Normandie Université

UNIVERSITÉ LE HAVRE NORMANDIE

THÈSE

Pour obtenir le diplôme de doctorat

Spécialité **MATHEMATIQUES**

Préparée au sein de l'**Université Le Havre Normandie**

Extreme Quantile Regression Based on Conditional GEV Models and Generalized Random Forests

Présentée et soutenue par

VIDAGBANDJI MAHUTIN LUCIEN

Thèse soutenue le 20/03/2026

devant le jury composé de :

M. BERRED ALEXANDRE	Professeur des Universités - ULHN - Université Le Havre Normandie	Directeur de thèse
M. BERTELLE CYRILLE	Professeur des Universités - ULHN - Université Le Havre Normandie	Co-directeur de thèse
M. NGATCHOU WANDJI JOSEPH	Professeur des Universités - Université de Lorraine	Président du jury
M. AMANTON LAURENT	Maître de conférences - ULHN - Université Le Havre Normandie	Membre du jury
MME BLANKE DELPHINE	Professeur des Universités - UNIVERSITE AVIGNON PAYS DU VAUCLUSE	Membre du jury
MME DABO SOPHIE	Professeur des Universités - Universite de Lille	Membre du jury
M. DOMBRY CLÉMENT	Professeur des Universités - UNIVERSITE DE FRANCHE-COMTE	Rapporteur du jury
MME GAYRAUD GHISLAINE	Professeur des Universités - UNIV TECHNOLOGIE COMPIEGNE UTC COMPIEGNE	Rapporteur du jury

Thèse dirigée par **BERRED ALEXANDRE** (Laboratoire de Mathématiques Appliquées du Havre) et **BERTELLE CYRILLE** (LABORATOIRE D'INFORMATIQUE DE TRAITEMENT DE L'INFORMATION ET DES SYSTEMES)

MIIΣ

LMAH

À ma famille, et spécialement à mes frères et sœurs,
cette thèse vous est dédiée. Qu'elle soit pour vous une source d'inspiration et d'émulation.

Remerciements

Au terme de ce travail de thèse du doctorat, il m'est particulièrement important d'exprimer ma profonde reconnaissance à l'ensemble des personnes qui ont contribué, de près ou de loin, à sa réalisation. Par leur encadrement, leurs conseils, leur soutien et leurs encouragements, elles ont largement participé à son aboutissement.

Je souhaite, en premier lieu, adresser mes sincères remerciements à mes directeurs de thèse, M. Alexandre Berred, M. Cyrille Bertelle et M. Laurent Amanton. Je les remercie pour la confiance qu'ils m'ont accordée en m'accueillant au sein de ce projet de recherche, ainsi que pour la qualité de leur encadrement. Leur disponibilité, leurs nombreuses relectures et leurs conseils avisés ont été essentiels à la progression et à l'aboutissement de ces travaux. Je leur suis particulièrement reconnaissant pour leurs qualités humaines, leur gentillesse à mon égard et l'exigence scientifique qu'ils ont su transmettre tout au long de cette thèse. À leurs côtés, point de stress persistant : la thèse fut une succession de moments de joie et d'instants plus éprouvants, mais grâce à eux, j'ai pu traverser ces périodes de turbulence avec courage et détermination. Merci d'avoir su trouver les mots pour maintenir mon moral dans les moments difficiles. Je me souviendrai non seulement de leurs précieux conseils, mais aussi, et surtout, de leurs blagues légendaires.

Je tiens à exprimer mes plus sincères remerciements à l'ensemble des membres du jury pour l'intérêt qu'ils ont porté à ce travail. Je remercie tout particulièrement les rapporteurs, M. Clément Dombry et Mme. Ghislaine Gayraud, pour l'attention qu'ils ont consacrée à la lecture de ce manuscrit, ainsi que pour leurs remarques et suggestions pertinentes, qui ont contribué à en améliorer la qualité. Je remercie également M. Joseph W. Ngatchou, président du jury, ainsi que Mme Sophie Dabo et Mme Delphine Blanke, d'avoir accepté de participer à ce jury et pour le temps précieux qu'ils ont consacré à l'évaluation de ces travaux. Leurs expertises respectives ont constitué un apport significatif à l'appréciation de cette thèse.

