binom{m}{i} \binom{i}{k} t^{i-1} = 0.$$

Grouping the terms by monomial leads to

$$
\begin{aligned}
&c_1\binom{m}{1}\binom{1}{1} + \left\{-c_1\binom{m}{2}\binom{2}{1} + 2c_2\binom{m}{2}\binom{2}{2}\right\} t \\
&+ \left\{c_1\binom{m}{3}\binom{3}{1} - 2c_2\binom{m}{3}\binom{3}{2} + 3c_3\binom{m}{3}\binom{3}{3}\right\} t^2 \\
&+ \ldots + \left\{\sum_{i=1}^{m}(-1)^{i-1} i c_i \binom{m}{m}\binom{m}{i}\right\} t^{m-1} = 0.
\end{aligned}
$$

Since the power basis is a linearly independent set, we have that

$$
\begin{cases}
c_1\binom{m}{1}\binom{1}{1} = 0 \\
-c_1\binom{m}{2}\binom{2}{1} + 2c_2\binom{m}{2}\binom{2}{2} = 0 \\
\ldots \\
\sum_{i=1}^{m}(-1)^{i-1} i c_i \binom{m}{m}\binom{m}{i} = 0.
\end{cases}
$$

which implies that $c_1 = c_2 = \ldots = c_m = 0$. But $c_k = \omega_{k,m} - \pi_{k,m}$, so this leads to $\omega_{k,m} = \pi_{k,m}$, for $k = 1, 2, \ldots, m$. ■

3.2 EGPD with Bernstein-beta density

We introduced in Section 3.1 the Bernstein-beta density estimator and we have seen its asymptotic properties. In this section we are concerned with the inference of the EGPD model based on the Bernstein-beta density. Thus, we first review the constraints imposed by the EGPD class and prove that they are satisfied when using a Bernstein-beta estimator. Then, we present the method used for the estimation of the parameters.

3.2.1 Constraints imposed by the EGPD class

Taking into consideration the Bernstein-beta pdf and cdf introduced in (2.3.4) and (2.3.6), given an i.i.d. random variable X, we obtain the cdf and pdf of the EGPD model based on Bernstein-beta density (EGPD-BB), as

$$
\begin{aligned}
F(x) &= G_{BB,m}\left\{H_\xi(x)\right\}, \\
f(x) &= g_{BB,m}\left\{H_\xi(x)\right\} h_\xi(x),
\end{aligned}
\tag{2.3.7}
$$

where H_ξ and h_ξ denotes the cdf and pdf of the GPD.

According to *Naveau et al.* (2016), an EGPD class assumes that the lower tail of the distribution behaves like a power function x^s, while the upper tail like a GPD. Thus, we want to check if $\lim_{x\to\infty} \frac{\bar{F}(x)}{\bar{H}_\xi(x)}$ and $\lim_{x\to 0} \frac{F(x)}{x^s}$ hold, *i.e.*, they are finite and positive. In Lemma 2.3.1 we show the behavior of the two tails in the context of the EGPD-BB model.

Lemma 2.3.1. *Given an i.i.d. random variable X, the lower and upper tail behavior of an EGPD model based on Bernstein-beta density as formulated in (2.3.7), is:*

1. $\lim_{x\to 0} \frac{F(x)}{x^s} = c\omega_{s,m}$, *where c is a constant, and s denotes the position of the first non-null weight in ω.*

2. $\lim_{x\to\infty} \frac{\bar{F}(x)}{\bar{H}_\xi(x)} = m\omega_{m,m}$

Proof.

1. $\lim_{x\to 0}\frac{F(x)}{x^s} = c\omega_{s,m}$, where c is a constant, and s denotes the position of the first non-null weight in ω.

   Starting from

   $$\lim_{x\to 0}\frac{F(x)}{x^s} = \lim_{x\to 0}\frac{G_{BB,m}\left\{H_\xi(x)\right\}}{x^s},$$

   and following the same reasoning as *Naveau et al.* (2016), we get

   $$\begin{aligned}\lim_{x\to 0}\frac{G_{BB,m}\left\{H_\xi(x)\right\}}{x^s} &= \lim_{x\to 0}\frac{G_{BB,m}\left\{x\cdot\frac{H_\xi(x)}{x}\right\}}{G_{BB,m}(x)}\frac{G_{BB,m}(x)}{x^s},\\ &= \lim_{x\to 0}\frac{G_{BB,m}\left\{x\cdot\omega(x)\right\}}{G_{BB,m}(x)}\frac{G_{BB,m}(x)}{x^s},\\ &= \lim_{x\to 0}\frac{G_{BB,m}\left\{x\cdot\omega(x)\right\}}{G_{BB,m}(x)}\lim_{x\to 0}\frac{G_{BB,m}(x)}{x^s}.\end{aligned}$$

   We must say that, as long as $\omega(x) = \frac{H_\xi(x)}{x}$, the term $\frac{G_{BB,m}\{x\cdot\omega(x)\}}{G_{BB,m}(x)}$ is always constant and non-null (*i.e.*, equal to 1) as $x\to 0$ due to the fact that $\omega(x) = 1 + o(x)$, as $x\to 0$ (Taylor expansion).

   That leaves the computation of the second term in the limit above, *i.e.*,

   $$\lim_{x\to 0}\frac{G_{BB,m}(x)}{x^s} = \lim_{x\to 0}\frac{g_{BB,m}(x)}{sx^{s-1}}.$$

   Considering that $g_{BB,m}(0) = m\omega_{1,m}$, the above equality holds (*i.e.*, l'Hopital rule can be applied) only if $g_{BB,m}(0) = m\omega_{1,m} = 0$ or $s = 1$. Therefore, in order to have a finite and positive limit, we have to force the initial weight(s) to be zero. The number of necessary initial null weights is given by the choice of the power s from the denominator.

   To be more clear, we take an example.

   (a) for $s = 1$, we must force the condition $\omega_{1,m} > 0$
   $\lim_{x\to 0}\frac{G_{BB,m}(x)}{x} = \lim_{x\to 0} g_{BB,m}(x) = m\omega_{1,m}$

   (b) for $s = 2$, we must force the condition $\omega_{1,m} = 0$ and $\omega_{2,m} > 0$
   $\lim_{x\to 0}\frac{G_{BB,m}(x)}{x^2} = \lim_{x\to 0}\frac{g_{BB,m}(x)}{2x} = \lim_{x\to 0}\frac{g'_{BB,m}(x)}{2} = \omega_{2,m}m(m-1)(m-2)$
   where $g'_{BB,m}$ is the derivative of $g_{BB,m}$.

   (c) from simple induction that can easily be checked, for $s = k$, we must thus force the condition $\omega_{i,m} = 0$, for $i = 1, ..., k-1$, and $\omega_{k,m} > 0$.

   Consequently, the lower tail behavior is imposed by the model by considering constraints on the initial weights.

2. $\lim_{x\to\infty}\frac{\bar{F}(x)}{\bar{H}_\xi(x)} = m\omega_{m,m}$

   Starting from

   $$\lim_{x\to\infty}\frac{\bar{F}(x)}{\bar{H}_\xi(x)} = \lim_{x\to\infty}\frac{\bar{G}_{BB,m}\left\{H_\xi(x)\right\}}{\bar{H}_\xi(x)},$$

and then, applying the change of variable $\bar{H}_\xi(x) = v$, we have

$$\lim_{x\to\infty} \frac{\bar{G}_{BB,m}\{H_\xi(x)\}}{\bar{H}_\xi(x)} = \lim_{v\to 0} \frac{\bar{G}_{BB,m}(1-v)}{v}.$$

Next, by applying l'Hopital's rule gives,

$$\begin{aligned}\lim_{v\to 0} \frac{\bar{G}_{BB,m}(1-v)}{v} &\overset{\frac{0}{0}}{=} \lim_{v\to 0} g_{BB,m}(1-v)\\ &= g_{BB,m}(1)\\ &= m\omega_{m,m}.\end{aligned}$$

Therefore, in order to satisfy the constraint regarding the upper tail behavior, we have to force the last weight to be non-null, *i.e.*, $\omega_{m,m} > 0$.

■

Remark 2.3.1. We will see in the case study section that, in practice, the degree of the power function (*i.e.*, s) is given by the model choice, and it is not necessary to constraint the estimation approach. More specifically, due to the fact that we consider a sparse mixture model, some of the initial weights will become null, and will drive the behavior of the lower tail.

3.2.2 Parameter estimation

As we have seen in Section 3.2.1, the parameters of our model are the weights ω of the Bernstein-beta estimator and the two GPD parameters, σ and ξ. Beside this, we must also set the degree m of the polynomial (*i.e.*, the number of components in the mixture), which from now on will be called the *hyperparameter*.

In Section 3.2.2.1 and Section 3.2.2.2 we discuss the methodology for estimating the parameters ω, σ, ξ for a fixed m, namely the maximum likelihood estimator. Following that, in Section 3.2.2.3 we address the subject of how to set the hyperparameter m. The last part of this section, that is Section 3.2.2.4, is dedicated to a discussion regarding an alternative estimation method.

3.2.2.1 Maximum likelihood estimator

The parameters of our model are estimated using the classical maximum likelihood estimator. The likelihood function can be easily obtained from (2.3.4) as

$$\mathrm{L}(\theta|x) = \prod_{i=1}^{n} f(x_i|\theta) = \prod_{i=1}^{n} g_{BB,m}\{H_\xi(x_i)\}\, h_\xi(x_i), \tag{2.3.8}$$

where $\theta = [\omega, \sigma, \xi]$ denotes the vector of parameters. The corresponding log-likelihood function is given in (2.3.9).

$$\begin{aligned}\log\mathrm{L}(\theta|x) &= \sum_{i=1}^{n} \log\{f(x_i|\theta)\}\\ &= \sum_{i=1}^{n} \log\left[g_{BB,m}\{H_\xi(x_i)\}\right] + \sum_{i=1}^{n} \log\{h_\xi(x_i)\}\end{aligned} \tag{2.3.9}$$

To obtain the estimated parameters $\hat{\theta}^{\mathrm{MLE}}$ we have to minimize the negative log-likelihood (or to maximize the log-likelihood). Moreover, we must also have in mind that some parameters

must be constrained, either by model formulation (*e.g.*, scale parameter in GPD must be finite positive, the weights must be positive) or induced in order to satisfy certain conditions (*e.g.*, constraints on weights, see Section 3.2.1).

Therefore, our optimization problem becomes

$$\begin{aligned} \hat{\theta}^{\text{MLE}} = \underset{\theta}{\text{argmin}} \quad & \{-\text{logL}(\theta|x)\} \\ \text{subject to} \quad & \omega \in \Delta^m \\ & \omega_{m,m} > 0 \\ & \sigma > 0 \\ & \xi \geq 0, \end{aligned} \tag{2.3.10}$$

where $\Delta^m = \{\omega \in \mathbb{R}^m : \omega_{k,m} \geq 0, \sum_{k=1}^m \omega_{k,m} = 1\}$ are called *simplex constraints*.

We cannot obtain an analytic form solution from (2.3.10) as the model involves many parameters and it is highly non-linear. Thus, the optimal solution $\hat{\theta}^{MLE}$ is computed numerically through non-linear optimization algorithms, detailed in Section 3.3.

3.2.2.2 Penalized maximum likelihood estimator

We have seen in Section 3.1.2 that, for a sufficiently high degree m of Bernstein-beta estimator, the solution of the mixture is sparse, *i.e.*, only a small number of the weights are different than zero. Therefore, we consider a standard approach for promoting sparsity that is widely used in literature, namely the l_1-regularization. Generally speaking, instead of optimizing the objective function $J(x)$, we now optimize $J(x) + \lambda||x||_1$ (or $J(x)$, s.t. $||x||_1 \leq \gamma$), where $||\cdot||_1$ is the l_1-norm or the sum of absolute values. The larger the penalty applied, *i.e.*, λ , the more the estimates are shrunk towards zero. This penalization criterion was first introduced by *Tibshirani* (1996) under the name Lasso criterion, and it is often used in linear regression problems for subset selection.

Returning now to our problem, we want to penalize the small values of the weights and eventually to set them to zero. Therefore, considering the formulation of the Lasso problem mentioned above, the penalized form of the Bernstein-beta log-likelihood is

$$\begin{aligned} \hat{\theta}^{\text{MLE}_{l1}} = \underset{\theta}{\text{argmin}} \quad & \{-\text{logL}(\theta|x)\} \\ \text{subject to} \quad & ||\omega||_1 \leq \gamma \\ & \omega \in \Delta^m \\ & \omega_{m,m} > 0 \\ & \sigma > 0 \\ & \xi \geq 0. \end{aligned} \tag{2.3.11}$$

However, standard l_1-regularization, *i.e.*, $||\omega||_1 \leq \gamma$ or $\sum_{k=1}^m |\omega_{k,m}| \leq \gamma$, becomes ineffective in the presence of the already necessary simplex constraints, *i.e.*, $\sum_{k=1}^m \omega_{k,m} = 1$ and $\omega_{k,m} \geq 0$ (see the constraints in red from (2.3.11)). We can notice that in fact the solution yielded by using the simplex constraints is part of the l_1-regularization set, but we cannot be sure that this solution offers the best possible sparsity pattern.

There are approaches reported in the literature that study the sparse optimization under simplex constraints. We mention in this sense the work of *Pilanci et al.* (2012) where the authors consider a regularization of the objective function under simplex constraint with the inverse l_∞-norm, or the work of *Li et al.* (2016) where a regularization based on negative l_2-norm is studied.

Aside from these approaches, the sparsity under simplex constraints can be also obtained by using a pragmatic 2-step estimation. One approach could be to, first i) ignore the simplex constraints and find the optimal solution on $\mathbb{R}^m$, then, ii) project this solution on the simplex. On the same line, another method is to i) find the optimal solution under the simplex constraints, and then, ii) further shrink it by using a suitable method. This method is called in the literature *thresholding*. In the work of *Li et al.* (2016) all the above mentioned methods, and others, are compared, and the general conclusion was that naive approaches like thresholding can provide rather effective and similar results when compared with more sophisticated methods.

Driven by these results and the fact that it is computationally faster, we will use the thresholding approach in our study. Thus, the optimization problem becomes

$$\begin{aligned} \text{Step 1:} \qquad \hat{\theta}^{\text{MLE}} = \underset{\theta}{\text{argmin}} \quad & \{-\text{logL}(\theta|x)\} \\ \text{subject to} \quad & \omega \in \Delta^m \\ & \omega_{m,m} > 0 \\ & \sigma > 0 \\ & \xi \geq 0, \end{aligned} \tag{2.3.12a}$$

$$\text{Step 2:} \qquad \hat{\omega}^{\text{MLE}}(\tau) = \left\{\hat{\omega}_{i,m}^{\text{MLE}} \cdot \mathbb{I}\left(\hat{\omega}_{i,m}^{\text{MLE}} \geq \tau\right)\right\}_{1 \leq i \leq m-1} \tag{2.3.12b}$$

where $\mathbb{I}(\cdot)$ is the indicator function and τ is the threshold value.

Remark 2.3.2. Attention must be paid to the last weight $\omega_{m,m}$ that must be kept in the model, according to the constraints from Section 3.2.1. Thus, we are not including $\omega_{m,m}$ in the thresholding step and leave it unchanged, even if its value is very small.

For the selection of the threshold τ from (2.3.12b) we use the approach suggested by *Li et al.* (2016), that is:

i. compute the sets $\Omega = \left\{\hat{\omega}^{\text{MLE}}(\tau_i), \tau_i \in T\right\}$,
   where $T = \left\{\hat{\omega}_{i,m}^{\text{MLE}}, i \in \mathbb{I}\left(\hat{\omega}_{i,m}^{\text{MLE}} \neq 0\right)\right\}$.

ii. pick the best threshold $\tau_i \in T$ according to a model selection criterion, such as generalized information criterion (GIC) presented in *Kim et al.* (2012) and defined by

$$\text{GIC}(\tau_i) = -\text{logL}\left\{\hat{\omega}^{\text{MLE}}(\tau_i)\middle| x, \hat{\omega}_{m,m,}^{\text{MLE}}, \hat{\sigma}^{\text{MLE}}, \hat{\xi}^{\text{MLE}}\right\} + \lambda\sigma_\epsilon^2(\tau_i)M(\tau_i) \tag{2.3.13}$$

   Here, $\lambda = 2\log(m)$, $M(\tau_i) = \sum_{k=1}^{m} \mathbb{I}\left(\hat{\omega}_{k,m}^{MLE-\tau_i} \neq 0\right)$ and $\sigma_\epsilon^2(\tau_i)$ is the variance of the error term (*i.e.*, deviation from the true model). As this value is unknown, as suggested by *Kim et al.* (2012), we replace $\sigma_\epsilon^2(\tau_i)$ with the mean squared error obtained by using the initial non-thresholded model from (2.3.12a), as a reference.

   Given the set of candidate models Ω, the preferred model is the one with the minimum GIC value.

Remark 2.3.3. The class of GIC includes many well known selection criteria, such as AIC or BIC that correspond to $\lambda = 2$ and $\lambda = 2\log(n)$, respectively.

After Step 2 from (2.3.12b), the weights will not unit sum, thus, our estimator will not integrate to one (condition required by a probability density function). Moreover, Step 2 does not take into consideration the GPD parameters (σ, ξ) and the last weight $\omega_{m,m}$, so even though changes might appear in ω, the three parameters are unchanged. Consequently, we need an extra step, shown in (2.3.14), that deals with these two issues: 1) ensure the unit sum of the weights, and 2) reestimate accordingly the remaining three parameters. Basically, we optimize again our problem, but this time we take into consideration the sparse log-likehood, *i.e.*, $\text{logL}(\theta_{J(\tau)}|x)$, where $\theta_{J(\tau)} = [\omega_{J(\tau)}, \sigma, \xi]$ and $J(\tau) = \{j \in \{1, ..., m\} : \omega_j \neq 0\}$.

$$\begin{aligned} \text{Step 2a:} \qquad \hat{\theta}^{\text{MLE}}_{J(\tau)} = \underset{\theta_\tau}{\text{argmin}} \quad & \left\{-\text{logL}(\theta_{J(\tau)}|x)\right\} \\ \text{subject to} \quad & \omega_{J(\tau)} \in \Delta^m_{J(\tau)} \\ & \omega_{m,m} > 0 \\ & \sigma > 0 \\ & \xi \geq 0, \end{aligned} \tag{2.3.14}$$

where $\Delta^m_{J(\tau)} = \{\omega \in \mathbb{R}^m : \omega_{k,m} \geq 0, \sum_{k \in J(\tau)} \omega_{k,m} = 1\}$.

Remark 2.3.4. In practice, the optimization problem from (2.3.12a) might be very hard to solve, and solutions that are in fact local minima might be returned. Therefore, to improve the estimates, we propose to split the problem in two subproblems: 1) fix the GPD parameters and, thus, reduce the problem just to the optimization of the first sum in (2.3.9), and 2) with the known sparsity (*i.e.*, non-null weights), optimize the sparse log-likehood (see (2.3.14)) and find the estimates for all parameters. We will see in the case study below how exactly to choose the GPD parameters required by the first subproblem.

3.2.2.3 Selection of the optimal degree

We have seen in the first part of this chapter that the degree m is a very important feature of the model as it gives the smoothness of the estimator. *Babu et al.* (2002) show that m should be of order $o\{n/\log(n)\}$ for consistent convergence results, where n is the data sample size. Also, it was concluded in the numerical study of the mentioned work that $m = \frac{n}{\log(n)}$ works well. However, their study covered only small sample sizes (up to 125 observations), so when working with larger samples, the degree $m = \frac{n}{\log(n)}$ could be too large for practical purposes.

Another interesting approach for the choice of the optimal degree is given by *Guan* (2016), who used a change-point detection method of the profile log-likelihood on a set of candidate values m_i. Their reasoning is based on the fact that the change in log-likelihood is always positive as m increases. However, this approach is not suitable if a sparse solution is considered, as the change in log-likelihood will not be necessary positive as m increases.