Je souhaite également remercier les membres de mon comité de suivi de thèse, Nathalie Corson et Sophie Dabo, pour leurs recommandations et leurs conseils avisés tout au long de ces années de doctorat. Mes remerciements s'adressent aussi aux enseignants qui m'ont accompagné dans mes premières expériences d'enseignement, notamment Valentina Lanza, Nathalie Corson, David Monceau et Cédric Joncour, pour leur disponibilité et leur accompagnement. Je tiens enfin à remercier l'ensemble des membres du laboratoire LMAH pour l'environnement de travail, à la fois stimulant et bienveillant, dont j'ai pu bénéficier. Je leur suis également reconnaissant pour les moments de convivialité partagés, qui ont contribué à rendre ces années

de doctorat particulièrement enrichissantes.

J'ai eu l'opportunité de présenter mes travaux lors de plusieurs conférences, tant nationales qu'internationales. Ces expériences ont été rendues possibles grâce aux soutiens financiers dont j'ai bénéficié, grâce au dynamisme de mes directeurs de thèse. Je remercie à ce titre Arnaud Ducrot, directeur du laboratoire LMAH, ainsi qu'Éric Sanlaville, directeur du LITIS (site du Havre), pour leur appui. Je remercie également l'UFR ST de l'Université Le Havre Normandie ainsi que le LabCom SmartLogiLab pour les financements accordés, qui ont contribué au bon déroulement de mes déplacements scientifiques et à l'enrichissement de mon parcours doctoral.

Je souhaite adresser mes remerciements à mes collègues doctorants de LMAH et de LITIS, avec qui j'ai eu le plaisir de partager ces années de thèse, pour les échanges scientifiques, l'entraide et les moments de convivialité. Merci pour les déjeuners partagés et les matchs organisés. Merci à mes anciens collègues de bureau, Irmand Mikiela et Alexandre Thorel, pour leur accueil chaleureux et leurs précieux conseils lors de mes premières années de thèse. Je remercie tout particulièrement mes nouveaux collègues de bureau et ceux du bureau d'à côté. Je tiens d'abord à remercier Abdeltif, mon « jumeau » de thèse, pour les discussions et pour avoir géré l'organisation de la salle du pot avec la famille. I would also like to thank the other PhD students who helped the family prepare the reception room. Thank you, dear colleagues Li, Wang, and Wenran. Je tiens également à remercier mon collègue Kasse Mamadou pour ses conseils, et surtout pour le bissap et la bouillie offerts pour le pot de thèse. Thank you very much, Quang-Vinh Tran, for the gift you brought me back from Vietnam. Kamal, Gouled, Diop, Pierre, Saïd et Abdoul, je ne vous oublie pas : merci à vous pour les moments conviviaux passés ensemble.

Je tiens à remercier mes amis avec qui j'ai fait le master et avec qui je continue de discuter de nos sujets de recherche respectifs, ce qui me permet d'apprendre un peu sur d'autres thèmes. Merci à Régis, Roland et Auguste pour votre réactivité quand on propose de dîner ensemble lors de mes passages à Paris. Merci à Grace Mayala pour les discussions intéressantes sur la méthode de forêt aléatoire appliquée à des données déséquilibrées. Je remercie ma collègue de tous les instants, Innocentia, merci pour ta présence à ma soutenance. Je tiens également à remercier les amis que j'ai rencontrés en conférence et avec qui je continue de discuter de tout et de recherche — principalement la communauté « extrêmes » rencontrée à Chapel Hill. Ils sont nombreux, je vous remercie tous pour les discussions et les moments passés ensemble. Je remercie mon ami depuis l'enfance, devenu avec le temps mon meilleur ami, mon frère Hugues Midingoyi. Je te remercie d'avoir contribué à l'organisation de ma soutenance, même depuis le Canada. Je tiens à remercier ici des personnes qui ont eu une importance capitale pour mon bien-être et qui ne cessent de m'encourager chaque jour. Merci à ma grande, que j'appelle docteure Anna. Merci pour les encouragements et les moments inoubliables passés ensemble ; j'espère qu'on fêtera aussi ta soutenance de thèse en santé, comme tu l'as toujours

souhaité. Merci d'être là. Je tiens aussi à remercier Véronique, qui porte le même prénom que moi : merci pour ta présence et pour les moments passés ensemble.

Il est certain que si je n'avais pas reçu autant d'amour des miens, rien de tout cela n'aurait été possible. Je tiens à remercier mes parents pour leur soutien inestimable. Mon père et ma mère, je vous remercie pour votre soutien. Merci à mes frères et sœurs : Richard, Clarisse, Patrice, Noëllie et Clémentine. Je trouve désormais une raison pour justifier mes absences répétées. Je tiens à remercier mon oncle — mon deuxième papa — ainsi que mes cousins et cousines pour leur soutien, et merci pour votre déplacement le jour de ma soutenance. Oui, je fais partie de la famille des docteurs, la famille Vidagbandji ; nous allons prochainement fêter d'autres soutenances de thèse à Paris, puis à Clermont-Ferrand, et je serai présent. Merci tout particulièrement à ma sœur Chimène pour sa disponibilité sans faille dans l'organisation de mes pots de soutenance. Il suffit de la contacter, et elle gère le reste.

Enfin, je tiens à exprimer du plus profond de mon cœur ma gratitude à l'ensemble des personnes qui, de près ou de loin, d'une façon ou d'une autre, ont participé à la réalisation de cette thèse. Que ce soit par un soutien scientifique, une relecture attentive, un mot d'encouragement dans un moment de doute, ou simplement par leur présence amicale, chacune a laissé une empreinte dans ce travail et dans mon parcours. Je suis conscient que cette liste de remerciements est nécessairement incomplète. J'ai sans doute omis quelques noms, et je regrette sincèrement ces oublis involontaires. Je demande pardon à celles et ceux qui ne se reconnaîtraient pas dans ces lignes. Qu'ils et elles soient certain(e)s que ma reconnaissance n'en est pas moindre, et que chacun reste profondément gravé dans mon cœur.

Résumé

Cette thèse s'inscrit dans le domaine de la modélisation statistique des valeurs extrêmes, un champ essentiel pour l'analyse et la prédiction d'événements rares aux conséquences majeures dans des secteurs tels que la finance, l'ingénierie ou la gestion des risques naturels. Les méthodes classiques d'estimation des quantiles extrêmes présentent plusieurs limites : instabilité dans la queue de distribution, faible flexibilité en présence de covariables multidimensionnelles et difficulté des approches non paramétriques à capturer adéquatement les comportements asymptotiques. Pour répondre à ces défis, cette thèse développe de nouvelles méthodes de régression quantile extrême alliant de manière cohérente la théorie des valeurs extrêmes et des techniques modernes d'apprentissage statistique. La première contribution propose un estimateur du quantile conditionnel extrême fondé sur une version pondérée du maximum de vraisemblance pour la distribution GEV conditionnelle. Les poids issus des forêts aléatoires généralisées permettent de mieux capturer les relations non linéaires et les interactions complexes entre les covariables, tout en atténuant les effets liés à la forte dimensionnalité. L'existence et la convergence de l'estimateur sont établies, mettant son intérêt pour des quantiles élevés et des covariables de grande dimension. La seconde contribution introduit une pénalisation de type $L2$ sur l'indice de forme ξ, améliorant la stabilité de l'estimation des quantiles extrêmes. Des simulations et une application à des données météorologiques illustrent ses performances. Enfin, une fonction de pénalité spécifiquement adaptée à la distribution GEV est introduite pour stabiliser davantage l'estimation de l'indice des valeurs extrêmes. Cette approche améliore les performances en petits échantillons tout en restant efficace sur de grands jeux de données. Les comparaisons avec des méthodes classiques d'apprentissage statistique montrent des gains substantiels en précision, stabilité et robustesse, confirmés par une application à des données salariales américaines de 1980. Ces contributions apportent des avancées méthodologiques significatives pour la régression quantile extrême en présence de covariables complexes et ouvrent des perspectives prometteuses pour l'analyse d'événements rares dans des cadres multivariés, spatiaux ou temporels.

Mots Clés : Distribution des valeurs extrêmes généralisée, régression quantile extrêmes, forêt aléatoire généralisée, Estimateur de maximum de vraisemblance, Méthode des maxima de blocs.

Abstract

This thesis lies within the field of statistical modeling of extreme values, an essential area for the analysis and prediction of rare events with major consequences in sectors such as finance, engineering, or natural risk management. Classical methods for estimating extreme quantiles suffer from several limitations: instability in the tail of the distribution, limited flexibility in the presence of high-dimensional covariates, and the inability of nonparametric approaches to adequately capture asymptotic behaviors. To address these challenges, this thesis develops new extreme quantile regression methods that coherently combine extreme value theory with modern statistical learning techniques. The first contribution proposes an estimator of the conditional extreme quantile based on a weighted maximum likelihood estimator for the conditional GEV distribution. The weights derived from generalized random forests allow for better capture of nonlinear relationships and complex interactions between covariates, while mitigating issues related to high dimensionality. The existence and consistency of the estimator are established, highlighting its relevance for high quantiles and high-dimensional covariates. The second contribution introduces an $L2$-type penalization on the shape index ξ, improving the stability of extreme quantile estimation. Simulations and an application to meteorological data illustrate its performance. Finally, a penalty function specifically tailored to the GEV distribution is introduced to further stabilize the estimation of the extreme value index. This approach enhances performance in small samples while remaining effective for large datasets. Comparisons with classical statistical learning methods show substantial gains in accuracy, stability, and robustness, confirmed by an application to U.S. wage data from 1980. These contributions provide significant methodological advances for extreme quantile regression in the presence of complex covariates and open promising perspectives for the analysis of rare events in multivariate, spatial, or temporal frameworks.