Having in mind the close relationship with the nonparametric kernel estimator, an immediate thought will be to use the same approaches for the choice of the hyperparameter m as for the bandwidth. Thus, one traditional method is the Least Square Cross Validation (LSCV) which is based on the minimization of the MISE. This method was used before by *Bouezmarni and Rolin* (2007), *Kakizawa* (2004), *Leblanc* (2010) in Bernstein density estimation problems.

The idea of LSCV is to construct an estimator for the MISE from the data and, after that, to minimize it over multiple values of m in order to get the optimal degree. So, the LSCV estimator and its optimal value are given by

$$\begin{aligned} \text{LSCV}(m) &= \textstyle\int_0^1 g^2_{BB,m}(t)dt - \frac{2}{n}\sum_{i=1}^m g^{(-i)}_{BB,m}(t_i), \\ m_{\text{LSCV}} &= \underset{m}{\arg\min}\left\{\text{LSCV}(m)\right\}. \end{aligned} \tag{2.3.15}$$

Here, $g^{(-i)}_{BB,m}(t)$ denotes the Bernstein-beta estimator constructed from all the data except t_i.

3.2.2.4 Discussion on parameter estimation

We have seen in this section how to estimate the parameters of our EGPD model by optimizing a penalized maximum likelihood. We concluded that the numerical algorithm might face difficulties due to the several constraints that had to be imposed, and, thus, we proposed to split the problem in two subproblems: first optimize the likelihood of the mixture by fixing the GPD parameters, and then, by using the sparse likelihood, reestimate all the parameters at once.

In this subsection we want to discuss another approach for estimating sparse solutions from a high dimensional mixture model, that is the first subproblem mentioned above, by means of a penalized empirical risk loss function. This method was introduced by *Bunea et al.* (2010) and we briefly mentioned it in the introduction in Section 1.1. In this work, the authors considered estimating a density function in $\mathbb{R}$ domain by a sparse linear combination of basis functions, which in this case, are gaussian densities with known means and variances. The sparse solution is obtained by a l_1-penalization of the empirical risk (ER), *i.e.*, $\text{ER} = \|f\|_2^2 - \frac{2}{n}\sum_{i=1}^n f(x_i)$.

This estimator can recover with high probability the true weights only under a local coherence regarding the basis functions, defined as

$$\rho(\omega) = \max_{i \in M(\omega)} \max_{j \neq i} |\rho_f(i,j)|, \tag{2.3.16}$$

where $\rho_f(i,j) = \frac{<f_i, f_j>}{\|f_i\|\|f_j\|}$, f_i are the basis function and $M(\omega)$ is the number of non-null weights ω. Briefly, it means that for proper weight identification, the dictionary (*i.e.*, the set of basis functions) must not be over-complete, that is they must be clearly separated.

To anticipate what follows, we want to draw attention to an interesting conclusion of the simulation study of the mentioned work. It was shown that, even though this condition is not satisfied, *i.e.*, the dictionary is over-complete and the hit rate (percentage of times the cardinality of the true weights vector is correctly identified) is zero, the density approximation error is not changing crucially compared with the case when the condition is satisfied and the hit rate is 100%.

Driven by the good theoretical properties of this estimator, we consider it for our sparse estimation. However, this estimator cannot be applied directly to our case as our density function is on the unit interval and not in the $\mathbb{R}$ domain, thus another basis must be used. In fact, the change of basis is not that straightforward as no density functions on $[0,1]$ satisfy the condition required in *Bunea et al.* (2010). For example, if a beta density with parameters (a,b) is considered as basis function, then $\rho(\omega) = 1$ as $\rho_f(i,j)$ shows perfect correlation.

Even so, in view of the conclusion mentioned above and adding the fact that we are more interested in the density approximation error rather than in a perfect hit rate, we pursued the weights estimation based on l_1-penalization of the empirical risk, when a beta density is the basis. Then, we compared this output with the one yielded by the maximum likelihood. Unfortunately, the approach from *Bunea et al.* (2010) shows a poor estimation compared to MLE. We show in Appendix B this comparison for one applied study.

In conclusion, the estimator proposed by *Bunea et al.* (2010) does not seem that straightforward when we move to the unit interval, so in what follows we will not pursue this approach. An option, though, could be to step away from dictionaries composed of density functions, and consider orthogonal basis on $[0,1]$, such as wavelets, orthogonal polynomials, etc. This approach, however, brings an extra step to the estimation which requires to project the solution on the positive space such that it integrates to 1 (density requirement).

3.3 Case study: simulated data

In this section we study the performance of Bernstein-beta EGPD (EGPD-BB) on simulated data. First, we conduct an analysis when the hyperparameter m is considered known, *i.e.*, it is set to its true value, and, thus, focus on the assessment of the remaining parameters. Then, we consider the case when m is unknown and it is set to the optimal value suggested by *Babu et al.* (2002), *i.e.*, $m = n/\log(n)$.

3.3.1 Simulation study with the hyperparameter $m = m_{\text{true}}$

We consider simulations from our model with the following setting:

- $m_{\text{true}} = 50$, considered known,
- $\omega_{\text{true}} = [\omega_{15} = 0.25, \omega_{25} = 0.2, \omega_{35} = 0.25, \omega_{45} = 0.25, \omega_{50} = 0.05]$, the remaining 45 elements of ω_{true} vector out of the 50 are zero; this means that we have only $M = 5$ non-null weights,
- $\sigma_{\text{true}} = 2$ and $\xi_{\text{true}} = 0.15$.

The density of a model with the above settings is illustrated in Figure 2.3.1. We consider simulations of three different sizes, *i.e.*, $n = 300$, 1000, 2000, and for each of them we generate 50 samples from this multimodal model.

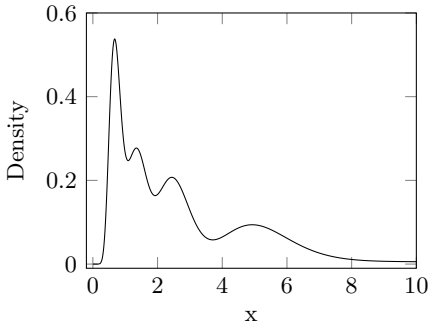

Figure 2.3.1: Shape of the density used in this simulation study, with $m = 50$, $\omega = [\omega_{15} = 0.25, \omega_{25} = 0.2, \omega_{35} = 0.25, \omega_{45} = 0.25, \omega_{50} = 0.05]$, $\sigma = 2$ and $\xi = 0.15$

We pursue the following inference methods for the EGPD-BB model:

1. Optimize the entire problem at once, as presented in Section 3.2.2.2 (E1)
2. Estimate by splitting the initial problem in two subproblems, where GPD parameters are fixed as follows:

   E2_{fix}: to their true values ($\sigma = \sigma_{\text{true}}$ and $\xi = \xi_{\text{true}}$),

   $\text{E2}_{\text{EGPD}_1}$: to their estimated values from the EGPD_1 model, *i.e.*, $G(u) = u^k$, see (2.1.16a) ($\sigma = \hat{\sigma}_{\text{EGP}D_1}$ and $\xi = \hat{\xi}_{\text{EGP}D_1}$),

   E2_{grid}: take a grid for each pair (σ, ξ) and choose the optimal one by minimizing BIC ($\sigma = \hat{\sigma}_{\text{grid}}$ and $\xi = \hat{\xi}_{\text{grid}}$). We consider a grid of 100 equally spaced points for each parameter, *i.e.*, $\sigma_{\text{grid}} = \{0.05, 0.1, ..., 5\}$ and $\xi_{\text{grid}} = \{0.01, 0.02, ..., 1\}$.

The notation for these methods means: "E"=Evaluation, "1/2"= entire problem/two subproblems optimization, and the subscripts "fix", "EGPD_1", or "grid" indicate the approach used for setting the GPD parameters. For clarity, we provide in Appendix A the estimation methodology for all the approaches used in this study.

The unrealistic situation E2_{fix} from point 2 above, where the GPD parameters are considered known, was included in order to highlight how the inclusion of the GPD parameters influences the estimation of the weights.

3.3.1.1 Parameter estimation

In this section we analyze the estimated parameters over the 50 replicates. We are particularly interested in the ability of EGPD-BB to identify the values of the true mixture weights and GPD parameters. Figure 2.3.2 and Figure 2.3.3 illustrate the estimated weights of the mixture and the boxplots of the GPD parameters, respectively, for each estimation approach. Table 2.3.1 summarizes the hit rate of the true mixture cardinality (*i.e.*, the percentage of times the true number of non-null weights M is equal to the estimated one $\hat{M}$, over the 50 replicates).

As presumed in the theoretical section, there are important differences in estimation while comparing the case when all the parameters are estimated at once (*i.e.*, E1) with the case when a division in subproblems is considered (*i.e.*, E2).

Table 2.3.1: Percentage of time $\hat{M} = M$ and the mean of $\hat{M}$ obtained from 50 replicates, for the sample sizes $n = 300, 1000, 2000$ and the estimation approaches E1, E2_{fix}, $\text{E2}_{\text{EGPD}_1}$ and E2_{grid} (M - true cardinality (*i.e.*, 5); $\hat{M}$ - estimated cardinality; $\bar{M}$ - mean of the estimated cardinality from 50 replicates)

	n=300		**n=1000**		**n=2000**	
	%	$\bar{M}$	%	$\bar{M}$	%	$\bar{M}$
$\mathbf{E2_{fix}}$	88%	4.88	96%	4.96	100%	5.00
$\mathbf{E2_{EGPD_1}}$	84%	4.84	88%	4.88	86%	4.86
$\mathbf{E2_{grid}}$	98%	4.98	100%	5.00	100%	5.00
E1	90%	4.90	94%	4.94	90%	4.90

As illustrated in Figure 2.3.2, the estimation approach E1 reports an overestimation of the scale parameter σ, and a significant instability in the estimation of the shape parameter ξ, regardless the sample size. Besides this, even though the hit rate is larger than 90%, as shown in Table 2.3.1, the positions of the estimated weights are not the true ones, as it can be seen from the last line of plots in Figure 2.3.3. More specifically, even though the estimated cardinality is equal to the true one (*i.e.,* 5) in most of the cases out of the 50 replicates, the positions of the estimated weights do not maintain the true location of the components. In fact, it can be noticed from the last line of plots in Figure 2.3.3, that the positions of the non-null estimated weights (in gray) are shifted to the left compared to the true ones (in blue), except the last weight ($\hat{\omega}_{m,m}$) which is estimated each time quite well, due to the constraint considered in the optimization problem, *i.e.*, $\omega_{m,m} > 0$.

The split in subproblems is improving both the estimation of the GPD parameters, and the identification and estimation of the weights. We focus, at first, on the situations when the GPD parameters are considered unknown, *i.e.,* estimation approaches $\text{E2}_{\text{EGPD}_1}$ and E2_{grid}. It can be observed from Table 2.3.1 that $\text{E2}_{\text{EGPD}_1}$ tends to underestimate the number of non-null weights, while E2_{grid} has a very good cardinality identification, achieving a 100% hit rate for large sample size datasets (*i.e.*, $n = 1000$ and $n = 2000$). Moreover, the estimation of the weights is improved compared to the E1 estimation case, especially when E2_{grid} is applied. More specifically, as illustrated in the second and third lines of plots from Figure 2.3.3, the five clusters formed by the weights' estimates are better defined, such that the estimated values of the weights

are not so scattered, and they are concentrated more closely to the true value of the weights parameters, especially in the $E2_{grid}$ case. Furthermore, the estimation of the GPD parameters, as illustrated in Figure 2.3.2, is considerably improved and more stable than the E1 case. While $E2_{EGPD_1}$ seems to still under- and overestimate the scale and shape parameters, respectively, $E2_{grid}$ achieves good estimation results. The identification and estimation performance of all parameters increases as the sample size becomes larger.

We have seen so far that when considering both GPD parameters and mixture weights to be evaluated at once, the estimations are poor, but they are improved when the optimization problem is split in two subproblems. It is also interesting to assess how the inclusion of GPD parameters influences the estimation of the weights. Thus, we analyze the estimation approach $E2_{fix}$, where the GPD parameters are considered known. As illustrated in the first line of plots from Figure 2.3.3, the positions of the estimated weights are very well identified, and the clusters formed by these estimates are very narrow, meaning that the estimated weights are in most of the cases on the true position. These results lead to the conclusion that the mixture weights can be identified and estimated with high accuracy, if the GPD parameters are well assessed.

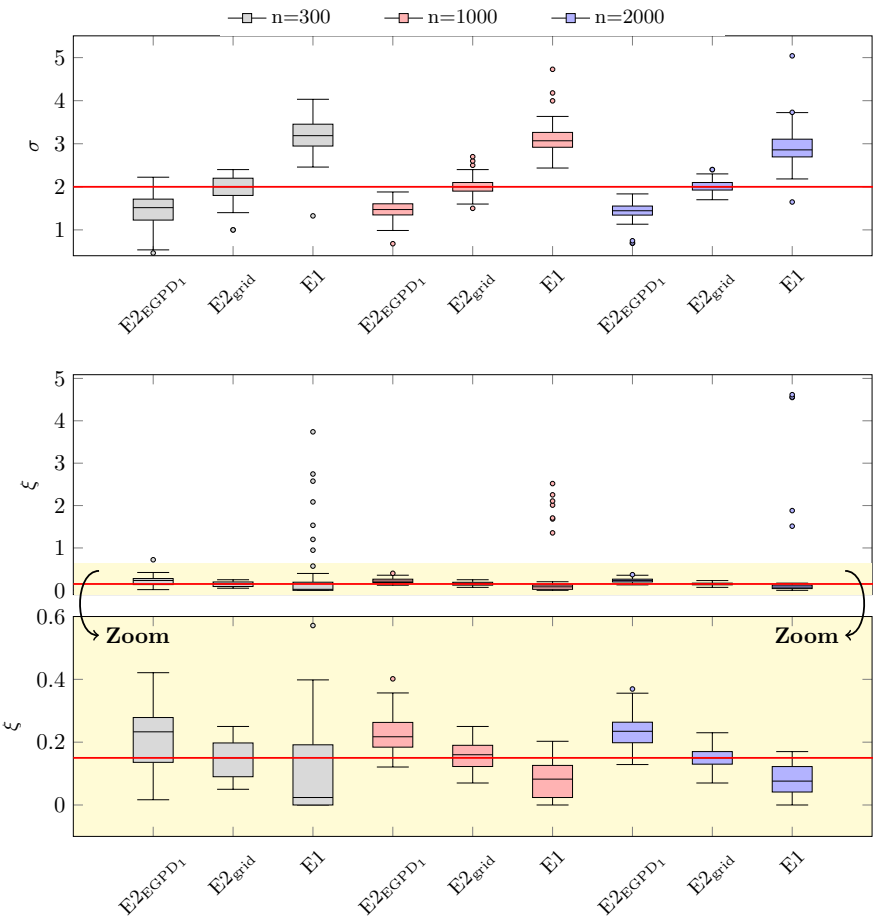

Figure 2.3.2: Boxplots of the estimated scale (top) and shape (bottom) GPD parameters for the 50 replicates, for the sample sizes $n = 300, 1000, 2000$ and the estimation approaches E1, $E2_{EGPD_1}$ and $E2_{grid}$ (the red horizontal line indicates the true value of the parameters)

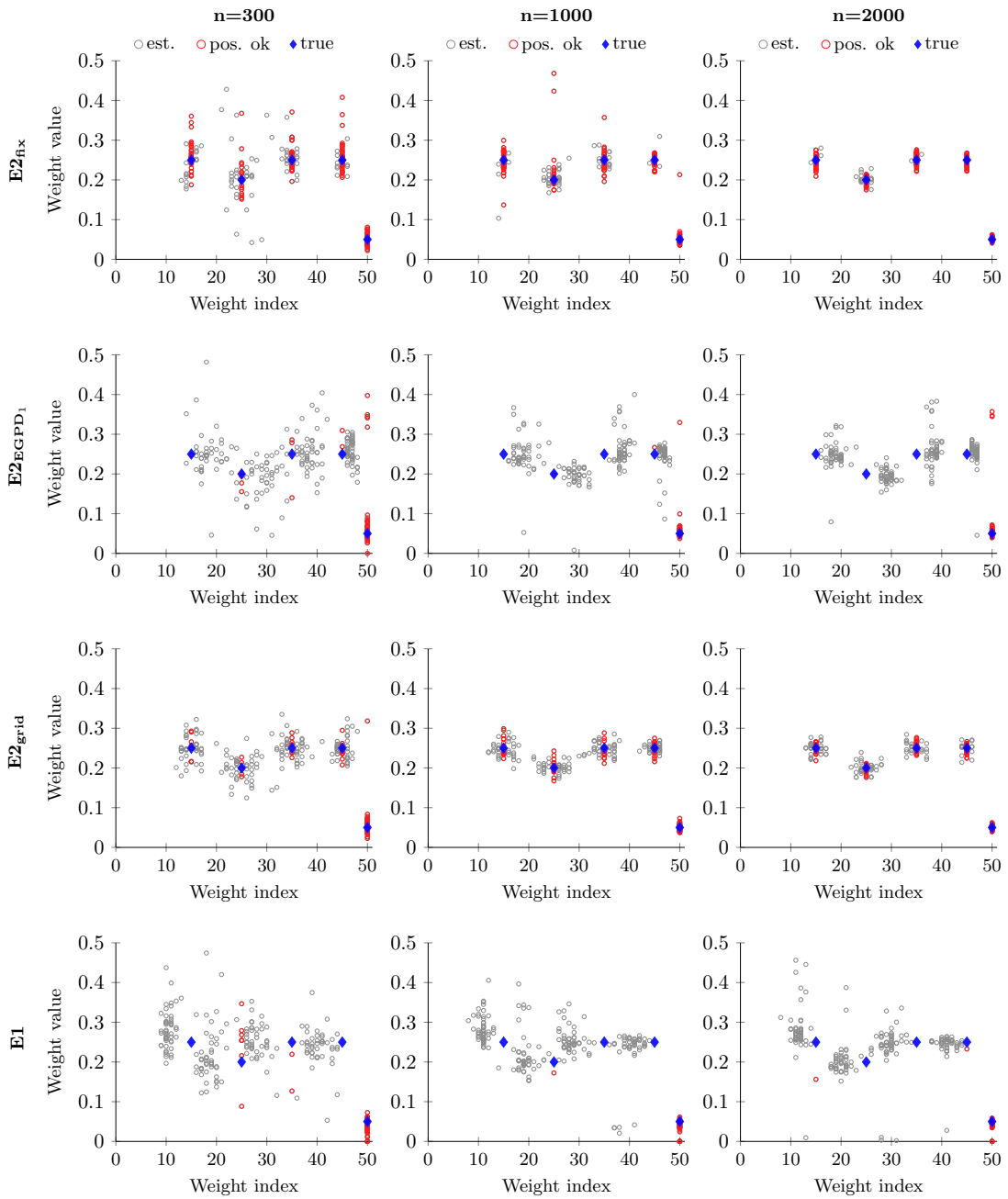

Figure 2.3.3: Estimated weights for the 50 replicates, for the sample sizes $n = 300, 1000, 2000$ and the estimation approaches E1, E2_{fix}, $\text{E2}_{\text{EGPD}_1}$ and E2_{grid} (blue diamond=true value, gray circle=estimated value, red circle=estimated value that is on the true position)

3.3.1.2 Quantile and density analysis

To further evaluate the performance of EGPD-BB, we analyze the estimated quantiles, with a particular interest in the extreme ones, and the density approximation error. Since the idea of using an EGPD-BB model is to add flexibility to the already existent parametric EGPD models

(see (2.1.16a)-(2.1.16d)), our analysis is based on a comparison with the EGPD_1 model.

For the quantile analysis, we consider estimating 99 quantiles, from 0.01 to 0.99, equally spaced. In order to measure the estimation accuracy of each model we compute the RMSE, and for comparing the estimations of two models, we compute the Ratio Root Mean Squared Error (R_{RMSE}). We have provided more details about these measures in Section 2.3.2.