Keywords: Generalized extreme value distribution, extreme quantile regression, generalized random forest, maximum likelihood estimator, block maxima method.

Contents

List of Tables

List of Figures

General Introduction

Extreme events (financial crises, floods, earthquakes, nuclear accidents, stock market crashes, etc.) occur across a wide variety of physical systems, dominate the news on a recurring basis, and capture our imagination because of their unpredictable nature and the severity of the damage they cause. For example, on the night of February 6, 2023, two major earthquakes (with magnitudes above 7) struck southwestern Türkiye and northern Syria only a few hours apart, resulting in more than 50,000 deaths. Understanding or predicting the occurrence of extreme events in complex nonlinear systems, whether natural or artificial, has therefore become essential, and still stands today as one of the major challenges in fields such as climate science, hydrology, finance, or engineering, to name only a few. Extreme Value Theory provides statistical tools designed to model such phenomena, particularly for estimating extreme quantiles or assessing the probability of occurrence of an extreme event that has never been observed before. Born in the interwar period, this branch of statistics—whose aim is no longer to study sample averages but rather their extreme values, that is, maxima or minima—was developed by a group of statisticians, including (Fisher and Tippett, 1928) and (Gnedenko, 1943). Today, the applications of this theory continue to expand across numerous fields.

The Netherlands has always lived under the constant threat of rising waters. On the night of January 31 to February 1, 1953, a violent storm caused an exceptional rise in sea level, which flooded the southwestern part of the country, resulting in the deaths of more than 1,800 people, the loss of thousands of animals, and the destruction of approximately 50,000 homes. Shocked by the disaster, the Dutch government launched a large-scale hydraulic engineering project, known as the Delta Plan. The commission tasked with reviewing safety standards recommended designing dikes in such a way that the probability of a flood overtopping them would be extremely low, on the order of an event occurring once every 10,000 years in the most exposed areas. To determine these safety levels, a team of scientists relied on Extreme Value Theory (EVT) to analyze historical tidal and flood data. Their goal was to estimate the probability distribution of the annual maximum water level. The results showed that these maxima could be described by a Gumbel distribution, and that the critical height required exceeded five meters. Based on these estimates, massive dams were constructed, which still protect the country today (see Figure 1). If we denote by X the random variable representing

Figure 1: Image of the Eastern Scheldt Barrier in the Netherlands during stormy weather, the main structure of the Delta Plan. (source: Wikipedia)

the annual maximum water level, the aim is to estimate a quantity h such that

$$P(X \leq h) = 1 - \frac{1}{10\,000}.$$

This quantity h is called the quantile of probability level $\tau = 1 - \frac{1}{10\,000}$. In hydrology, the term "return level" is often used to refer to this quantity. Estimating h, however, is not trivial, since the distribution of X is generally unknown and must be inferred from the available data.

Other equally interesting problems can also be considered. For instance, after observing seismic records in a given region for a century, with magnitudes ranging from 0 to 5, can one estimate the probability that an earthquake of magnitude greater than 7 will occur during the next century? Similarly, in public health, a particularly severe influenza epidemic can lead to a large number of hospitalizations or require a substantial amount of antivirals. How can we then assess the probability that an epidemic more intense than any previously observed will occur next year? Answering these types of questions amounts to estimating quantiles of the distribution, which may never have been observed in the available data. When the probability level τ is sufficiently large, that is, $\tau > 1 - \frac{1}{n}$ where n represents the sample size used for the analysis, the corresponding quantile is referred to as an extreme quantile, in other words, a quantile associated with a probability level τ close to 1. These quantiles lie in the tail of the distribution, where data are scarce, making their estimation particularly challenging. This

thesis is positioned within this context.