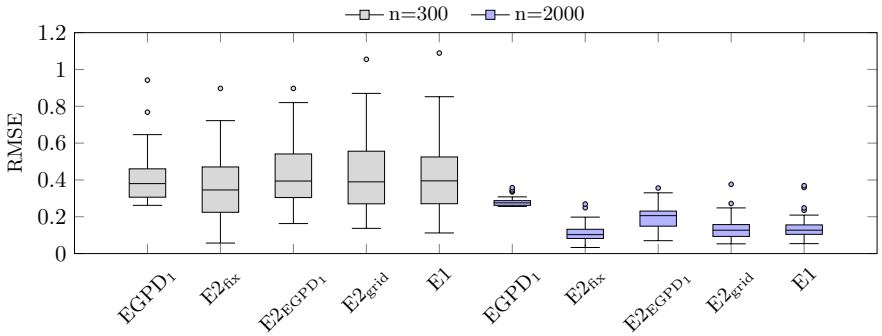

Figure 2.3.4: Boxplots of the estimated quantiles' RMSEs of both EGPD_1 and EGPD-BB (with estimation approaches E1, E2_{fix}, $\text{E2}_{\text{EGPD}_1}$ and E2_{grid} and $m = m_{\text{true}}$) over 50 replicates, for the sample sizes $n = 300, 2000$.

Figure 2.3.4 illustrates the boxplots of the RMSE values of the estimated quantiles from the 50 replicates (*i.e.*, each boxplot contains 50 RMSEs, one for each replicate), for EGPD_1 and EGPD-BB including each estimation approach (E1, E2_{fix}, $\text{E2}_{\text{EGPD}_1}$ and E2_{grid}). We show only the output for $n = 300$ (in gray) and $n = 2000$ (in blue), the other sample size (*i.e.*, $n = 1000$) providing an in-between interpretation. A brief examination of these plots shows that the accuracy of all models improves as the sample size is increased, such that for $n = 2000$ the RMSEs become smaller and less variable. Moreover, while for a sample size $n = 300$, the performance of EGPD_1 and EGPD-BB is more or less similar, for $n = 2000$ the RMSEs of the EGPD-BB are smaller than those of EGPD_1. Thus, EGPD-BB appear to be influenced more by the increase in the sample size compared with EGPD_1, but this behavior is not unexpected in a semiparametric framework.

A more clear overview of the contrast in quantile estimation error between EGPD_1 and EGPD-BB is provided in Table 2.3.2. Here, we display the percentage of time when the ratio between the RMSE of EGPD_1 and EGPD-BB (*i.e.*, $\text{R}(sim_i) = \frac{\text{RMSE}_{\text{EGPD}_1}(sim_i)}{\text{RMSE}_{\text{EGPD-BB}}(sim_i)}$) is larger than 1, that is, we account for the times when $\text{RMSE}_{\text{EGPD-BB}}$ is smaller than $\text{RMSE}_{\text{EGPD}_1}$. Thus, while for a small sample size EGPD_1 achieves smaller RMSEs for approximately half of the replicates, when the sample size increases, the EGPD-BB chieves a better performance for almost all replicates.

Table 2.3.2: Percentage of time the ratio between $\text{RMSE}_{\text{EGPD}_1}(sim_i)$ and $\text{RMSE}_{\text{EGPD-BB}}(sim_i)$ is larger than 1, for $i = 1, 2, ..., 50$ replicates, for the sample sizes $n = 300, 1000, 2000$ and EGPD-BB estimation approaches E1, E2_{fix}, $\text{E2}_{\text{EGPD}_1}$ and E2_{grid}, for the case $m = m_{\text{true}}$

	$\mathbf{E2_{fix}}$	$\mathbf{E2_{EGPD_1}}$	$\mathbf{E2_{grid}}$	**E1**
n=300	68%	40%	54%	50%
n=1000	100%	68%	92%	90%
n=2000	100%	90%	98%	96%

We have compared so far the accuracy of EGPD_1 with EGPD-BB with respect to the quantile estimation. Further on, we analyze the performance of the four estimation approaches

considered for EGPD-BB, *i.e.*, E1, E2_{fix}, $\text{E2}_{\text{EGPD}_1}$ and E2_{grid}. As illustrated in Figure 2.3.4, while for a sample size $n = 300$ there are no significant differences in RMSEs, for a larger sample size (*i.e.*, $n = 2000$), the $\text{E2}_{\text{EGPD}_1}$ errors become more variable than for the other approaches, but it still provides a reduced estimation error compared to EGPD_1. Moreover, the efficiency in quantile estimation of E1 and E2_{grid} is very close to the one of E2_{fix} (*i.e.*, the case when the GPD parameters are considered known), meaning that the inclusion of GPD parameters in the estimation approach does not cause a decrease in quantiles estimation performance. The ratio RMSE from Table 2.3.2 display the same conclusion.

Until now, we have analyzed the overall performance of the estimated quantiles, but it is essential to also measure the models' efficiency with respect to large quantiles, such as 90-, 95-, 99-th quantile (q_{90}, q_{95}, q_{99}), *i.e.*, to assess the performance of the models on the upper tail.

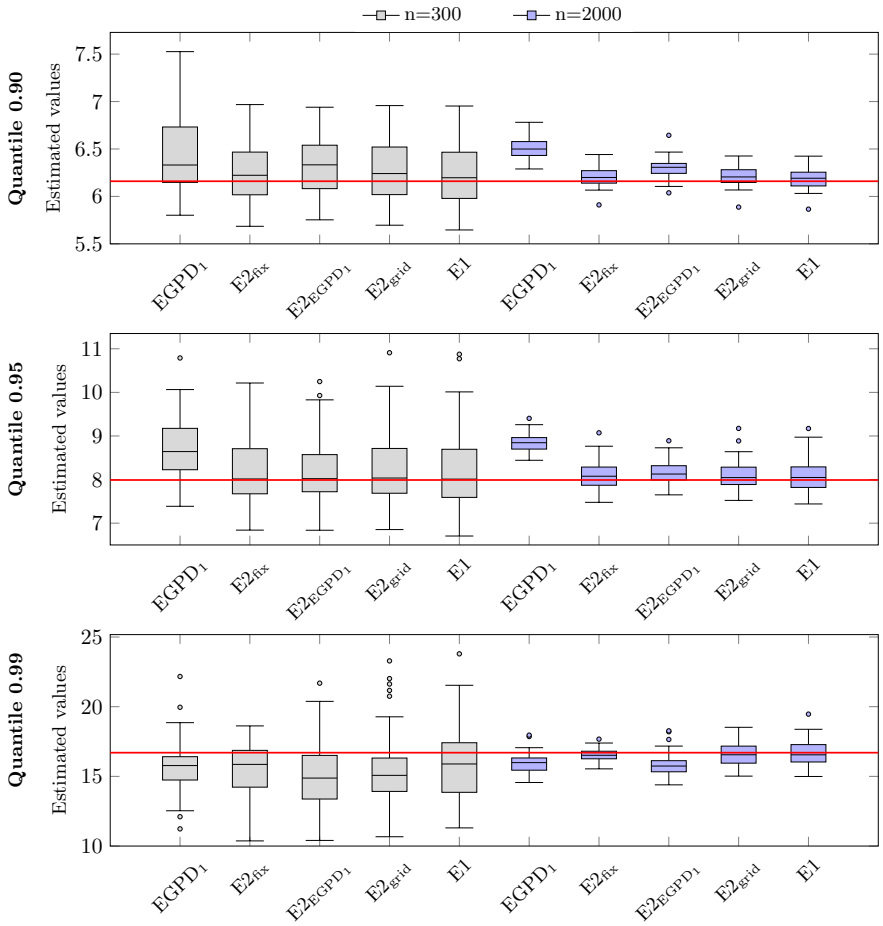

Figure 2.3.5: Boxplots of the estimated $q_{0.9}, q_{0.95}, q_{0.99}$ quantiles of both EGPD_1 and EGPD-BB (with estimation approaches E1, E2_{fix}, $\text{E2}_{\text{EGPD}_1}$ and E2_{grid}, and $m = m_{\text{true}}$) over 50 replicates, for the sample sizes $n = 300$ and 2000 (the red horizontal line indicates the true value of the quantiles)

Figure 2.3.5 illustrates the boxplots of the estimated values for the three extreme quantiles from the 50 replicates, for each model and two different sample size datasets, *i.e.*, $n = 300$ and $n = 2000$. Here, it can be observed that the sample size greatly affects the quantile estimation, more specifically, the estimates associated with a larger sample size ($n = 2000$) are less variable

compared to the ones from a small sample size ($n = 300$). Moreover, there is an overestimation tendency for q_{90}, while on the contrary q_{99} is slightly underestimated, when compared to the true value (red horizontal line), especially at EGPD_1 and EGPD-BB with $\text{E2}_{\text{EGPD}_1}$ estimation approach, but the overall assessment is adequate.

Table 2.3.3: Ratio between the RMSEs of EGPD_1 and EGPD-BB (with estimation approaches E1, E2_{fix}, $\text{E2}_{\text{EGPD}_1}$ and E2_{grid}, and $m = m_{\text{true}}$) for $q_{0.9}, q_{0.95}, q_{0.99}$ quantiles over 50 replicates, for the sample sizes $n = 300, 1000, 2000$ (in red the cases where EGPD_1 performs better than EGPD-BB)

	$\mathbf{E2_{fix}}$			$\mathbf{E2_{EGPD_1}}$			$\mathbf{E2_{grid}}$			**E1**		
	$q_{0.90}$	$q_{0.95}$	$q_{0.99}$	$q_{0.90}$	$q_{0.95}$	$q_{0.99}$	$q_{0.90}$	$q_{0.95}$	$q_{0.99}$	$q_{0.90}$	$q_{0.95}$	$q_{0.99}$
n=300	1.540	1.160	1.027	1.460	1.221	0.781	1.490	1.120	0.774	1.598	1.019	0.805
n=1000	2.716	1.688	1.621	1.974	1.921	0.761	2.409	1.723	0.980	2.660	1.541	0.994
n=2000	3.364	2.566	2.096	2.156	2.878	0.811	3.243	2.529	1.207	3.495	2.371	1.126

To compare the performance of EGPD_1 with EGPD-BB more clearly, Table 2.3.3 displays the ratio between the $\text{RMSE}_{\text{EGPD}_1}$ and the $\text{RMSE}_{\text{EGPD-BB}}$ of the estimates (*i.e.*, $\text{R}(q_p) = \frac{\text{RMSE}_{\text{EGPD}_1}(q_p)}{\text{RMSE}_{\text{EGPD-BB}}(q_p)}$) for each of the three extreme quantiles, over the 50 replicates. Recall that a ratio larger than 1 implies that EGPD-BB has a better performance in estimating that quantile when compared to EGPD_1. The ratios presented here show that EGPD_1 provides better estimates for q_{99} when the sample size is small or medium (red colored cells), but these estimates are outperformed by EGPD-BB when the sample size increases to $n = 2000$, except for the case of $\text{E2}_{\text{EGPD}_1}$.

One last point that we want to study is the approximation error of the density. We use the MIAE, see details of this performance indicator in Section 2.3.2, to check which model, EGPD_1 or EGPD-BB, provides a smaller error, and also which EGPD-BB estimation approach yields a more accurate smoothing.

In Table 2.3.4 we can see that EGPD_1 has a larger MIAE when compared to EGPD-BB, regardless of the estimation approach. If we focus on the three practical estimation methods, *i.e.*, $\text{E2}_{\text{EGPD}_1}$, E2_{grid} and E1, we see that the best output is given by E2_{grid} (red in the mentioned table), which in fact in not very far from the "ideal" case E2_{fix} (the case when the GPD parameters are considered known). The other two methods behave almost the same when compared to each other, but E1 tends to provide improved estimates, especially when the sample size increases. We highlight in blue the second best approximation. This is even more clearly illustrated in Figure 2.3.6, where we display the fitted densities, for both EGPD_1 and EGPD-BB (with estimation approaches E1, E2_{fix}, $\text{E2}_{\text{EGPD}_1}$ and E2_{grid}) models. While EGPD_1 does not fit accurately the bulk of the distributions, EGPD-BB is considerably more flexible in this sense.

Table 2.3.4: MIAE of EGPD_1 and EGPD-BB (with estimation approaches E1, E2_{fix}, $\text{E2}_{\text{EGPD}_1}$ and E2_{grid}, and $m = m_{\text{true}}$) over 50 replicates, for the sample sizes $n = 300, 1000, 2000$ (in red the best and in blue the second best approximation error for each sample size)

	$\mathbf{EGPD_1}$	$\mathbf{E2_{fix}}$	$\mathbf{E2_{EGPD_1}}$	$\mathbf{E2_{grid}}$	**E1**
n=300	0.311	0.128	0.145	0.137	0.146
n=1000	0.305	0.071	0.094	0.080	0.094
n=2000	0.303	0.052	0.079	0.061	0.074

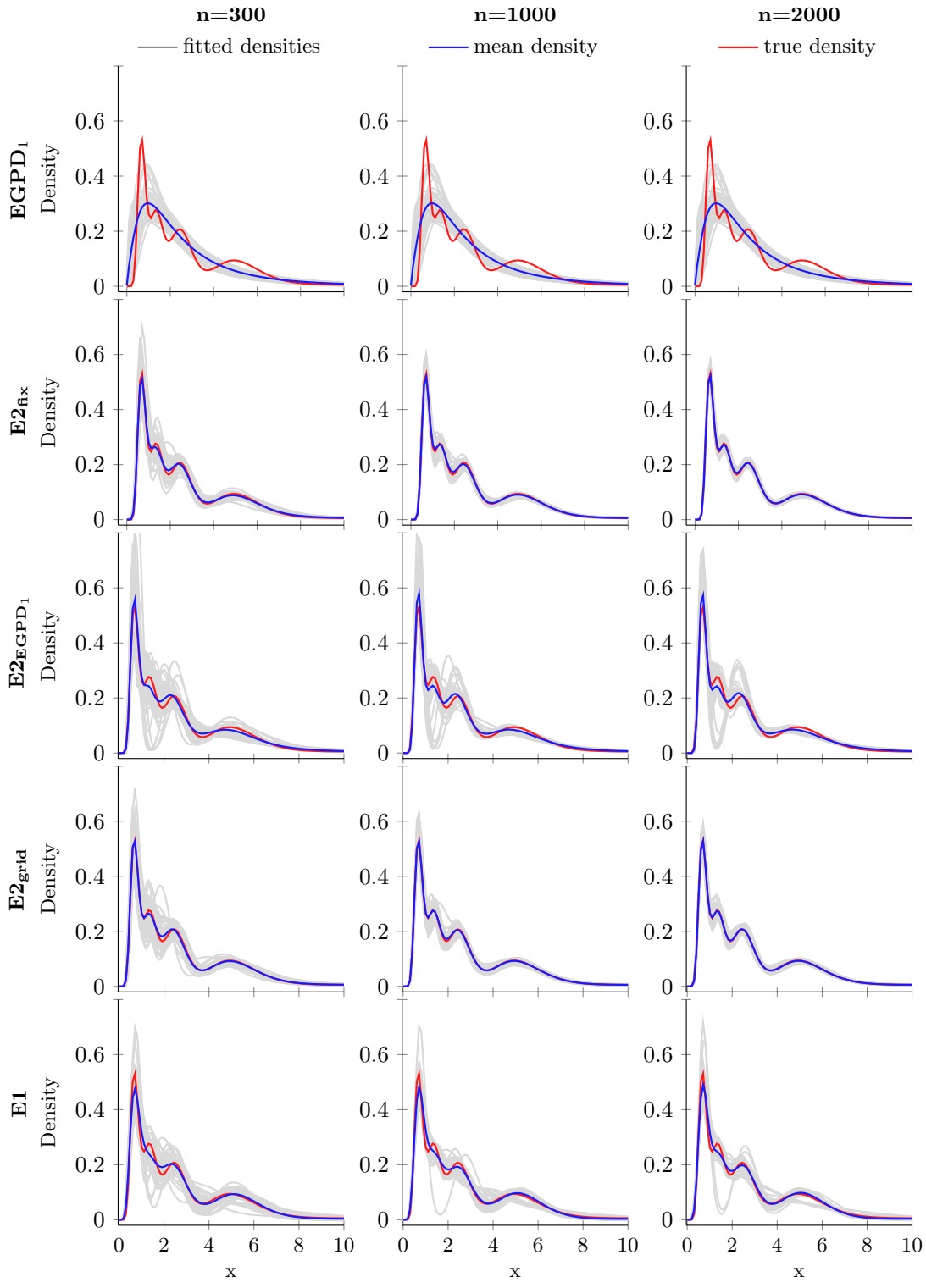

Figure 2.3.6: Fitted densities of the EGPD-BB (with estimation approaches E1, E2_{fix}, $\text{E2}_{\text{EGPD}_1}$ and E2_{grid}, and $m = m_{\text{true}}$) and the EGPD_1 models, over 50 replicates, for the sample sizes $n = 300, 1000, 2000$ (red = the true density, gray = the density fit of each replicate, blue = the mean of the 50 fitted densities)

3.3.2 Simulation study with the hyperparameter $m = m_{opt}$

The purpose of this study is to assess the performance of the EGPD-BB when the hyperparameter m is set to the optimal value suggested by *Babu et al.* (2002), *i.e.*, $m_{\text{opt}} = \frac{n}{\log(n)}$, where n is the sample size. We compare our model with EGPD_1. Due to the similarities in quantile estimation and density approximation between EGPD-BB when E2_{grid} and E1 are applied, but also taking into account their computational time, in this section we focus only on the performance of $\text{E2}_{\text{EGPD}_1}$ and E1, and not E2_{grid}.

As mentioned in the theoretical part, for large sample size datasets, the optimal value suggested by *Babu et al.* (2002) for the hyperparameter m might be too large, and thus, it can create an over-fitting of the density. For this reason, in this simulation study we analyze the performance of the models for the case when $m = m_{\text{opt}}$, but also the case when this degree is $m = \frac{m_{\text{opt}}}{2}$. The latter setting serves only as a contrasting example and it does not stand as a solution for estimating m.

We consider for this simulation study 100 replicates from a mixture of two gamma densities and a GPD (Mix2GaGPD), with different sample sizes: $n = 300, 600, 1000$. Recall that a mixture of two gamma densities is defined as $f_{2\text{Ga}}(x; \alpha, \beta, p) = pf_1(x; \alpha_1, \beta_1) + (1-p)f_2(x; \alpha_2, \beta_2)$, where f_i is the pdf of the gamma distribution with shape and scale parameters α_i and β_i, respectively. To preserve the continuity at the jointure point with GPD, *i.e.*, at u, the scale parameter σ of the GPD is considered as $\sigma = \frac{1-F_{2\text{Ga}}(u;\alpha,\beta,p)}{f_{2\text{Ga}}(u;\alpha,\beta,p)}$. We refer to Section 2.3.1 from the previous chapter for more details about this type of mixtures.

We use the following parameters' setting for this simulation study:

- Gamma mixture parameter: $\alpha = (2, 4)$, $\beta = (1, 2)$, $p = (0.4, 0.6)$,
- GPD parameters: $(\xi, u) = (0.1, q_{90})$, where q_{90} denotes the 90-th quantile of the two gamma mixture.

The corresponding optimal degrees for each sample size are: $m_{\text{opt}}^{n=300} = 53$, $m_{\text{opt}}^{n=600} = 94$ and $m_{\text{opt}}^{n=1000} = 145$.

We study the performance of both EGPD-BB and EGPD_1 in approximating the true density and estimating three extreme quantiles, *i.e.*, q_{90}, q_{95}, q_{99}. Figure 2.3.8 and Figure 2.3.7 illustrate the fitted densities and the boxplots of the estimated quantiles over the 100 replicates, respectively, for each considered case. Table 2.3.5 and Table 2.3.6 display, respectively, the MIAEs of the density approximations and the ratio between EGPD_1 and EGPD-BB of the quantiles RMSEs (*i.e.*, $\text{R}_{\text{RMSE}}(q_p) = \frac{\text{RMSE}_{\text{EGPD}_1}(q_p)}{\text{RMSE}_{\text{EGPD-BB}}(q_p)}$), for each case.

Table 2.3.5: MIAE of the EGPD_1 and EGPD-BB (with estimation approaches $\text{E2}_{\text{EGPD}_1}$ and E1) over 100 replicates, when $m = m_{\text{opt}}$ or $m = \frac{m_{\text{opt}}}{2}$, for the sample sizes $n = 300, 600, 1000$ (in red is highlighted the best approximation error between EGPD_1, E2_{EGPD} and E1, for each sample size and scenario)

	$m = m_{opt}$			$m = m_{opt}/2$		
	$\textbf{EGPD}_1^{(1)}$	$\textbf{E2}_{\textbf{EGPD}_1}$	**E1**	$\textbf{EGPD}_1^{(1)}$	$\textbf{E2}_{\textbf{EGPD}_1}$	**E1**
n=300	0.163	0.162	0.161	0.163	0.164	0.163
n=600	0.159	0.159	0.151	0.159	0.113	0.111
n=1000	0.156	0.222	0.182	0.156	0.106	0.109

[1] EGPD_1 does not depend on the choice of m, so the two columns (*i.e.*, 1^{st} and 4^{th}) that display the results of the EGPD_1) model are the same; we use the second EGPD_1 column to better display the results for the case when $m = m_{opt}/2$

Table 2.3.6: Ratio between the RMSEs of EGPD_1 and EGPD-BB (with estimation approaches E1 and $\text{E2}_{\text{EGPD}_1}$) for $q_{0.9}, q_{0.95}, q_{0.99}$ quantiles over 100 replicates, when $m = m_{\text{opt}}$ or $m = \frac{m_{\text{opt}}}{2}$, for the sample sizes $n = 300, 600, 1000$ (in red the cases where EGPD_1 performs better than EGPD-BB)

	$m = m_{\text{opt}}$						$m = m_{\text{opt}}/2$					
	$\text{E2}_{\text{EGPD}_1}$			E1			$\text{E2}_{\text{EGPD}_1}$			E1		
	$q_{0.9}$	$q_{0.95}$	$q_{0.99}$	$q_{0.9}$	$q_{0.95}$	$q_{0.99}$	$q_{0.9}$	$q_{0.95}$	$q_{0.99}$	$q_{0.9}$	$q_{0.95}$	$q_{0.99}$
n=300	1.125	1.558	1.079	1.131	1.532	1.090	1.175	1.422	0.769	0.969	0.783	0.302
n=600	1.298	1.716	0.827	1.144	1.685	0.641	1.301	1.888	1.252	1.315	1.934	1.184
n=1000	0.406	0.354	0.130	1.255	1.942	1.170	1.516	2.176	1.201	1.339	2.348	1.569

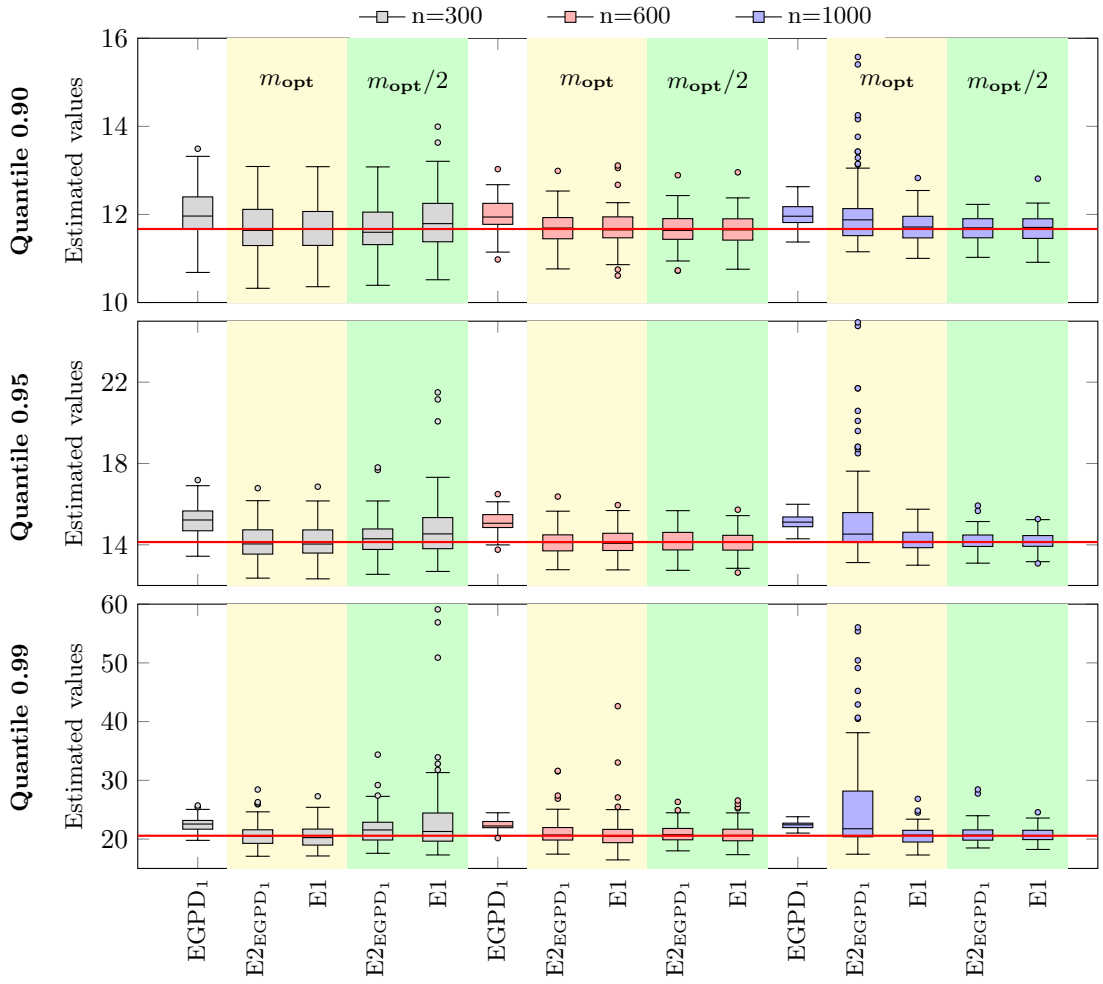

Figure 2.3.7: Boxplots of the estimated $q_{0.9}, q_{0.95}, q_{0.99}$ quantiles of both EGPD_1 and EGPD-BB (with estimation approaches E1 and $\text{E2}_{\text{EGPD}_1}$) over 100 replicates, when $m = m_{\text{opt}}$ or $m = \frac{m_{\text{opt}}}{2}$, for the sample sizes $n = 300, 600, 1000$ (the red horizontal line indicates the true value of the quantiles)

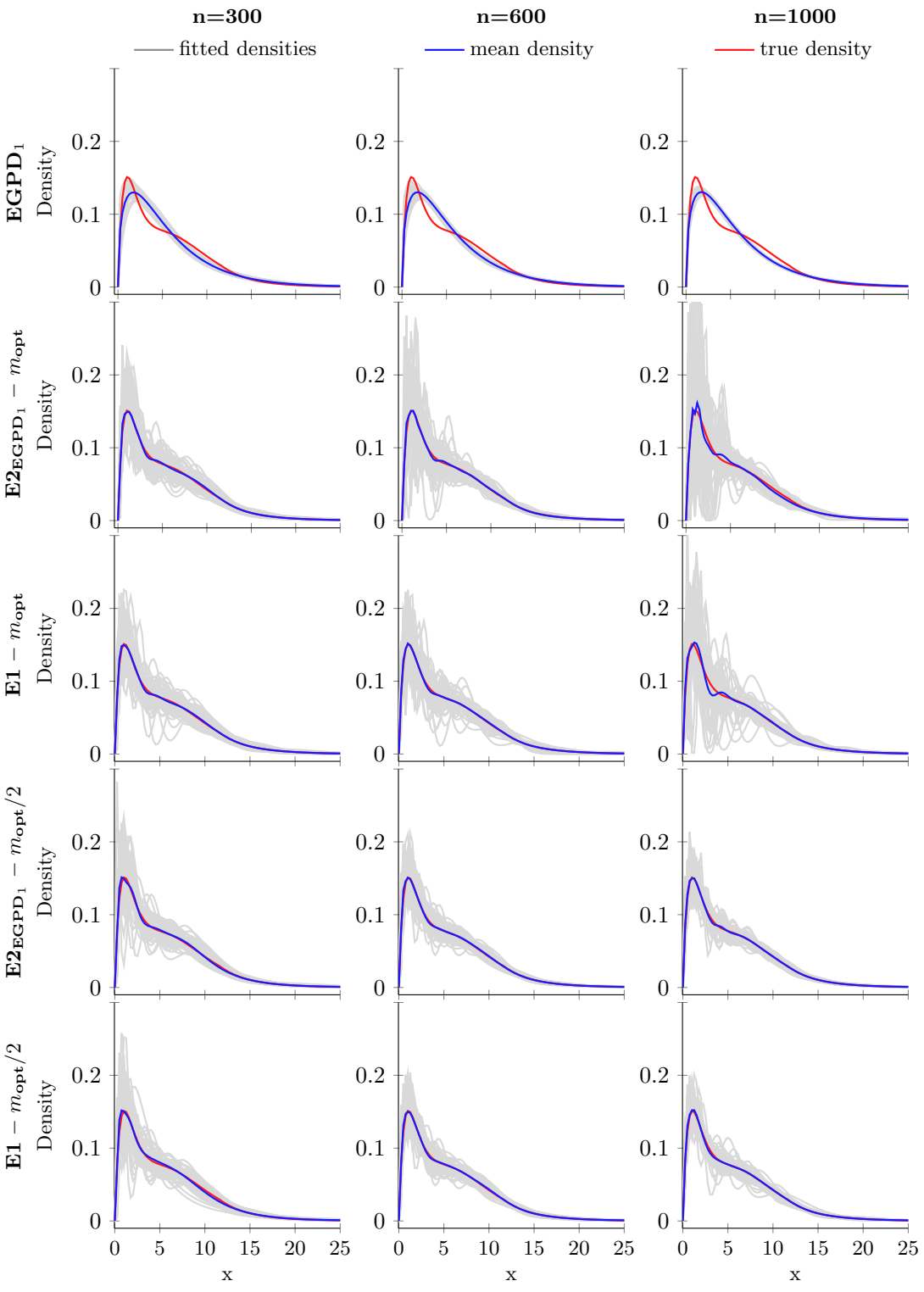

Figure 2.3.8: Fitted densities of the EGPD-BB (with estimation approaches $\text{E2}_{\text{EGPD}_1}$ and E1) and the EGPD_1 models, over 100 replicates, when $m = m_{\text{opt}}$ or $m = \frac{m_{\text{opt}}}{2}$, for the sample sizes $n = 300, 600, 1000$ (red = the true density, gray = the density fit of each replicate, blue = the mean of the 100 fitted densities)

We focus, first, on the performance of the models when the hyperparameter is set to the optimal value, *i.e.*, $m = m_{opt}$. With respect to the sample size, we have that:

1. **Small sample size ($n = 300$ and $m_{\mathbf{opt}} = 53$)**

   Both EGPD_1 and EGPD-BB have similar MIAE and quantile estimation error (ratio is close to 1), but the EGPD-BB model, regardless of the estimation approach applied, E1 or $\text{E2}_{\text{EGPD}_1}$, tends to perform slightly better. A more clear overview of the efficiency of each model in approximating the density is illustrated in Figure 2.3.8. Here, it can be observed that EGPD_1 does not capture well the bulk part of the true density (red curve), while EGPD-BB yields, in average, an approximation (see blue curve) very similar to the true one (red curve), even though the estimated curves (in gray) have more variability than the ones provided by the EGPD_1 model. Figure 2.3.7 displays the estimated q_{90}, q_{95}, q_{99} quantiles. It can be seen that EGPD_1 overestimates all three quantiles, but also that, even though, EGPD-BB corrects this aspect and provide improved quantile estimates, it shows again an increased variability in the estimations when compared to EGPD_1. Finally, there is no clear difference between the two estimations approaches of the EGPD-BB model, *i.e.*, $\text{E2}_{\text{EGPD}_1}$ and E1.

2. **Medium sample size ($n = 600$ and $m_{\mathbf{opt}} = 94$)**

   Despite the similarities in the MIAE of EGPD_1 and EGPD-BB, Figure 2.3.8 shows that the density approximations (in gray) yielded by EGPD-BB are much more noisy, thus we might have a first sign of over-fitting due to a too large degree m. EGPD_1 is clearly still overestimating the true quantiles and provides an increased estimation error compared to EGPD-BB, *i.e.*, the ratios from Table 2.3.6 are larger than 1. There is, though, an exception at q_{99}, where due to a more variable and unstable quantile estimation, see Figure 2.3.7, EGPD-BB performs worse than EGPD_1 (see red color cells in Table 2.3.6), even if the latter one still overestimates the true quantile.

3. **Large sample size ($n = 1000$ and $m_{\mathbf{opt}} = 145$)**

   As the sample size is larger, there is even more evidence that EGPD-BB is over-fitting the true density, as can be seen from Figure 2.3.8 and the MIAEs from Table 2.3.5. Moreover, the EGPD-BB fit of the extreme quantiles (q_{90}, q_{95}, q_{99}) appears to be greatly affected by this over-fitting when the estimation approach $\text{E2}_{\text{EGPD}_1}$ is applied, yielding poor and variable quantile estimations as illustrated in Figure 2.3.7. On the contrary, EGPD-BB with the estimation approach E1 provides an accurate fit of the extreme quantiles and it appears not to be affected in this sense by this over-fitting.

The above analysis shows that the setting of the hyperparameter m to the optimal value suggested by *Babu et al.* (2002) might not be a reliable solution, especially for large sample sizes where this optimal value might increase considerably, *e.g.*, $m_{\text{opt}} = 145$ for $n = 1000$. In the following, we evaluate the approximation error and quantile fit in the case where m is set to $m_{\text{opt}}/2$. Thus, with respect to the sample size, we have that:

1. **Small sample size ($n = 300$ and $m_{\mathbf{opt}}/2 = 27$)**

   The density fit and its MIAEs, displayed in Figure 2.3.8 and Table 2.3.5, respectively, do not appear to be significantly affected by the decrease of the degree m. On the contrary, a smaller degree, thus a fewer number of components in the mixture, greatly influences the fit of the extreme quantiles, as can be seen from Figure 2.3.7 and Table 2.3.6, for both estimations approaches ($\text{E2}_{\text{EGPD}_1}$ and E1).

2. **Medium sample size ($n = 600$ and $m_{\mathbf{opt}}/2 = 47$)**

Both the density and the quantiles fit are improved compare to the case when $m = m_{\text{opt}}$, the estimations are less variable and they provide this time a more accurate output than EGPD_1 model.

3. **Large sample size** ($n = 1000$ **and** $m_{\text{opt}}/2 = 73$)

   The difference between the fit of the models with $m = m_{\text{opt}}$ and $m = \frac{m_{\text{opt}}}{2}$ is even more clear in this case. The density and quantiles fits are greatly improved, such that the MIAEs are halved, as displayed in Table 2.3.5, and the quantile estimation errors are significantly decreased, outperforming the EGPD_1 model in all the considered cases.

In conclusion, the setting of the hyperparameter m to the optimal value $n/\log(n)$ appears to be suitable only for small sample size datasets, and, thus it is recommended that, for medium and, especially for large samples, to employ a different choice or estimation method to set this degree. Such method can be, for example, the LSCV methodology briefly described in Section 3.2.2.3. Moreover, the E1 estimation approach is less sensitive to the choice of m for large sample size datasets, yielding good estimates even for $m = m_{\text{opt}}$.

3.4 Case study: rainfall data

We now apply EGPD-BB on the two rainfall datasets used in the case study from Chapter 2. As a quick reminder, one dataset contains hourly precipitation from 1996 to 2011 recorded at the Lyon station, and the other one mean areal daily precipitation from 1948 to 2010 recorded at the Durance station. The sample sizes of our datasets, after removing the dry events, are:

Lyon: Spring-282, Summer-251, Fall-336, Winter-184

Durance: Spring-726, Summer-755, Fall-600, Winter-590.

3.4.1 Rainfall at the Lyon station

Motivated by our simulation Study 2, the degree of EGPD-BB is set to $m_{\text{opt}} = \frac{n}{\log(n)}$, *i.e.*, $\hat{m}_{\text{opt}}^{Spring} = 50$, $\hat{m}_{\text{opt}}^{Summer} = 46$, $\hat{m}_{\text{opt}}^{Fall} = 58$ and $\hat{m}_{\text{opt}}^{Winter} = 36$.

We evaluate the three estimation approaches analyzed in the simulation study for EGPD-BB, *i.e.*, E1, $\text{E2}_{\text{EGPD}_1}$ and E2_{grid}, for each season separately, and, then, we compare these estimates with EGPD_1. For E2_{grid}, we use a grid of 30 equally spaced points within the intervals $\sigma = [0.5, 4]$ and $\xi = [0, 0.5]$. We must mention that wider intervals were tested at first, but the best output was all the time included in a more narrow one. Thus, to fine tune the grid and reduce the computational time, we considered this limited set of values.

Table 2.3.7 displays for each estimation approach, the estimated GPD parameters $(\hat{\sigma}, \hat{\xi})$, the sparsity degree $\hat{M}$ (*i.e.*, the number of non-null weights) and the first non-null position $(\hat{s})$ in the estimated weights vector. It can be noticed from the sparsity degree that around 80% of the weights are set to zero. The values of the estimated GPD parameters for the three approaches are more or less the same, but special attention must be paid to approach E1, where the 95% confidence intervals tend to be larger than for the other estimation methods. Thus, E1 might encounter problems in finding the optimal result, fact that was already pointed out in the simulation study.

Table 2.3.7: Estimated GPD parameters with the associated 95% confidence intervals, sparsity degree $\hat{M}$ and the first non-null position ($\hat{s}$) in the estimated weights vector, for the EGPD-BB model with $m = m_{\text{opt}}$ fitted to the hourly Lyon rainfall data (1996-2011) by three estimations approaches: E1, $\text{E2}_{\text{EGPD}_1}$ and E2_{grid}, for each season

	Spring ($\hat{m}_{\text{opt}} = 50$)				**Summer** ($\hat{m}_{\text{opt}} = 45$)			
	$\hat{M}$	$\hat{s}$	$\hat{\sigma}$	$\hat{\xi}$	$\hat{M}$	$\hat{s}$	$\hat{\sigma}$	$\hat{\xi}$
$\textbf{E2}_{\textbf{EGPD}_1}$	9	8	$1.40_{[1.34,1.55]}$	$0.06_{[0.00,0.11]}$	10	5	$2.44_{[2.23,2.61]}$	$0.12_{[0.01,0.19]}$
$\textbf{E2}_{\textbf{grid}}$	8	8	$1.26_{[0.85,1.55]}$	$0.07_{[0.01,0.10]}$	8	6	$2.05_{[2.00,2.46]}$	$0.06_{[0.05,0.20]}$
E1	12	7	$1.53_{[1.16.1.90]}$	$0.00_{[0.00,0.21]}$	10	5	$2.58_{[1.41,3.70]}$	$0.10_{[0.00,0.66]}$

	Fall ($\hat{m}_{\text{opt}} = 57$)				**Winter** ($\hat{m}_{\text{opt}} = 36$)			
	$\hat{M}$	$\hat{s}$	$\hat{\sigma}$	$\hat{\xi}$	$\hat{M}$	$\hat{s}$	$\hat{\sigma}$	$\hat{\xi}$
$\textbf{E2}_{\textbf{EGPD}_1}$	12	8	$1.54_{[1.50,1.76]}$	$0.26_{[0.12,0.28]}$	6	9	$0.77_{[0.76,0.87]}$	$0.28_{[0.15,0.41]}$
$\textbf{E2}_{\textbf{grid}}$	11	8	$1.58_{[1.50,1.96]}$	$0.21_{[0.15,0.30]}$	7	10	$0.65_{[0.55,0.85]}$	$0.27_{[0.25,0.40]}$
E1	13	8	$1.50_{[1.14,2.98]}$	$0.10_{[0.00,0.30]}$	9	6	$1.64_{[0.64,2.23]}$	$0.12_{[0.00,0.36]}$

Figure 2.3.9 illustrates the histograms with the fitted densities for both EGPD_1 and EGPD-BB with the three estimation approaches, as well as the QQ-plots. The fitted densities and quantiles of the two models, regardless of the estimation approach in the EGPD-BB case, are very close to each other. EGPD-BB has an improved estimation compared to EGPD_1 for the bulk of the distribution for Spring, Summer and Winter, but it still does not capture completely all the peaks and valleys from Summer. A visual inspection of the fitted densities and quantiles brings the conclusion that E2_{grid} is achieving the best estimation. We also corroborate this conclusion with the classification given by the BIC displayed in Table 2.3.8. The best model selected by BIC is E2_{grid} (in red), for three out of the four seasons. In Winter, the BIC criteria of $\text{E2}_{\text{EGPD}_1}$ and E2_{grid} are very close, and in this case the method with the fewest number of parameters is favored and selected, *i.e.*, $\text{E2}_{\text{EGPD}_1}$.