Often, the phenomena we seek to model are not directly observable. Instead, we have indirect or conditional information through one or more explanatory variables. Formally, let $Y \in \mathcal{Y} \subset \mathbb{R}$ denote the random variable representing the phenomenon of interest, and $X \in \mathcal{X} \subset \mathbb{R}^p$ a vector of covariates. Answering the questions mentioned earlier amounts to estimating the conditional quantiles of Y given $X = x$. In other words, it consists in determining a function $Q_{Y|X=x}(\tau)$ such that

$$P\left(Y \leq Q_{Y|X=x}(\tau) \mid X = x\right) = \tau, \quad \text{for all } x \in \mathcal{X}$$

where $\tau \in (0,1)$ denotes the probability level.

In many situations, summarizing the behavior of the variable of interest by its conditional mean alone is insufficient. The mean can mask asymmetries, variability effects, or extreme behaviors. To better understand the underlying structure of the data, especially rare or atypical events, it is necessary to study the entire conditional distribution. Introduced by (Koenker and Bassett, 1978), quantile regression provides a powerful statistical framework to estimate different quantiles of a conditional distribution as a function of explanatory variables. It is based on the estimation of the conditional quantile and thus allows a more complete description of the relationship between Y and X, capturing heteroskedastic effects, nonlinear dependencies, and behaviors in the tails of the distribution.

In the current context, characterized by the explosion of data and the increase in computational power, classical statistical methods sometimes show their limitations when it comes to modeling complex and nonlinear relationships. Statistical learning methods thus offer a powerful and flexible alternative. They allow capturing subtle interactions and nonparametric structures within high-dimensional data. By combining quantile regression with these modern approaches, such as random forests, boosting, or neural networks, it becomes possible to estimate conditional quantiles in a more robust and adaptive way, even in contexts where classical parametric assumptions are not satisfied.

The objective of this thesis is to contribute to the advancement of extreme quantile regression methods by combining extreme value theory with statistical learning approaches. This combination aims to improve the estimation of high-probability-level quantiles while leveraging the flexibility and predictive power of modern models.

Thesis Outline

- **Chapters 1 and 2** are dedicated to the state of the art and the presentation of the fundamental concepts necessary to understand the basic notions related to extreme value theory and statistical learning. In the first chapter, we introduce univariate extreme value

theory, presenting the main approaches used for the analysis of extreme phenomena. We then discuss the notion of domains of attraction and their associated quantiles. Finally, the last part of the chapter is devoted to a review of extreme quantile estimation methods, first in the unconditional framework and then in the conditional framework. The second chapter is devoted to statistical learning. We present the fundamental principles of this field, defining what a learning algorithm is and describing some commonly used methods, notably those employed in our work. Finally, we introduce extreme quantile regression based on extreme value theory and statistical learning methods, explaining the motivations and objectives of our study.

♣ **Chapter 3** presents our main theoretical contributions. We introduce a weighted maximum likelihood estimator for the conditional generalized extreme value distribution within the framework of extreme quantile regression. We begin by showing how the proposed approach overcomes key limitations of classical quantile regression, particularly in the estimation of very high quantiles. We then establish the existence and consistency of the weighted maximum likelihood estimator, where the weights are derived from the generalized random forest methodology.

♣ **Chapter 4** presents our first applicative-oriented contribution. In this chapter, we develop an extreme quantile regression (QR) method that combines extreme value theory and statistical learning in order to overcome the limitations of classical quantile regression. Following the block maxima framework from extreme value theory, we model the conditional distribution of block maxima using the generalized extreme value (GEV) distribution, whose parameters are allowed to depend on the covariates. These parameters are then estimated through a weighted maximum likelihood estimator, where the weights are provided by generalized random forests. To control the variability of the shape parameter, we introduce an L_2-type penalty acting on the fluctuations of the function $x \mapsto \xi(x)$ across the predictor space $\mathcal{X}$. Simulation studies demonstrate that the proposed method effectively addresses the challenges inherent to classical quantile regression and outperforms several statistical-learning-based quantile regression approaches. Finally, we apply our methodology to the daily weather data from the Fort Collins weather station in Colorado, USA, thus illustrating its practical relevance in real-world contexts.

♣ **Chapter 5** presents our second applicative-oriented contribution. In this work, we integrate a penalty function on the extreme value index within the weighted maximum likelihood estimator. This penalty, originally introduced by (Coles and Dixon, 1999) specially for unconditional GEV distributions, addresses the instabilities of the maximum likelihood estimator in small-sample settings, while preserving its asymptotic efficiency and performance for large samples. The effectiveness and robustness of the proposed

methodology are demonstrated through comparative simulation studies and validated on a real dataset, namely the 1980 U.S. wage data.

We concludes this manuscript and outlining several perspectives that naturally emerge from the work conducted in this thesis. Finally, we note that the last three chapters, which present our contributions, can be read independently. Each chapter begins with a brief review of the essential notions, providing the reader with the necessary background to understand the results without requiring a linear reading of the entire manuscript.

List of Publications

- Vidagbandji, L. M., Berred, A., Bertelle, C., & Amanton, L. (2025). Generalized random forest for extreme quantile regression. *Communications in Statistics - Simulation and Computation*, 1–24. https://doi.org/10.1080/03610918.2025.2543854.
- Vidagbandji, L. M., Berred, A., Bertelle, C., & Amanton, L. (2026). Penalized estimation of GEV parameters for extreme quantile regression. *Accepted for publication in Journal of Statistical Theory and Practice*
- Vidagbandji *et al.* (2026). Consistency of Weighted maximum likelihood estimator for extreme quantile regression. To be submitted.

List of communications

1. Vidagbandji *et al.* Local Weighted Maximum Likelihood Estimator for Extreme Quantile Regression. *Atelier des doctorants des laboratoires LMI et LMRS*, Université de Rouen, 10 Mars, 2026.
2. Vidagbandji *et al.* Extreme quantile regression using generalized random forests and block maxima approach. *International Conference on Extreme Value Analysis, Probabilistic and Statistical Models and their Application* (EVA 2025). University of North Carolina at Chapel Hill, USA, 22-27 June, 2025.
3. Vidagbandji *et al.* Penalized estimation of GEV parameters for extreme quantile regression. *56e Journées des statistiques de la SFDS*. Université Aix Marseille, Campus Saint-Charles, France, 02-06 juin, 2025.
4. Vidagbandji *et al.* Combining Extreme Value Theory and Random Forests for High Quantile Regression. *17e Journée de la Féderation Normandie Mathématiques*. INSA Rouen Normandie, France, 26 mai, 2025.

5. Vidagbandji *et al.* Generalized random forest approach for GEV extreme quantile regression. *Rencontres des jeunes chercheurs africains en france, sixième édition*. Institut Henri Poincaré (Paris), France, 12-13 december, 2024.

6. Vidagbandji *et al.* Generalized random forest for extreme quantile regression. *The 26th international conference on computational statistics*. University of Giessen, Allemagne, 27-30 august, 2024, 31.

7. Vidagbandji *et al.* Parameter estimation of generalized extreme value distribution using generalized random forest method. *16e Journée de la Fédération Normandie Mathématiques*. Université de Rouen Normandie, France, 05 juillet, 2024.

8. Vidagbandji *et al.* GEV-Extremal random forest. *55e Journées de statistiques de la SFDS*. Campus de la victoire de l'Université de Bordeaux, France, 27-31 mai, 2024, 1098–1105.

9. Vidagbandji *et al.* Quantile regression, An approach based on GEV distribution and machine learning. *The french regional conference on complex systems*. Université Le Havre Normandie, France, 31 mai-2 juin, 2023.

CHAPTER UN

GENERALITIES ON EXTREME VALUE THEORY

Contents

This chapter is based on the books (Coles, 2001), (Reiss et al., 1997), (De Haan and Ferreira, 2006), (Beirlant et al., 2006) and the articles (Fisher and Tippett, 1928), (Gumbel, 1941), (Fisher and Tippett, 1928), and (Balkema and De Haan, 1974).

1.1 Introduction

Most standard statistical methods aim to characterize the central behavior of a distribution, where the majority of observations lie and where notions such as the mean or variance play a decisive role. Yet, many phenomena of interest, whether natural, financial, or industrial risks, are governed not by typical behavior but by rare events occurring in the tails of the distribution. Extreme Value Theory (EVT) is precisely concerned with studying these extreme regions and providing a rigorous framework for their modeling. The emergence of this theory dates back to 1928, when Leonard Tippett, a researcher at the British Cotton Industry Research Association, showed that the strength of a cotton yarn was determined by its strongest fiber. This insight highlighted the limitations of mean-centered approaches in describing phenomena dominated by extreme observations. A few years later, the foundational work of (Gumbel, 1941) marked a major milestone, especially in hydrology for flood modeling, thereby initiating the first concrete applications of EVT. Since then, Extreme Value Theory has become an essential tool for analyzing extreme risks across numerous fields: meteorology and climatology for the study of extreme precipitation or temperature (Coles and Tawn, 1996), (Abdelaziz, 2013), (Afroz et al., 2021); insurance for evaluating large claims (Bousebata, 2022), (Cohen Sabban, 2022), (Abad et al., 2014); finance for the assessment of extreme risks (Gilli and këllezi, 2006), (Bensalah, 2000); epidemiology (Butler et al., 1998); and anomaly detection (Bochenek and Ustrnul, 2022), (Al-Behadili et al., 2016). This chapter introduces the fundamental concepts of EVT, presenting key asymptotic results and the main models used to describe tail behavior.