Table 2.3.8: BIC output of EGPD-BB model for the hourly Lyon rainfall data (1996-2011), for each of the four seasons and the three estimation approaches: E1, $\text{E2}_{\text{EGPD}_1}$ and E2_{grid} (in red is indicated the best estimation approach for each season).

	Spring	**Summer**	**Fall**	**Winter**
$\textbf{E2}_{\textbf{EGPD}_1}$	833.541	970.854	1107.918	408.108
$\textbf{E2}_{\textbf{grid}}$	825.998	958.338	1102.086	409.688
E1	850.442	970.708	1112.258	430.454

Figure 2.3.10 illustrates the histograms with the fitted densities on the transformed data for each season and both EGPD_1 and EGPD-BB. Recall that for a random variable X, the transformed data for the EGPD model is $U = H_\xi(X)$, where H_ξ is the cdf of GPD. Thus, we display the histograms of the transformed data U together with the fit of the Bernstein-beta density estimator $g_{BB,m}(t)$ from (2.3.4) for EGPD-BB and the power function (*i.e.*, $g(u) = \kappa u^{\kappa-1}$) of the EGPD_1 model. It can be seen that EGPD_1 is very close to a straight line (due to the fact that the estimated k parameters are near 1, thus exact GPD model), while EGPD-BB captures well the peaks and valleys of the distribution, thus adds more flexibility for the bulk of the distribution. Moreover, $\text{E2}_{\text{EGPD}_1}$ and E2_{grid} have a very similar fit, while E1 is slightly different, especially in Winter, fact that can be seen also from the estimated GPD parameters in Table 2.3.7.

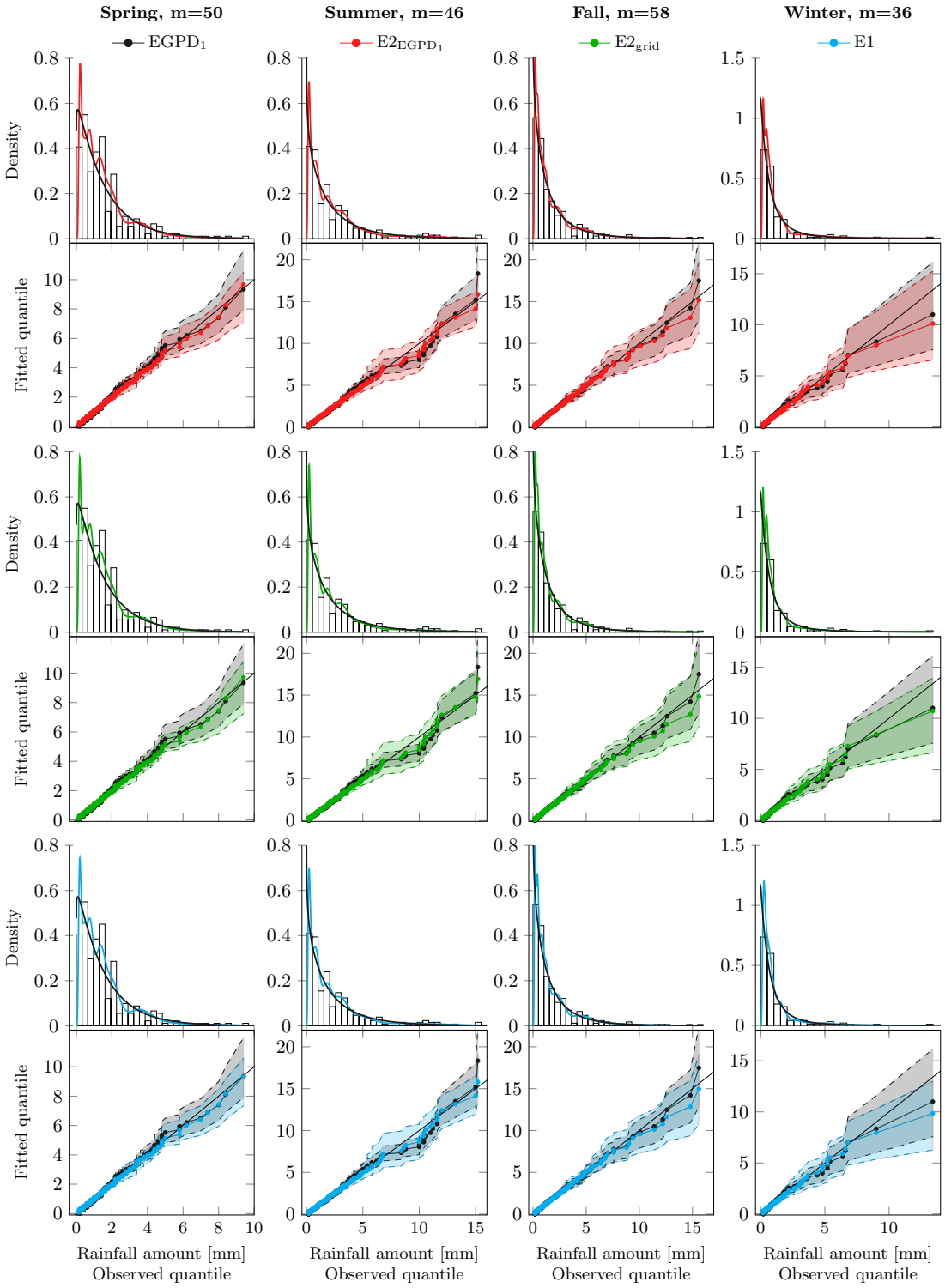

Figure 2.3.9: Histograms with the fitted densities and QQ-plots with the associated 95% confidence intervals for EGPD_1 and EGPD-BB with $m = m_{\text{opt}}$ for hourly Lyon rainfall data (1996-2011), for each of the four seasons. Three approaches are considered for the estimation of the EGPD-BB model: E1, $\text{E2}_{\text{EGPD}_1}$ and E2_{grid}

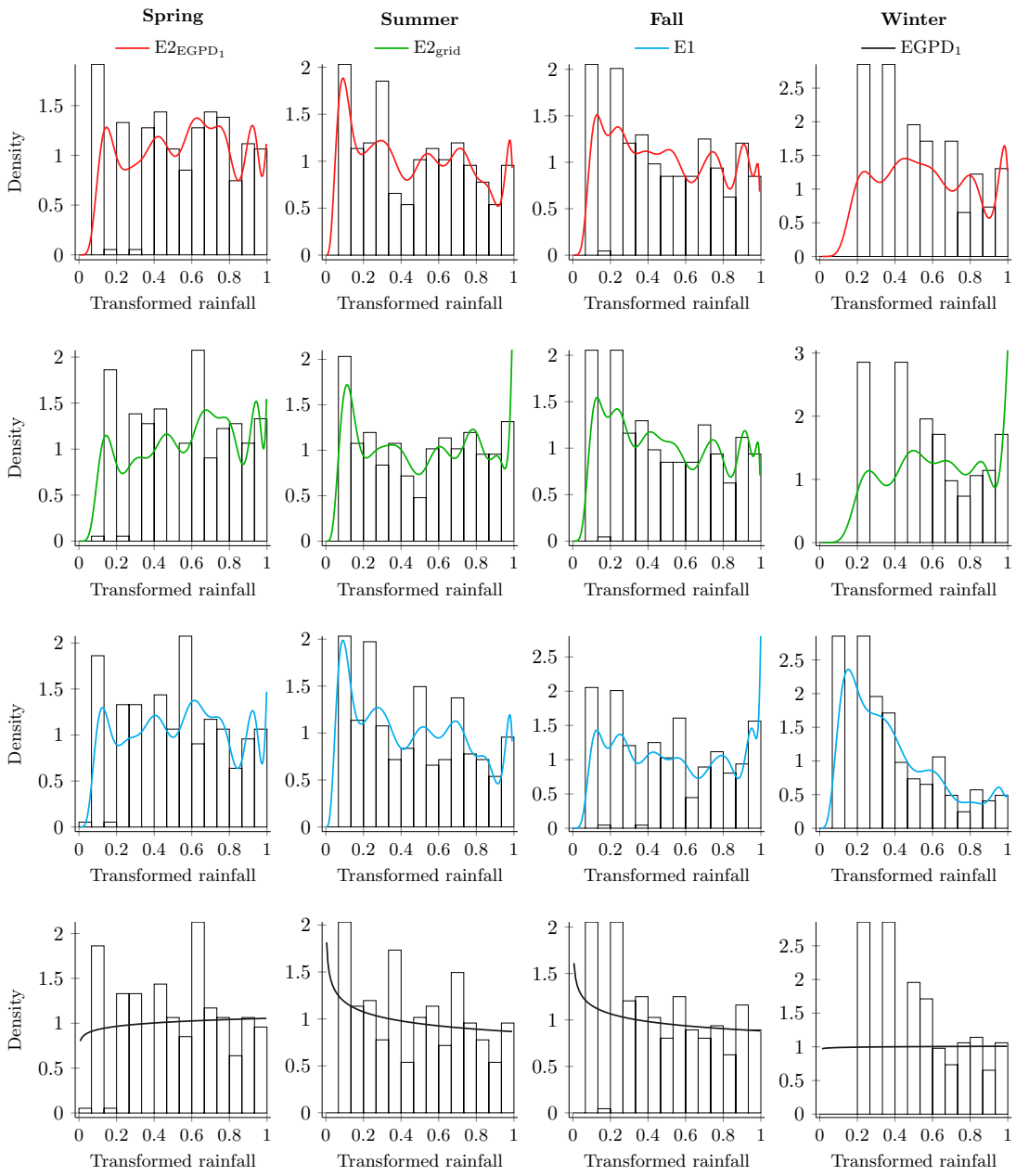

Figure 2.3.10: Histograms with the fitted densities for EGPD_1 and EGPD-BB with $m = m_{\text{opt}}$ of the transformed hourly Lyon rainfall data (1996-2011), *i.e.*, $U = H_\xi(X)$, for each of the four seasons. Three approaches are considered for the estimation of the EGPD-BB model: E1, $\text{E2}_{\text{EGPD}_1}$ and E2_{grid}.

Until now, the focus was mainly directed towards the performance evaluation of extreme quantiles from the right tail, but we neglect the small ones from the left tail. Figure 2.3.11 shows a zoom on the small values of the QQ-plots for both EGPD_1 and EGPD-BB. We illustrate only the plots for the estimation approach $\text{E2}_{\text{EGPD}_1}$, the other two approaches providing the same output. Consequently, we can see that EGPD-BB has a better fit compared to EGPD_1 for the

small values, thus, the fact that the lower tail constraint (see Section 3.2.1) was not directly included in the optimization problem, is not an issue, as the sparse behavior of the EGPD-BB model is forcing this constraint to be satisfied.

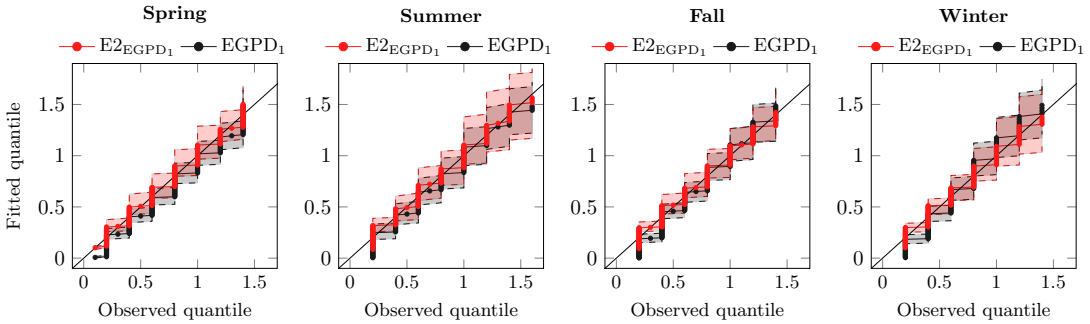

Figure 2.3.11: Zoom on the small values of the QQ-plots for EGPD_1 and EGPD-BB, with $m = m_{\text{opt}}$ and estimation approach $\text{E2}_{\text{EGPD}_1}$, for hourly Lyon rainfall data (1996-2011) for each of the four seasons

Finally, even though we concluded in the simulation study that for small sample sizes a setting of m to $m_{\text{opt}} = \frac{n}{\log(n)}$ brings good estimates, we also use the LSCV technique, detailed in Section 3.2.2.3, to find the optimal degree. We compute the LSCV value for each degree from the set $[25, 26, ..., m_{\text{opt}}]$. The selected hyperparameters m_{LSCV} for each season are, either the same, or close to the ones given by m_{opt}, *i.e.*, these degrees are $\hat{m}_{\text{LSCV}}^{Spring} = 50$, $\hat{m}_{\text{LSCV}}^{Summer} = 45$, $\hat{m}_{\text{LSCV}}^{Fall} = 57$ and $\hat{m}_{\text{LSCV}}^{Winter} = 34$. We do not present the fitted models for these cases as their performance is almost identical to the one already discussed above.

3.4.2 Rainfall at the Durance station

The performance evaluation for the Durance rainfall data is handled similarly to the Lyon case study. There is, though, an important difference regarding the sample size. While in the previous case study we worked with estimations on small sample sizes, in the Durance study we have a medium sample size case. We have seen in the simulation study that, for medium sample sizes, by the setting $m_{\text{opt}} = \frac{n}{\log(n)}$ might not be appropriate, due to an overestimation of this degree. However, as there is no precise boundary between a small and a medium sample size, we assess first the performance of the fitted models when $m = m_{\text{opt}}$. We obtain for each season the following degrees: $\hat{m}_{\text{opt}}^{Spring} = 111$, $\hat{m}_{\text{opt}}^{Summer} = 114$, $\hat{m}_{\text{opt}}^{Fall} = 94$ and $\hat{m}_{\text{opt}}^{Winter} = 93$. Moreover, for the estimation approach E2_{grid} we use 30 equally spaced values from the following intervals: $\sigma = [3, 15]$ and $\xi = [0, 0.5]$. We apply the same reasoning in setting these bounds as for the Lyon case study.

Table 2.3.9 displays the estimated values for the GPD parameters $(\hat{\sigma}, \hat{\xi})$, the sparsity degree $\hat{M}$ and the first non-null position $(\hat{s})$ in the estimated weights vector. While $\text{E2}_{\text{EGPD}_1}$ and E2_{grid} have similar estimates, one can clearly see a decline in the inference of E1. This estimation approach shows great instability in estimating the parameters, especially the shape parameter ξ, fact that can be easily checked by looking at the extremely wide confidence intervals.

Table 2.3.9: Estimated GPD parameters with the associated 95% confidence intervals, sparsity degree $\hat{M}$ and the first non-null position ($\hat{s}$) in the estimated weights vector, for the EGPD-BB model with $m = m_{\text{opt}}$ fitted to the daily Durance rainfall data (1948-2010) by three estimations approaches: E1, $\text{E2}_{\text{EGPD}_1}$ and E2_{grid}, for each season

	Spring ($\hat{m}_{\text{opt}} = 111$)				**Summer** ($\hat{m}_{\text{opt}} = 114$)			
	$\hat{M}$	$\hat{s}$	$\hat{\sigma}$	$\hat{\xi}$	$\hat{M}$	$\hat{s}$	$\hat{\sigma}$	$\hat{\xi}$
$\textbf{E2}_{\textbf{EGPD}_1}$	17	3	$7.77_{[7.40,9.47]}$	$0.09_{[0.04,0.16]}$	14	4	$4.46_{[4.19,4.55]}$	$0.30_{[0.22,0.32]}$
$\textbf{E2}_{\textbf{grid}}$	16	3	$8.84_{[6.00,9.00]}$	$0.18_{[0.10,0.20]}$	16	4	$5.11_{[5.01,6.00]}$	$0.24_{[0.15,0.35]}$
E1	12	3	$8.10_{[3.59,16.45]}$	$3.49_{[0.32,13.35]}$	10	3	$6.46_{[2.34,14.25]}$	$4.50_{[0.00,14.24]}$

	Fall ($\hat{m}_{\text{opt}} = 94$)				**Winter** ($\hat{m}_{\text{opt}} = 93$)			
	$\hat{M}$	$\hat{s}$	$\hat{\sigma}$	$\hat{\xi}$	$\hat{M}$	$\hat{s}$	$\hat{\sigma}$	$\hat{\xi}$
$\textbf{E2}_{\textbf{EGPD}_1}$	18	2	$12.24_{[11.71,12.92]}$	$0.09_{[0.05,0.14]}$	20	2	$10.53_{[9.75,10.74]}$	$0.00_{[0.00,0.13]}$
$\textbf{E2}_{\textbf{grid}}$	14	2	$11.26_{[9.98,14.00]}$	$0.13_{[0.05,0.15]}$	11	3	$9.00_{[9.00,11.84]}$	$0.07_{[0.01,0.10]}$
E1	18	5	$2.78_{[4.06,68.40]}$	$0.77_{[0.00,5.73]}$	11	3	$8.00_{[2.63,24.54]}$	$1.97_{[0.00,5.13]}$

Figure 2.3.12 shows the histograms with the fitted densities and the QQ-plots with the 95% confidence intervals, for both EGPD_1 and EGPD-BB with the three estimation approaches. The QQ-plots indicate that the right tail is poorly estimated by EGPD_1 for all seasons, while EGPD-BB corrects this behavior, especially during Fall. Further on, the three estimation approaches bring a rather diverse output for all seasons except Spring. E2_{grid} provides the most accurate fit, capturing very well both the extreme values and the bulk of the distribution. The estimation approach $\text{E2}_{\text{EGPD}_1}$ brings similar output, but with slightly smaller precision when it comes to modeling the extreme values from Summer and Winter. However, a different and less accurate fit is given by E1, for Fall and Winter. Even though its fit seems more accurate than the one yielded by EGPD_1, the confidence intervals are very wide in this case, thus the uncertainty of the estimations is larger.

The visual inspection of Figure 2.3.12 and the analysis of the estimated parameters from Table 2.3.9 is not supported by the classification given by BIC, displayed in Table 2.3.10. Except Fall, the best fit indicated by BIC is the one from E1. This behavior is not yet surprising as BIC favors the models with less parameters, thus it selects E1 which has a smaller sparsity degree.

Table 2.3.10: BIC output of EGPD-BB model for the daily Durance rainfall data (1948-2010), for each of the four seasons and the three estimation approaches: E1, $\text{E2}_{\text{EGPD}_1}$ and E2_{grid} (in red is indicated the best estimation approach for each season)

	Spring	**Summer**	**Fall**	**Winter**
$\textbf{E2}_{\textbf{EGPD}_1}$	4309.612	4243.730	4071.095	3715.641
$\textbf{E2}_{\textbf{grid}}$	4304.201	4244.859	4048.001	3672.778
E1	4289.813	4214.804	4067.789	3670.028

As part of the EGPD-BB inference, we want to also point our attention to the fit on the transformed data. Therefore, in Figure 2.3.13 we display the histograms of the transformed data U (*i.e.*, $U = H_\xi(X)$) together with the fit of the Bernstein-beta density estimator $g_{BB,m}$ for the EGPD-BB and the power function (*i.e.*, $g(u) = \kappa u^{\kappa-1}$) of the EGPD_1 model. The fact that E1 has different GPD estimated parameters compared to $\text{E2}_{\text{EGPD}_1}$ and E2_{grid} , is also illustrated by these plots. A simple visual diagnostic of these plots shows that EGPD-BB adds considerable flexibility to the fit of the transformed data compared to the power function from the EGPD_1 model, capturing very well the bulk of the distribution, as well as the tails.

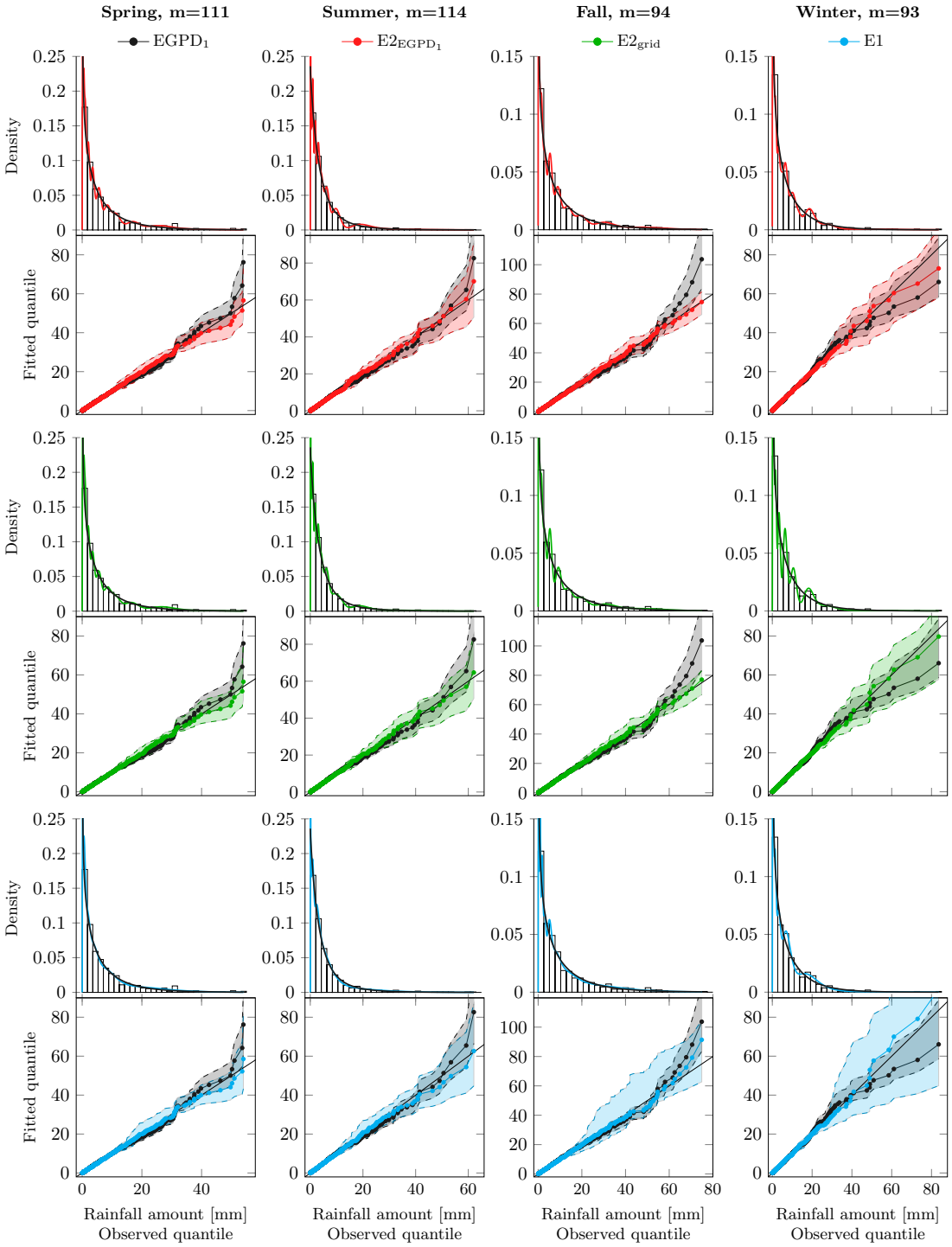

Figure 2.3.12: Histograms with the fitted densities and QQ-plots with the associated 95% confidence intervals for EGPD_1 and EGPD-BB with $m = m_{\text{opt}}$ for daily Durance rainfall data (1948-2010), for each of the four seasons. Three approaches are considered for the estimation of the EGPD-BB model: E1, $\text{E2}_{\text{EGPD}_1}$ and E2_{grid}

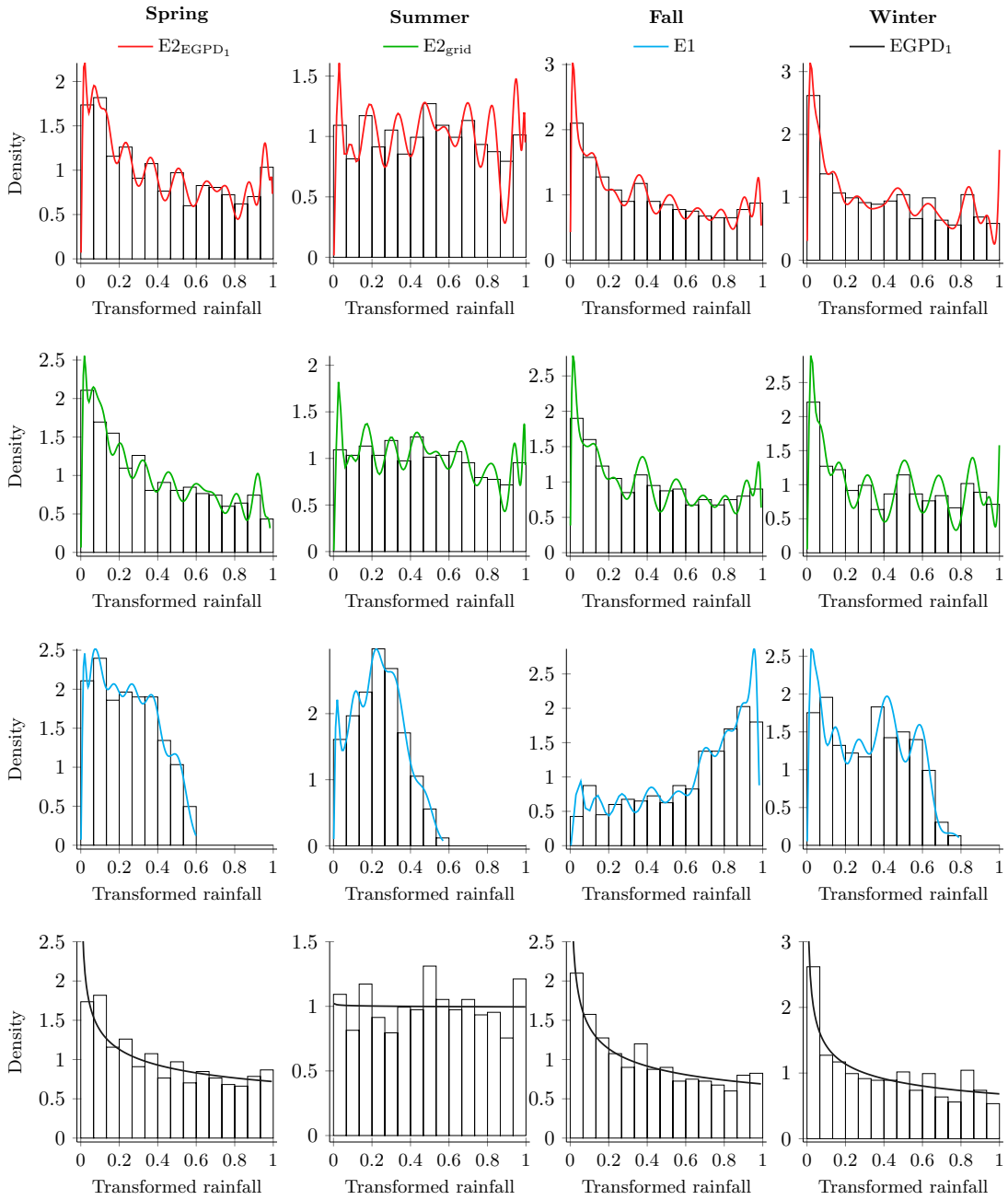

Figure 2.3.13: Histograms with the fitted densities for EGPD_1 and EGPD-BB with $m = m_{\text{opt}}$ of the transformed daily Durance rainfall data (1948-2010), *i.e.*, $U = H_\xi(X)$, for each of the four seasons. Three approaches are considered for the estimation of the EGPD-BB model: E1, $\text{E2}_{\text{EGPD}_1}$ and E2_{grid}.

Concerning the inference of the small values, in Figure 2.3.14 we show the zoom on the small values of the QQ-plots (all estimation approaches have similar outputs, so we show just $\text{E2}_{\text{EGPD}_1}$). Even though, both EGPD_1 and EGPD-BB capture well the small values, EGPD-BB outperforms EGPD_1, especially for Spring, Fall and Winter.

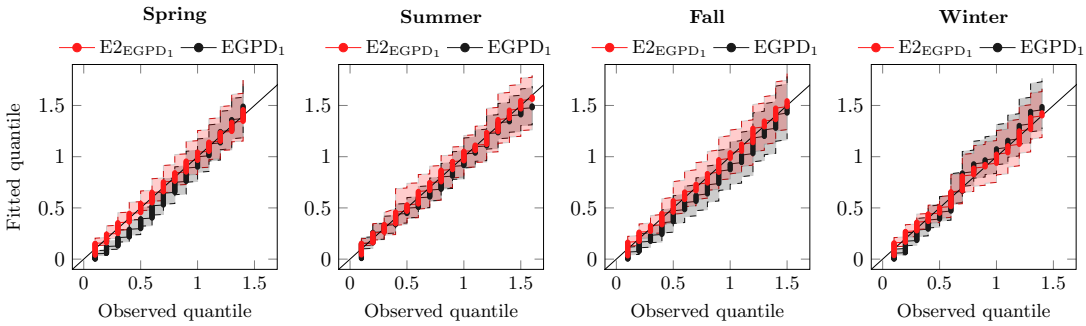

Figure 2.3.14: Zoom on the small values of the QQ-plots of EGPD_1 and EGPD-BB, with $m = m_{\text{opt}}$ and estimation approach $\text{E2}_{\text{EGPD}_1}$, for daily Durance rainfall data (1948-2010) for each of the four seasons

In what follows, we check: 1) if our choice of degree is indeed an optimal one, and 2) if improved estimates can be obtained by changing this degree. For this, we use for each season the LSCV approach to find the optimal hyperparameter m. We consider just the estimation approaches E1 and $\text{E2}_{\text{EGPD}_1}$, disregarding E2_{grid}, for two reasons: i) we already have a good fit for E2_{grid} with the degree m_{opt}, and ii) E2_{grid} is already very costly from a computational point of view, so by doing a cross-validation it becomes even more time consuming. The resulted LSCV optimal degrees (considering a grid $[25, 26, ..., m_{\text{opt}}]$) are: $\hat{m}_{\text{LSCV}}^{Spring} = 97$, $\hat{m}_{\text{LSCV}}^{Summer} = 106$, $\hat{m}_{\text{LSCV}}^{Fall} = 93$ and $\hat{m}_{\text{LSCV}}^{Winter} = 53$. While for Spring, Summer, and Fall there is no significant difference between m_{LSCV} and m_{opt}, for Winter the degree is almost halved.

Table 2.3.11: Estimated GPD parameters with the associated 95% confidence intervals, sparsity degree $\hat{M}$ and the first non-null position ($\hat{s}$) in the estimated weights vector, for the EGPD-BB model with $m = m_{\text{LSCV}}$, fitted to the daily Durance rainfall data (1948-2010) by two estimations approaches: E1 and $\text{E2}_{\text{EGPD}_1}$

	Spring ($\hat{m}_{\text{LSCV}} = 97$)				**Summer** ($\hat{m}_{\text{LSCV}} = 106$)			
	$\hat{M}$	$\hat{s}$	$\hat{\sigma}$	$\hat{\xi}$	$\hat{M}$	$\hat{s}$	$\hat{\sigma}$	$\hat{\xi}$
$\mathbf{E2_{EGPD_1}}$	18	3	$7.70_{[7.48,8.14]}$	$0.13_{[0.05,0.17]}$	17	4	$4.30_{[4.16,4.54]}$	$0.28_{[0.22,0.33]}$
E1	13	3	$8.06_{[3.52,19.25]}$	$0.90_{[0.02,9.67]}$	22	3	$6.78_{[3.30,15.61]}$	$5.08_{[0.00,7.66]}$

	Fall ($\hat{m}_{\text{LSCV}} = 93$)				**Winter** ($\hat{m}_{\text{LSCV}} = 53$)			
	$\hat{M}$	$\hat{s}$	$\hat{\sigma}$	$\hat{\xi}$	$\hat{M}$	$\hat{s}$	$\hat{\sigma}$	$\hat{\xi}$
$\mathbf{E2_{EGPD_1}}$	21	2	$12.42_{[11.93,13.43]}$	$0.09_{[0.05,0.13]}$	12	2	$10.48_{[9.92,10.82]}$	$0.03_{[0.00,0.11]}$
E1	13	2	$19.79_{[8.65,67.36]}$	$0.40_{[0.00,7.83]}$	12	2	$11.54_{[3.49,13.85]}$	$0.00_{[0.00,1.59]}$

Table 2.3.11 and Figure 2.3.15 display the estimated parameters and the fit of the models, respectively, when using the degree m_{LSCV}. Generally, the performance of these new models is the same or better than the case of m_{opt} for both approaches. We want to point, though, some particularities for each method compared to the case when $m = m_{\text{opt}}$:

- E1 improves the estimations for Fall and Winter, and the model becomes more stable having narrower confidence interval. However, we must note that the uncertainty in estimation is still high.
- $\text{E2}_{\text{EGPD}_1}$ improves the estimation for Fall, but with the cost of three more parameters, while in Winter it maintains the same fit, yet by reducing the number of weights to 12 compared to 20 before.

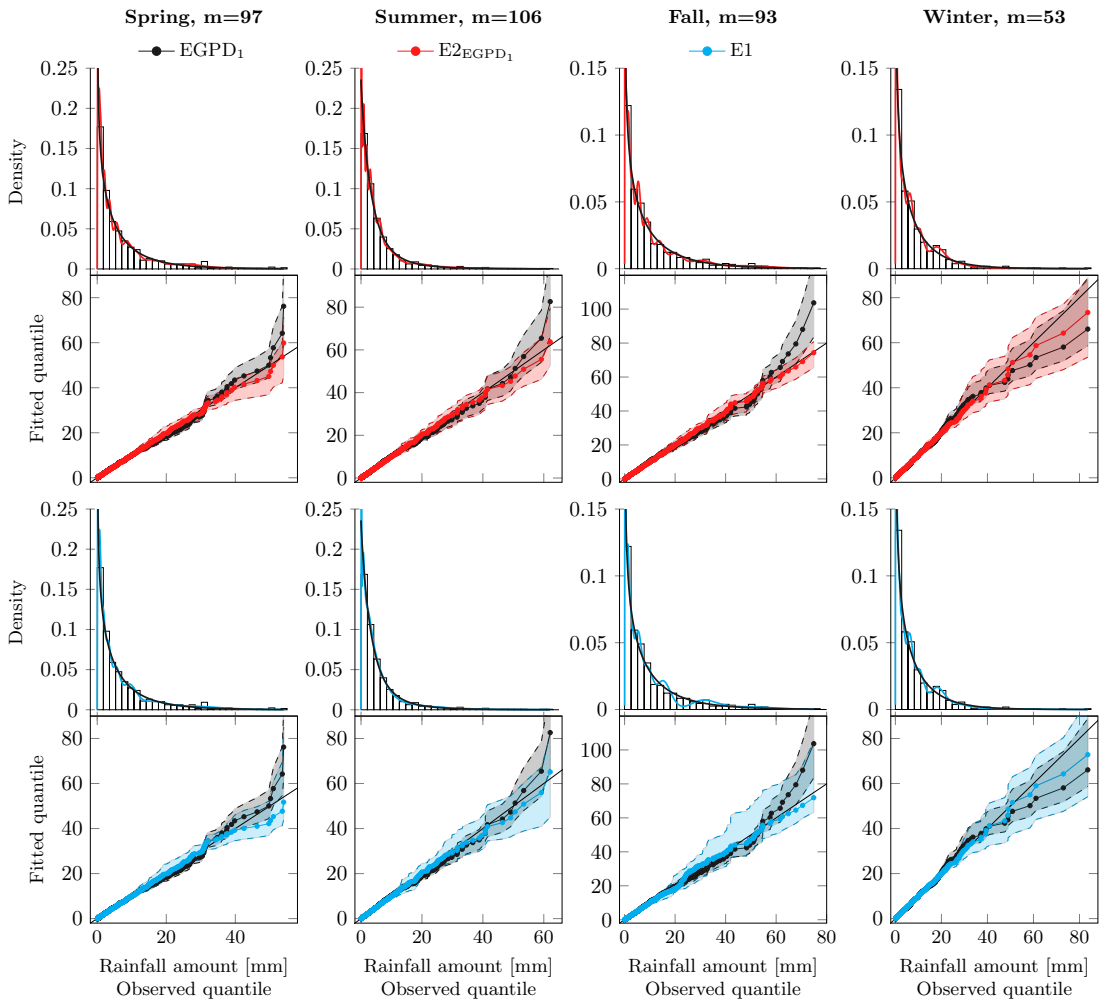

Figure 2.3.15: Histograms with the fitted densities and QQ-plots with the associated 95% confidence intervals for EGPD_1 and EGPD-BB with $m = m_{\text{LSCV}}$ for daily Durance rainfall data (1948-2010), for each of the four seasons. Two approaches are considered for the estimation of the EGPD-BB model: E1 and $\text{E2}_{\text{EGPD}_1}$

Consequently, for the Durance precipitation with a medium sample size dataset, the setting of the hyperparameter to $m = m_{\text{opt}}$ is, generally, a good option. We have estimation problems with the approach E1, but as was already stated in the theoretical part, these issues come from the likelihood optimization and not necessary from the choice of m.

To summarize, the EGPD-BB model yields good estimates for both Lyon and Durance rainfall datasets, and it outperforms the fit of EGPD_1 by adding more flexibility for the modeling of the bulk of the distribution.

Remark 2.3.5. The analysis presented in the case studies of this chapter was performed in both Matlab and R Software. On the one hand, the estimation of the EGPD-BB model was handled in Matlab by means of the Optimization Toolbox for nonlinear constraint optimization. More specifically, we used the *fmincon* function with the interior point algorithm. On the other hand, the fit of the EGPD_1 model was performed in R Software using the packages *mev*, particularly the *egp2* and *egp2.fit* functions.

Remark 2.3.6. The computational time, on a machine with an Intel Core 2 Duo 2.53 GHz processor, for an E1 estimation ranges from 15-35 seconds for Lyon case study and from 80-130 seconds for Durance one, depending on the season and, thus, the degree of the polynomial. There is not notable difference between E1 and $\mathrm{E2}_{\mathrm{EGPD}_1}$. However, $\mathrm{E2}_{\mathrm{grid}}$ is clearly much time consuming as the optimization problem must be run 30x30 times, but by using parallel computing we can decrease considerably this duration.

Chapter 4

Discussions, conclusions and perspectives

4.1 Discussions and conclusions

The work presented in this part of the thesis addresses statistical modeling of the entire-range of precipitation amounts. Due to the fact that precipitation data are heavily skewed to the right, establishing a probability distribution that provides a good fit has proven to be a challenging task. Different distributions, such as exponential, gamma, Weibull or lognormal have been considered as possible candidates, but without a great success in capturing well both the bulk and the tails of the distribution.

As the upper tail of the distribution holds crucial information that characterizes extreme events, many researchers focus only on the behavior of the largest rainfall amounts. The popular framework of EVT, more exactly the GPD, has been quickly adopted in this sense. However, besides the considerable reduction in sample size, a major drawback of this approach is the need of a threshold selection, *i.e,* the limit between moderate and large rainfall intensities. Defining this threshold is a delicate work in the field of EVT, since it has a major impact on the capability of the models in describing the extreme events.