1.2 Univariate Extreme Values

Consider a sequence of independent and identically distributed ($i.i.d.$) random variables $X_1, X_2, \ldots,$ with common distribution function F. Let denote

$$M_n = \max\{X_1, \ldots, X_n\}.$$

The distribution function of M_n is given by

$$F_{M_n}(x) = \Pi_{i=1}^{n} [P(X_i \leq x)] = F^n(x). \tag{1.1}$$

This relation shows that the distribution of the maximum depends directly on the underlying distribution F. However, in practice, F is not always known. Even when it is known, the distribution of M_n may be difficult to compute explicitly. A natural approach consists in estimating F from observed data using classical statistical methods, then substituting this estimate into Equation (1.1) to obtain an approximation of the distribution of the maximum. However, this

method has a major limitation: a small error in the estimation of F can lead to significant discrepancies between the true distribution of the maximum and its approximation. Extreme Value Theory (EVT) aims precisely to overcome this difficulty by studying the possible limiting distributions of M_n (after normalization). The pinoreering work of (Fisher and Tippett, 1928) on the limiting distributions of the normalized maximum of an $i.i.d.$ sequence marks the starting point of this theory. These results were later refined by (Gnedenko, 1943), who formalized the conditions for convergence. Subsequently, a unified formulation of the limiting laws was proposed thanks to the contributions of (Von Mises, 1936) and (Jenkinson, 1955). The fundamental theorem of this theory, known as the *Extreme Value Theorem* (see Theorem 1.1), builds on these major contributions. Before stating the theorem, we recall the notion of a *degenerate distribution*.

Definition 1.1. ***Degenerate Distribution***
Let $(\Omega, \Im, \mathbb{P})$ be a probability space and $(\chi, \mathbb{B})$ an observation space. A random variable $X : \Omega \to \chi$ is said to be degenerate*, or to have a degenerate distribution F_X, if X is almost surely constant.*

Theorem 1.1. *(Fisher and Tippett, 1928; Gnedenko, 1943)*
Let $X_1, X_2, \ldots$ be an $i.i.d.$ sequence with distribution function F. If there exist normalizing sequences $a_n > 0$ and $b_n \in \mathbb{R}$, and a non-degenerate distribution function G such that

$$\lim_{n \to +\infty} F^n(a_n y + b_n) = G(y), \tag{1.2}$$

then G must be one of the following three families of extreme value distributions

1. ***Gumbel (Type I):*** $\Lambda(x) = \exp(-\exp(-x))$, *for* $x \in \mathbb{R}$;

2. ***Fréchet (Type II):*** $\Phi_{\frac{1}{\xi}}(x) = \begin{cases} 0, & x \leq 0, \\ \exp(-x^{-\frac{1}{\xi}}), & x > 0, \end{cases}$ *with* $\xi > 0$;

3. ***Weibull (Type III):*** $\Psi_{-\frac{1}{\xi}}(x) = \begin{cases} \exp(-(-x)^{-\frac{1}{\xi}}), & x < 0, \\ 1, & x \geq 0, \end{cases}$ *with* $\xi < 0$.

This theorem shows that, regardless of the distribution of the random variables X_i, the possible limiting distribution of the normalized maximum M_n necessarily belongs to one of the three extreme value distribution families.

1.3 Generalized Extreme Value Distribution

The three classical distributions presented previously can be combined into a single parameterization, introduced by (Jenkinson, 1955), of the form

$$G_\xi(x) = \begin{cases} \exp\left(-(1+\xi x)^{-\frac{1}{\xi}}\right), & \xi \neq 0,\ 1+\xi x > 0, \\ \exp(-\exp(-x)), & \xi = 0, \end{cases} \qquad \forall x \in \mathbb{R}. \tag{1.3}$$