While characterizing extreme events is essential, also one cannot totally neglect the small and moderate precipitation amounts. Several applications such as water resources management requires not only a clear understanding of extreme events, but rather a global assessment of the rainfall data. Various extreme mixture models have been proposed in the literature in this sense, but in this work we focus on the class of extended generalized Pareto distributions (EGPD) introduced by *Naveau et al.* (2016). This class of models is in compliance with the EVT for both lower and upper tails, and at the same time, it allows a smooth transition between the two ends without the need of selecting a threshold. More specifically, a proper model from the EGPD class has the lower tail (small values) behaving like a power function, while the upper tail (large values) is a GPD. The work that introduced this approach, proposed four possible models from the EGPD class based on parametric families, all of them complying with the two constraints on the tails. However, these parametric EGPD models seem to lack flexibility in modeling the bulk of the distribution.

To address this issue, we have proposed first a new semiparametric model that allows a more flexible fit for the bulk of the EGPD, namely, a transformed kernel (TK) density estimator based on an EGPD transformation. This model is obtained by transforming the data with the EGPD cdf, and then, estimating the density of the transformed data with a nonparametric kernel density estimator. We have proved that the theoretical constraints describing the behavior of a rainfall distribution, are not satisfied by this model for neither of the two tails. However, in

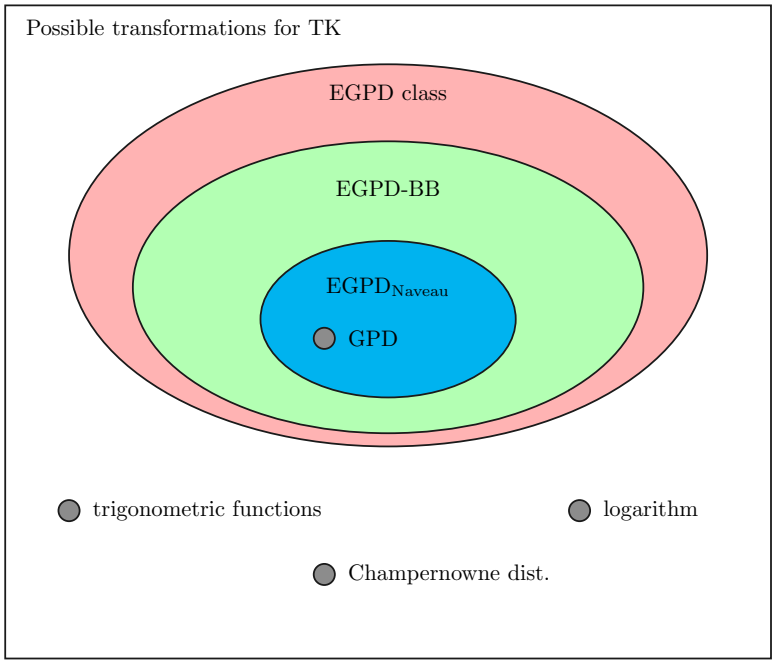

Figure 2.4.1: The general EGPD framework and its relationship with the proposed model: EGPD-BB and TK

practice the performance of the model seems not to be influenced by this aspect.

A second method that has been developed in this work, is in fact a new model from the general EGPD class, *i.e.*, we have considered a semiparametric EGPD based on Bernstein polynomials (EGPD-BB), more specifically a sparse mixture of beta densities. This model has proved to be a proper EGPD model as it complies with both the lower and the upper limiting constraints imposed by the EGPD class.

Figure 2.4.1 illustrates a diagram of how all these models interact within the general EGPD class and with the TK models. EGPD represents the general class of GPD extensions of the form $F(x) = G\{H_\xi(x)\}$, where H_ξ denotes the cdf of GPD. Depending on the choice of the G function, this class can generate several particular cases, such as EGPD-BB, EGPD$_{Naveau}$[1] or even a GPD. More specifically, EGPD$_{Naveau}$ represents a subclass of the EGPD-BB model proposed in this thesis, *i.e.*, each of the four parametric G functions are particular cases of the Bernstein density estimator. Moreover, given some specific parameters[2], both EGPD-BB and EGPD$_{Naveau}$ can yield the GP distribution. Furthermore, along with some commonly used transformations, such as logarithm, trigonometric function, different cdfs, the general EGPD class (with its subclasses) can also be used as a transformation function for the TK density estimator.

The performance of both proposed approaches, the transformed kernel (TK) and the Bernstein-beta EGPD (EGPD-BB) models, has been evaluated in case studies on simulated data, as well as on rainfall datasets. The general conclusion of these studies is that the proposed models add flexibility to the fit of the bulk of the distributions when compared to the EGPD$_{Naveau}$ models, without affecting the behavior of the two tails. Moreover, the sample size of the data was proved to be a very important factor in the performance of both models, *i.e.*, a larger sample size provided a more accurate estimation. This is not surprising as both

[1] the four parametric EGPD models introduced in *Naveau et al.* (2016), *i.e.*, EGPD$_{1,2,3,4}$
[2] *e.g.*, $m = 1$ for EGPD-BB, or $\kappa = 1$ for EGPD$_1$

models rely on nonparametric smoothing which is data driven.

An interesting remark regarding the TK model is related to the limiting constraints on the tails. We showed through the case studies on both simulated and rainfall data that, despite the fact that these limits were not satisfied, the fit of the tails was accurate even for a small sample size (*e.g.*, 300 observations), improving considerably for a larger one (*e.g.*, 1000 observations).

If we focus now on the EGPD-BB model, first, we have shown in one of the simulation studies that the weights identifiability is highly influenced by the inclusion of the GPD parameters in the estimation process. That is, the maximum likelihood estimator (MLE) had troubles when all parameters were estimated at once, yielding an inaccurate identifiability of the weights and variable estimations of the GPD parameters. This result motivated us to consider an alternative estimation method based on subproblems split, which improved the weights identification and also stabilized the estimations of the GPD parameters.

Another point of concern regarding the EGPD-BB model is the choice of the Bernstein polynomial degree m, *i.e.*, the number of components in the mixture. This hyperparameter is often compared with the bandwidth of a kernel density estimator, in the way that, a higher degree (smaller bandwidth) allows for more flexibility in smoothing a function, but attention must be paid as a too large degree can produce over-fitting. While the literature on the kernel bandwidth estimation is rather developed, there are not many studies related to Bernstein polynomial degree estimation. In this sense, we have mainly relied on the work of *Babu et al.* (2002), where it was showed that the optimal degree could be $m_{\text{opt}} = n/\log n$. However, we have determined in the simulation studies that this choice could lead to over-fitting for a large sample size, such as $n = 1000$. Therefore, in the application on rainfall data we also applied an alternative method for estimating m, *i.e.*, the least squares cross validation (LSCV). While, for small sample size data, LSCV and Babu's choice yielded the same setting, for larger samples, LSCV provided an improved estimation of the degree. However, LSCV is inconvenient to use in practice as it is computationally intensive.

Overall, we provide two different semiparametric methods that can accurately model the entire range of precipitation amounts without the need of selecting a threshold. While we managed to approach aspects such as, the optimal selection of the bandwidth or the polynomial degree, the identifiability analysis in mixture models, or parameter estimation through a penalized MLE, a lot of research and development is still required in this area. Some of the ideas that we did not get to approach in this thesis, but we consider they might be of interest, are described in the next section.

4.2 Perspectives

An important aspect that should be considered as future work, is the selection of the optimal degree of the Bernstein polynomial. The close connection of this degree with the bandwidth of the kernel density estimator, can provide ideas for new estimation methods. As approaches like LSCV are not convenient to use in practice since they are time consuming, simple and fast methods such as rule-of-thumb or plug-in selectors could be interesting to research.

Also, the inference of the EGPD-BB model is at the moment troublesome. Moreover, the two alternative estimation approaches that we proposed in this sense, *i.e.*, the split of the estimation process in two subproblems, are either computationally slow as we estimate over a grid (*i.e.*, E2_{grid} approach), or rely on the performance of another model such as the parametric EGPD (*i.e.*,E2_{EGPD} alternative). Therefore, finding an improved optimization methodology for the MLE or even a new inference method that can accurately estimate all the parameters at once would add value to the applicability of the EGPD-BB model.

Another interesting perspective, related in fact to the inference process, is to further pursue the work introduced by *Bunea et al.* (2010) regarding sparse mixture models on the real domain with l_1-penalization of the empirical risk (ER). We have briefly seen that the adaptation of this

method to a compact support such as the unit interval, particularly when considering the Bernstein-beta basis, is not straightforward and does not provide accurate estimates. In this sense, an option could be to step away from dictionaries composed of density functions, and consider orthogonal basis on $[0, 1]$, such as wavelets, orthogonal polynomials, etc.

Furthermore, one new study that would be worth pursuing is the TK's performance when the transformation cdf is not an EGPD_1 model, but rather the EGPD-BB. As EGPD based on Bernstein-beta density estimator showed an improved performance compared with the one based on parametric families (*e.g.*, the power function), the overall performance of TK might be enhanced. Another possible direction could be to assess the estimation efficiency of TK, when Bernstein-beta density estimator is considered for modeling the transformed data on the unit interval instead of the copula-based kernel. This perspective evolved from the fact that Bernstein-beta has a smaller asymptotic boundary bias and variance compared to the copula kernel, and thus, it might provide a better fit of the data at the end points of the $[0, 1]$ interval.

In addition to this, as in this work we focus only on the statistical modeling of precipitation amounts, a future work would be to couple the modeling of precipitation occurrences and amounts. Since accurate stochastic simulations of precipitation are required for assessment studies, like sensitivity of floods, erosion or crops models, incorporating EGPD in the stochastic models could lead to a significant improvement of their accuracy in reproducing extremes.

Finally, as the modeling of the entire-range of precipitation amounts at multiple sites is scarce in the literature at the moment, it would be interesting to develop a multivariate version of both TK and EGPD models. On the same line, the development of regional precipitation models based on EGPD could offer a more robust parameter estimation. More specifically, *Evin et al.* (2016) showed that a regional model can considerable improve the estimation of the GPD shape parameter.

Appendix A - Parameter estimation EGPD-BB

A.1 Algorithm - Fixed m

Step1: Estimate all parameters under all constraints

$$\begin{aligned} \hat{\theta}^{\text{MLE}} = \underset{\theta}{\text{argmin}} \quad & \{-\text{logL}(\theta|x)\} \\ \text{subject to} \quad & \omega \in \Delta^m \\ & \omega_{m,m} > 0 \\ & \sigma > 0 \\ & \xi \geq 0, \end{aligned}$$

where $\Delta^m = \{\omega \in \mathbb{R}^m : \omega_{k,m} \geq 0, \sum_{k=1}^m \omega_{k,m} = 1\}$

Step2: Threshold the estimated weights from Step 1

$$\hat{\omega}^{\text{MLE}}(\tau) = \left\{\hat{\omega}_{i,m}^{\text{MLE}} \cdot \mathbb{I}\left(\hat{\omega}_{i,m}^{\text{MLE}} \geq \tau\right)\right\}_{1 \leq i \leq m-1}$$

where $\mathbb{I}(\cdot)$ is the indicator function and τ is the optimal threshold value.

The optimal threshold τ is found as follows:

$$\text{GIC}(\tau_i) = -\text{logL}\left\{\hat{\omega}^{\text{MLE}}(\tau_i)\middle| x, \hat{\omega}_{m,m,}^{\text{MLE}}, \hat{\sigma}^{\text{MLE}}, \hat{\xi}^{\text{MLE}}\right\} + \lambda\sigma_\epsilon^2(\tau_i)M(\tau_i)$$
$$\tau_{opt} = \min_{\tau} GIC(\tau)$$

where $\lambda = 2\log(m)$, $M(\tau_i) = \sum_{k=1}^m \mathbb{I}\left(\hat{\omega}_{k,m}^{MLE-\tau_i} \neq 0\right)$ and $\sigma_\epsilon^2(\tau_i)$ is the variance of the error term (*i.e.*, deviation from the true model). As this value is unknown, we replace $\sigma_\epsilon^2(\tau_i)$ with the mean squared error obtained by using the initial non-thresholded model from the previous step, as a reference.

Step3: Readjust parameters using the sparse log-likelihood function $\text{logL}(\theta_{J(\tau)}|x)$

$$\begin{aligned} \hat{\theta}_{J(\tau)}^{\text{MLE}} = \underset{\theta_\tau}{\text{argmin}} \quad & \left\{-\text{logL}(\theta_{J(\tau)}|x)\right\} \\ \text{subject to} \quad & \omega_{J(\tau)} \in \Delta_{J(\tau)}^m \\ & \omega_{m,m} > 0 \\ & \sigma > 0 \\ & \xi \geq 0, \end{aligned}$$

where $\Delta_{J(\tau)}^m = \{\omega \in \mathbb{R}^m : \omega_{k,m} \geq 0, \sum_{k \in J(\tau)} \omega_{k,m} = 1\}$.

A.2 Algorithm - Fixed m, σ, ξ

Subproblem1: Fix GPD parameters and estimate the weights

Step 1: Estimate the weights ω

$$\begin{aligned} \hat{\theta}^{\text{MLE}} = \underset{\theta}{\text{argmin}} \quad & \left\{ -\text{logL}\left(\omega^{MLE} \middle| x, \sigma, \xi \right) \right\} \\ \text{subject to} \quad & \omega \in \Delta^m \\ & \omega_{m,m} > 0 \end{aligned}$$

where $\Delta^m = \{\omega \in \mathbb{R}^m : \omega_{k,m} \geq 0, \sum_{k=1}^{m} \omega_{k,m} = 1\}$

Step 2: Threshold the estimated weights from Step 1

$$\hat{\omega}^{\text{MLE}}(\tau) = \left\{ \hat{\omega}_{i,m}^{\text{MLE}} \cdot \mathbb{I}\left(\hat{\omega}_{i,m}^{\text{MLE}} \geq \tau \right) \right\}_{1 \leq i \leq m-1}$$

where $\mathbb{I}(\cdot)$ is the indicator function and τ is the optimal threshold value.
The optimal threshold τ is found as follows:

$$\begin{aligned} \text{GIC}(\tau_i) = -\text{logL}\left\{ \hat{\omega}^{\text{MLE}}(\tau_i) \middle| x, \hat{\omega}_{m,m,}^{\text{MLE}}, \hat{\sigma}^{\text{MLE}}, \hat{\xi}^{\text{MLE}} \right\} + \lambda \sigma_\epsilon^2(\tau_i) M(\tau_i) \\ \tau_{opt} = \min_{\tau} GIC(\tau) \end{aligned}$$

where $\lambda = 2\log(m)$, $M(\tau_i) = \sum_{k=1}^{m} \mathbb{I}\left(\hat{\omega}_{k,m}^{MLE-\tau_i} \neq 0 \right)$ and $\sigma_\epsilon^2(\tau_i)$ is the variance of the error term (*i.e.*, deviation from the true model). As this value is unknown, we replace $\sigma_\epsilon^2(\tau_i)$ with the mean squared error obtained by using the initial non-thresholded model from the previous step, as a reference.

Subproblem2: Knowing the sparsity pattern from Subproblem 1, optimize the sparse log-likelihood and find the estimates for all parameter.

$$\begin{aligned} \hat{\theta}_{J(\tau)}^{\text{MLE}} = \underset{\theta_\tau}{\text{argmin}} \quad & \left\{ -\text{logL}(\theta_{J(\tau)} | x) \right\} \\ \text{subject to} \quad & \omega_{J(\tau)} \in \Delta_{J(\tau)}^m \\ & \omega_{m,m} > 0 \\ & \sigma > 0 \\ & \xi \geq 0, \end{aligned}$$

where $\Delta_{J(\tau)}^m = \{\omega \in \mathbb{R}^m : \omega_{k,m} \geq 0, \sum_{k \in J(\tau)} \omega_{k,m} = 1\}$.

A.3 Algorithm - Fixed m and (σ, ξ) from a grid

Subproblem1: Fix GPD parameters and estimate the weights

Step 1: Iterate Subproblem 1 from *Algorithm - Fixed* m, σ, ξ for all the pairs (σ, ξ) from the grid

Step 2: Choose the optimal pair (σ, ξ) by minimizing BIC index

$$\begin{aligned} \mathrm{BIC}(\sigma, \xi) &= -2\mathrm{logL}\left(\hat{\omega}^{\mathrm{MLE}}(\tau)\middle| x, \sigma, \xi\right) + M(\sigma, \xi) \cdot log(n) \\ (\sigma_{opt}, \xi_{opt}) &= \min_{(\sigma,\xi)} \mathrm{BIC}(\sigma, \xi) \end{aligned}$$

where $M(\sigma, \xi) = \sum_{k=1}^{m} \mathbb{I}\left(\hat{\omega}_{k,m}^{\mathrm{MLE}} \neq 0\right) + 2$.

Subproblem2: Knowing the sparsity pattern from Subproblem 1, optimize the sparse log-likelihood and find the estimates for all parameter.

$$\begin{aligned} \hat{\theta}_{J(\tau)}^{\mathrm{MLE}} = \underset{\theta_\tau}{\mathrm{argmin}} \quad & \left\{-\mathrm{logL}(\theta_{J(\tau)}|x)\right\} \\ \text{subject to} \quad & \omega_{J(\tau)} \in \Delta_{J(\tau)}^{m} \\ & \omega_{m,m} > 0 \\ & \sigma > 0 \\ & \xi \geq 0, \end{aligned}$$

where $\Delta_{J(\tau)}^{m} = \{\omega \in \mathbb{R}^m : \omega_{k,m} \geq 0, \sum_{k \in J(\tau)} \omega_{k,m} = 1\}$.

Appendix B - Additional results for the EGPD-BB model estimation

Simulation setting:

- $m_{\text{true}} = 50$, considered known,
- $\omega_{\text{true}} = [\omega_{15} = 0.25, \omega_{25} = 0.2, \omega_{35} = 0.25, \omega_{45} = 0.25, \omega_{50} = 0.05]$, the remaining 45 elements of ω_{true} vector out of the 50 are zero; this means that we have only $M = 5$ non-null weights,

Table B.1: MIAE, hit rate and mean cardinality $\bar{M}$ of non-null weights for 50 replicates

	MIAE	**hit rate**	$\bar{M}$
MLE	0.128	88%	4.88
ERM	0.284	0%	9.10

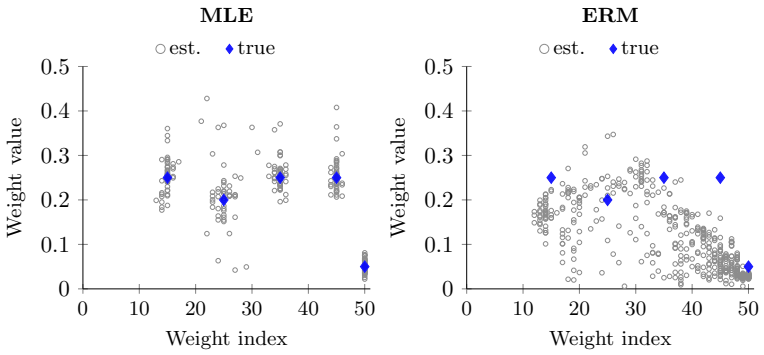

Figure B.1: Estimated weights for 50 replicates

References

H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.

R. Alemany, C. Bolancé, and M. Guillén. A nonparametric approach to calculating value-at-risk. *Insurance: Mathematics and Economics*, 52(2):255–262, 2013.

B. A. Amisigo and N. C. van de Giesen. Using a spatio-temporal dynamic state-space model with the EM algorithm to patch gaps in daily riverflow series. *Hydrology and Earth System Sciences*, 9(3):209–224, 2005.

S. Apipattanavis, G. Podestá, B. Rajagopalan, and R.W. Katz. A semiparametric multivariate and multisite weather generator. *Water Resources Research*, 43(11), 2007.

G.J. Babu, A.J. Canty, and Y.P. Chaubey. Application of Bernstein polynomials for smooth estimation of a distribution and density function. *Journal of Statistical Planning and Inference*, 105(2):377–392, 2002.

A. Bárdossy and G. Pegram. Infilling missing precipitation records - A comparison of a new copula-based method with other techniques. *Journal of Hydrology*, 519:1162–1170, 2014.

C.N. Behrens, H.F. Lopes, and D. Gamerman. Bayesian analysis of extreme events with threshold estimation. *Statistical Modelling*, 4(3):227–244, 2004.

S. Bercu and F. Proïa. A SARIMAX coupled modelling applied to individual load curves intraday forecasting. *Journal of Applied Statistics*, 40(6):1333–1348, 2013.

S. Bernstein. Démonstration du théorème de Weierstrass fondée sur le calcul des probabilités. *Communications de la Société mathématique de Kharkow*, 13(1):1–2, 1912.

D.P. Bertsekas and J.N. Tsitsiklis. *Introduction to probability*. Athena Scientific, Belmont, 2nd edition, 2008.

C. Bolancé, M. Guillén, and J.P. Nielsen. Inverse beta transformation in kernel density estimation. *Statistics & Probability Letters*, 78(13):1757–1764, 2008.

T. Bouezmarni and J.M. Rolin. Bernstein estimator for unbounded density function. *Journal of Nonparametric Statistics*, 19(3):145–161, 2007.

A.W. Bowman, P. Hall, and D.M. Titterington. Cross-validation in nonparametric estimation of probabilities and probability densities. *Biometrika*, 71(2):341–351, 1984.

G.E.P. Box and G.M. Jenkins. *Time series analysis: forecasting and control*. Holden-Day, 1976.

G.E.P. Box and D.A. Pierce. Distribution of residual autocorrelations in autoregressive-integrated moving average time series models. *Journal of the American Statistical Association*, 65(332):1509–1526, 1970.

T. Buch-Kromann, M. Englund, J. Gustafsson, J. Perch Nielsen, and F. Thuring. Non-parametric estimation of operational risk losses adjusted for under-reporting. *Scandinavian Actuarial Journal*, 2007(4):293–304, 2007.

T. Buch-Larsen, J.P. Nielsen, M. Guillén, and C. Bolancé. Kernel density estimation for heavy-tailed distributions using the champernowne transformation. *Statistics*, 39(6):503–516, 2005.

F. Bunea, A.B. Tsybakov, M.H. Wegkamp, and A. Barbu. SPADES and mixture models. *The Annals of Statistics*, 38(4):2525–2558, 2010.

J. Carreau and Y. Bengio. A hybrid Pareto model for asymmetric fat-tailed data: the univariate case. *Extremes*, 12(1):53–76, 2008.

J. Carreau and Y. Bengio. A hybrid pareto mixture for conditional asymmetric fat-tailed distributions. *IEEE Transactions on Neural Networks*, 20(7):1087–1101, 2009.

Enrique Castillo, Ali S. Hadi, N. Balakrishnan, and Jose M. Sarabia. *Extreme value and related models with applications in engineering and science*. Wiley, 2005.

D.G. Champernowne. A model of income distribution. *The Economic Journal*, 63(250):318, 1953.

S.X. Chen. Beta kernel estimators for density functions. *Computational Statistics & Data Analysis*, 31(2):131–145, 1999.

S.X. Chen. Probability density function estimation using gamma kernels. *Annals of the Institute of Statistical Mathematics*, 52(3):471–480, 2000.

S. Coles. *An introduction to statistical modeling of extreme values*. Springer, 2001.

D. Cooley, D. Nychka, and P. Naveau. Bayesian spatial modeling of extreme precipitation return levels. *Journal of the American Statistical Association*, 102(479):824–840, 2007.

L. Devroye and L. Györfi. *Nonparametric density estimation: the L1 view*. Wiley, 1985.

F.F. do Nascimento, D. Gamerman, and H.F. Lopes. A semiparametric Bayesian approach to extreme value estimation. *Statistics and Computing*, 22(2):661–675, 2011.

R. Duin. On the choice of smoothing parameters for Parzen estimators of probability density functions. *IEEE Transactions on Computers*, C-25(11):1175–1179, 1976.

A.A. Elshorbagy, U.S. Panu, and S.P. Simonovic. Group-based estimation of missing hydrological data: I. Approach and general methodology. *Hydrological Sciences Journal*, 45(6), 2000.

A.A. Elshorbagy, S.P. Simonovic, and U.S. Panu. Estimation of missing streamflow data using principles of chaos theory. *Journal of Hydrology*, 255:123–133, 2002.

P. Embrechts, C. Klüppelberg, and T. Mikosch. *Modelling extremal events for insurance and finance*. Springer, 1997.

K. Eng, G.D. Tasker, and P.C.D. Milly. An analysis of region-of-influence methods for flood regionalization in the Gulf-Atlantic rolling plains. *Journal of the American Water Resources Association*, 41(1):135–143, 2005.

G. Evin, J. Blanchet, E. Paquet, F. Garavaglia, and D. Penot. A regional model for extreme rainfall based on weather patterns subsampling. *Journal of Hydrology*, 541:1185–1198, 2016.

R.T. Farouki. The Bernstein polynomial basis: A centennial retrospective. *Computer Aided Geometric Design*, 29(6):379–419, 2012.

R.A. Fisher and L.H.C. Tippett. Limiting forms of the frequency distribution of the largest or smallest member of a sample. *Mathematical Proceedings of the Cambridge Philosophical Society*, 24(02):180, 1928.

A. Frigessi, O. Haug, and H. Rue. A dynamic mixture model for unsupervised tail estimation without threshold selection. *Extremes*, 5(3):219–235, 2002.

E.M. Furrer and R.W. Katz. Improving the simulation of extreme precipitation events by stochastic weather generators. *Water Resources Research*, 44(12), 2008.

F. Garavaglia, J. Gailhard, E. Paquet, M. Lang, R. Garcon, and P. Bernardara. Introducing a rainfall compound distribution model based on weather patterns sub-sampling. *Hydrology and Earth System Sciences*, 14(6):951–964, 2010.

S. Ghosal. Convergence rates for density estimation with Bernstein polynomials. *The Annals of Statistics*, 29(5):1264–1280, 2001.

B. Gnedenko. Sur la distribution limite du terme maximum d'une serie aleatoire. *The Annals of Mathematics*, 44(3):423, 1943.

F. Gottardi, C. Obled, J. Gailhard, and E. Paquet. Statistical reanalysis of precipitation fields based on ground network data and weather patterns: Application over French mountains. *Journal of Hydrology*, 432:154–167, 2012.

J.B. Greenhouse, R.E. Kass, and R.S. Tsay. Fitting nonlinear models with ARMA errors to biological rhythm data. *Statistics in Medicine*, 6(2):167–183, 1987.

Z. Guan. Efficient and robust density estimation using Bernstein type polynomials. *Journal of Nonparametric Statistics*, 28(2):250–271, 2016.

D. Gujarati and D. Porter. *Basic Econometrics*. McGraw Hill, New York, 5th edition, 2008.

H.V. Gupta, H. Kling, K.K. Yilmaz, and G.F. Martinez. Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *Journal of Hydrology*, 377(1-2):80–91, 2009.

P. Gyau-Boakye and G.A. Schultz. Filling gaps in runoff time series in West Africa. *Hydrological Sciences Journal*, 39(6):621–636, 1994.

C.L. Harvey, H. Dixon, and J. Hannaford. An appraisal of the performance of data-infilling methods for application to daily mean river flow records in the UK. *Hydrology Research*, 43 (5):618–636, 2012.

R.M. Hirsch. An evaluation of some record reconstruction techniques. *Water Resources Research*, 15(6):1781–1790, 1979.

K.-L. Hsu, H.V. Gupta, and S. Sorooshian. Artificial neural network modeling of the rainfall-runoff process. *Water Resources Research*, 31(10):2517–2530, 1995.

E. Imbeaux. *La Durance: régime, crues et inondations*. Vve Ch. Dunod, Paris, 1892.

Y. Ji, C. Wu, P. Liu, Ji. Wang, and K.R. Coombes. Applications of beta-mixture models in bioinformatics. *Bioinformatics*, 21(9):2118–2122, 2005.

M.C. Jones. Simple boundary correction for kernel density estimation. *Statistics and Computing*, 3(3):135–146, 1993.

M.C. Jones and D.A. Henderson. Miscellanea kernel-type density estimation on the unit interval. *Biometrika*, 94(4):977–984, 2007.

Y. Kakizawa. Bernstein polynomial probability density estimation. *Journal of Nonparametric Statistics*, 16(5):709–729, 2004.

H.M. Kang and F. Yusof. Homogeneity tests on daily rainfall series. *Int. J. Contemp. Math. Sciences*, 7(1):9–22, 2012.

R.W. Katz. Precipitation as a chain-dependent process. *Journal of Applied Meteorology*, 16(7): 671–676, 1977.

R.W. Katz and M.B. Parlange. Generalizations of chain-dependent processes: application to hourly precipitation. *Water Resources Research*, 31(5):1331–1341, 1995.

R.W. Katz, M.B. Parlange, and P. Naveau. Statistics of extremes in hydrology. *Advances in Water Resources*, 25(8):1287–1304, 2002.

L. Kaufman and P.J. Rousseeuw. *Finding groups in gata: an introduction to cluster analysis*. Wiley Series in Probability and Statistics, Wiley, 1990.

M. Khalil, U.S. Panu, and W.C. Lennox. Groups and neural networks based streamflow data infilling procedures. *Journal of Hydrology*, 241:153–176, 2001.

J.-W. Kim and Y.A. Pachepsky. Reconstructing missing daily precipitation data using regression trees and artificial neural networks for SWAT streamflow simulation. *Journal of Hydrology*, 394(3-4):305–314, 2010.

Y. Kim, S. Kwon, and H. Choi. Consistent model selection criteria on high dimensions. *Journal of Machine Learning Research*, 13:1037–1057, 2012.

A. Kuentz. *Un siècle de variabilité hydro-climatique sur le bassin de la Durance*. PhD thesis, AgroParisTech, 2013.

A. Kuentz, T. Mathevet, J. Gailhard, C. Perret, and V. Andréassian. Over 100 years of climatic and hydrologic variability of a Mediterranean and mountainous watershed: the Durance River. In *Cold and Mountain Region Hydrological Systems Under Climate Change: Towards Improved Projections Proceedings*, pages 19–25, Gothenburg, Sweden, 2013. International Association of Hydrological Sciences.

A. Kuentz, T. Mathevet, D. Coeur, C. Perret, J. Gailhard, L. Guérin, Y. Gash, and V. Andréassian. Historical hydrometry and hydrology of the Durance river watershed. *La Houille Blanche*, (4):57–63, 2014.

M.H. Kutner, C. Nachtsheim, and J. Neter. *Applied linear regression models*. McGraw-Hill Irwin, 4th edition, 2004.

D. Kwiatkowski, P.C.B. Phillips, P. Schmidt, and Y. Shin. Testing the null hypothesis of stationarity against the alternative of a unit root. *Journal of Econometrics*, 54(1-3):159–178, 1992. ISSN 6.

A. Leblanc. A bias-reduced approach to density estimation using Bernstein polynomials. *Journal of Nonparametric Statistics*, 22(4):459–475, 2010.

A. Leblanc. On estimating distribution functions using Bernstein polynomials. *Annals of the Institute of Statistical Mathematics*, 64(5):919–943, 2012a.

A. Leblanc. On the boundary properties of Bernstein polynomial estimators of density and distribution functions. *Journal of Statistical Planning and Inference*, 142(10):2762–2778, 2012b.

D. Lee, W.K. Li, and T.S.T. Wong. Modeling insurance claims via a mixture exponential model combined with peaks-over-threshold approach. *Insurance: Mathematics and Economics*, 51 (3):538–550, 2012.

E.L. Lehmann and H.J.M. D'Abrera. *Nonparametrics: statistical methods based on ranks*. Springer, 2006.

P. Li, S.S. Rangapuram, and M. Slawski. Methods for sparse and low-rank recovery under simplex constraints. 2016.

A.E. MacDonald. *Extreme value mixture modelling with medical and industrial applications*. PhD thesis, University of Canterbury, New Zealand, 2011.

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

A.E. MacDonald, C.J. Scarrott, and D.S. Lee. Boundary correction, consistency and robustness of kernel densities using extreme value theory. 2011b.

G.J. MacLachlan and D. Peel. *Finite mixture models*. Wiley, New York, 2000.

S.G. Makridakis, S.C. Wheelwright, and R.J. Hyndman. *Forecasting: methods and mpplications*. Wiley, 1998.

N. Markovich. *Nonparametric analysis of univariate heavy-tailed data: research and practice*. John Wiley & Sons, London, 2007.

N. Markovitch and U.R. Krieger. Nonparametric estimation of long-tailed density functions and its application to the analysis of World Wide Web traffic. *Performance Evaluation*, 42 (2):205–222, 2000.

B.V.M. Mendes and H.F. Lopes. Data driven estimates for mixtures. *Computational Statistics & Data Analysis*, 47(3):583–598, 2004.

S.-P. Miaou. A stepwise time series regression procedure for water demand model identification. *Water Resources Research*, 26(9):1887–1897, 1990.

D.C. Montgomery, E.A. Peck, and G.G. Vining. *Introduction to linear regression analysis*. John Wiley & Sons, Inc., New Jersey, 5th edition, 2012.

S. Nadarajah. Extremes of daily rainfall in west central Florida. *Climatic Change*, 69(2-3): 325–342, 2005.

J.E. Nash and J.V. Sutcliffe. River flow forecasting through conceptual models part I - A discussion of principles. *Journal of Hydrology*, 10(3):282–290, 1970.

P. Naveau, R. Huser, P. Ribereau, and A. Hannart. Modeling jointly low, moderate, and heavy rainfall intensities without a threshold selection. *Water Resources Research*, 52(4):2753–2769, 2016.

F.J. Nogales, J. Contreras, A.J. Conejo, and R. Espinola. Forecasting next-day electricity prices by time series models. *IEEE Transactions on Power Systems*, 17(2):342–348, 2002.

A. Pankratz. *Forecasting with dynamic regression models*. Wiley-Interscience, 1991.

U.S. Panu, M. Khalil, and A.A. Elshorbagy. Streamflow data infilling techniques based on concepts of groups and neural networks. In *Artificial Neural Networks in Hydrology*, pages 235–258. Springer Netherlands, 2000.

I. Papastathopoulos and J.A. Tawn. Extended generalised Pareto models for tail estimation. *Journal of Statistical Planning and Inference*, 143(1):131–143, 2013.

E. Parzen. On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33(3):1065–1076, 1962.

E. Passow. Polynomials with positive coefficients: uniqueness of best approximation. *Journal of Approximation Theory*, 21:352–355, 1977.

S. Petrone. Bayesian density estimation using Bernstein polynomials. *Canadian Journal of Statistics*, 27(1):105–126, 1999.

G.M. Phillips. *Interpolation and approximation by polynomials*. Springer-Verlag New York, 2003.

M. Pilanci, L.E. Ghaoui, and V. Chandrasekaran. Recovery of sparse probability measures via convex programming. In *Advances in Neural Information Processing Systems 25*, pages 2420–2428, 2012.

H. Raman, S. Mohan, and P. Padalinathan. Models for extending streamflow data: a case study. *Hydrological Sciences Journal*, 40(3):381–393, 1995.

C.W. Richardson. Stochastic simulation of daily precipitation, temperature, and solar radiation. *Water Resources Research*, 17(1):182–190, feb 1981.

B.D. Ripley. Time series in R 1.5.0. *The R Journal*, 2(2):2–7, 2002.

M. Rosenblatt. Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*, 27(3):832–837, 1956.

P.J. Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987.

M. Rudemo. Empirical choice of histograms and kernel density estimators. *Scandinavian Journal of Statistics*, 9(2):65–78, 1982.

S.E. Said and D.A. Dickey. Testing for unit roots in autoregressive-moving average models of unknown order. *Biometrika*, 71(3):599–607, 1984.

T. Schneider. Analysis of incomplete climate data: estimation of mean values and covariance matrices and imputation of missing values. *Journal of Climate*, 14(5):853–871, 2001.

E.F. Schuster. Incorporating support constraints into nonparametric estimators of densities. *Communications in Statistics - Theory and Methods*, 14(5):1123–1136, 1985.

G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

B.W. Silverman. *Density estimation for statistics and data analysis*. Chapman and Hall, 1986.

R.D. Stern and R. Coe. A model fitting analysis of daily rainfall data. *Journal of the Royal Statistical Society. Series A (General)*, 147(1):1, 1984.

Suhartono. Time series forecasting by using seasonal autoregressive integrated moving average: subset, multiplicative or additive model. *Journal of Mathematics and Statistics*, 7(1):20–27, 2011.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, 58(1):267–288, 1996.

RS Tsay. Regression models with time series errors. *Journal of the American Statistical Association*, 79(385):118–124, 1984.

S.I. Vagropoulos, G.I. Chouliaras, E.G. Kardakos, C.K. Simoglou, and A.G. Bakirtzis. Comparison of SARIMAX, SARIMA, modified SARIMA and ANN-based models for short-term PV generation forecasting. In *2016 IEEE International Energy Conference (ENERGYCON)*, pages 1–6. IEEE, 2016.

R. Vezzoli, S. Pecora, E. Zenoni, and F. Tonelli. Data analysis to detect inhomogeneity, change points, trends in observations: an application to Po river discharge extremes. *CMCC Research Paper*, 138, 2012.

R.A. Vitale. A Bernstein polynomial approach to density function estimation. *Statistical inference and related topics*, 2:87–99, 1975.

E. Voronovskaja. Determination de la forme asymptotique d'approximation des fonctions par les polynomes de M. Bernstein. *CR Acad. Sci. URSS*, 79:79–85, 1932.

M. Vrac and P. Naveau. Stochastic downscaling of precipitation: from dry events to heavy rainfalls. *Water Resources Research*, 43(7):1–13, 2007.

M. Vrac, M. Stein, and K. Hayhoe. Statistical downscaling of precipitation through nonhomogeneous stochastic weather typing. *Climate Research*, 34(3):169–184, 2007.

J.R. Wallis, D.P. Lettenmaier, and E.F. Wood. A daily hydroclimatological data set for the continental United States. *Water Resources Research*, 27(7):1657–1663, 1991.

M.P. Wand and M.C. Jones. *Kernel smoothing*. Chapman & Hall, 1995.

M.P. Wand, J.S. Marron, and D. Ruppert. Transformations in density estimation. *Journal of the American Statistical Association*, 86(414):343, jun 1991.

J.B. Wijngaard, A.M.G. Klein Tank, and G.P. Konnen. Homogeneity of 20th century European daily temperature and precipitation series. *International Journal of Climatology*, 23(6):679–692, 2003.

D.S. Wilks. Conditioning stochastic daily precipitation models on total monthly precipitation. *Water Resources Research*, 25(6):1429–1439, 1989.

D.S. Wilks. Interannual variability and extreme-value characteristics of several stochastic daily precipitation models. *Agricultural and Forest Meteorology*, 93(3):153–169, 1999.

D.S. Wilks. *Statistical methods in the atmospheric sciences*. Academic Press, 2011.

C.A. Woodhouse, S.T. Gray, and D.M. Meko. Updated streamflow reconstructions for the Upper Colorado River Basin. *Water Resources Research*, 42(5):1–16, 2006.

D.A. Woolhiser, G.G.S. Pegram, D.A. Woolhiser, and G.G.S. Pegram. Maximum likelihood estimation of fourier coefficients to describe seasonal variations of parameters in stochastic daily precipitation models. *Journal of Applied Meteorology*, 18(1):34–42, 1979.

L. Yang and J.S. Marron. Iterated transformation-kernel density estimation. *Journal of the American Statistical Association*, 94(446):580–589, 1999.

C. Yozgatligil and C. Yazici. Comparison of homogeneity tests for temperature using a simulation study. *International Journal of Climatology*, 36(1):62–81, 2016.

X. Zhao, C. Scarrott, L. Oxley, and M. Reale. Extreme value modelling for forecasting market crisis impacts. *Applied Financial Economics*, 20(1-2):63–72, 2010.

W. Zucchini and P.T. Adamson. *The occurrence and severity of droughts in South Africa*. South Africa Water Research Commission, 1984.