The parameter ξ is called the *extreme value index* and controls the heaviness of the tail of the distribution. This distribution is known as the *Generalized Extreme Value* distribution, abbreviated GEV. The set of distribution functions F satisfying equation (1.2) forms the *domain of attraction* of G_ξ, denoted $\mathcal{D}(G_\xi)$. The most general form of the GEV distribution is given by

$$G_{\mu,\sigma,\xi}(x) = \begin{cases} \exp\left(-\left(1+\xi\dfrac{x-\mu}{\sigma}\right)_+^{-\frac{1}{\xi}}\right), & \xi \neq 0, \\ \exp\left(-\exp\left(-\dfrac{x-\mu}{\sigma}\right)\right), & \xi = 0, \end{cases} \qquad \forall x \in \mathbb{R}, \tag{1.4}$$

where $\mu \in \mathbb{R}$, $\sigma > 0$, and $\xi \in \mathbb{R}$ are respectively the location, scale, and shape parameters. We denote $a_+ = \max\{0, a\}$. This distribution combines the three classes of extreme value laws. Mainly

- when $\xi > 0$, one recovers the Fréchet distribution;
- when $\xi < 0$, one obtains the Weibull distribution;
- when $\xi = 0$, the Gumbel law is recovered as the limiting case of (1.4).

Readers interested in the proof of this result, or wishing to explore the concepts introduced in this chapter in greater depth, may refer to the monograph by (De Haan and Ferreira, 2006). Let us recall that the previous results may also be adapted to the study of minima using the following transformation

$$\min(X_1, \ldots, X_n) = -\max(-X_1, \ldots, -X_n).$$

In the remainder of this manuscript, we restrict attention to the case of maxima; the corresponding results for minima follow from an analogous reasoning.

Before proceeding further, we introduce some notations that will be used throughout the chapter. If X denotes a real-valued random variable with distribution function F, the right endpoint of F, that is, the upper bound of its support, is defined by

$$x^* = \sup\{x \in \mathbb{R} : F(x) < 1\},$$

with the convention $\sup\{\varnothing\} = \infty$. We also define

- the generalized inverse of F, given by

$$F^{\leftarrow}(x) = \inf\{y \in \mathbb{R} : F(y) \geq x\};$$

- the function $U(\cdot)$, defined as the generalized inverse of $\dfrac{1}{1-F(\cdot)}$, that is

$$U(t) = \left(\frac{1}{1-F}\right)^{\leftarrow}(t);$$

- the asymptotic equivalence relation, denoted $f(x) \sim g(x)$ as $x \to a$, if

$$\lim_{x\to a} \frac{f(x)}{g(x)} = 1.$$

We may now state the necessary and sufficient condition for a distribution function F to belong to the domain of attraction of the generalized extreme value distribution, denoted $\mathcal{D}(G_\xi)$.

Theorem 1.2. *Let U be the generalized inverse of the function $\dfrac{1}{1-F}$. Then $F \in \mathcal{D}(G_\xi)$ if and only if there exists a positive function $a(\cdot)$ such that*

$$\lim_{t\to+\infty} \frac{U(tx)-U(t)}{a(t)} = \frac{x^\xi - 1}{\xi}, \quad \textit{for all } x > 0. \tag{1.5}$$

This relation, known as the extreme value condition, forms the basis of many estimators of the extreme value index ξ, a key parameter in the analysis of rare events. Among the most classical estimators are: Hill's estimator (Hill, 1975), Pickands' estimator (Pickands III, 1975), and the moment estimator.

Definition 1.2. *Let $X_1, X_2, \ldots$ be an i.i.d. sequence of random variables with distribution function F. We say that F is max-stable if, for every $n \geq 2$, there exist constants $a_n > 0$ and $b_n \in \mathbb{R}$ such that*

$$\lim_{n\to+\infty} F^n(a_n x + b_n) = G_\xi(x), \quad \forall x \in \mathbb{R}. \tag{1.6}$$

A possible choice of the function $a(t)$ in Equation (1.5) is

- $a(t) = \xi U(t)$, if $\xi > 0$;
- $a(t) = -\xi\big(U(\infty) - U(t)\big)$, if $\xi < 0$;
- $a(t) = U(t) - \dfrac{1}{t}\displaystyle\int_0^t U(s)\,ds$, if $\xi = 0$.

The normalization constants introduced above are then

$$a_n = a(n), \quad b_n = U(n).$$

A distribution is *max-stable* if and only if it follows a generalized extreme value distribution (Coles, 2001). There are two main approaches commonly used in the analysis of extreme values: the block maxima (BM) method and the Peak-Over-Threshold (POT) method. In the following, we present these two approaches and derive the corresponding quantile expressions.

1.4 Extreme Value Analysis

1.4.1 Block Maxima Method and Associated Quantile

Let $X_1, \cdots, X_N$ be a sequence of $i.i.d.$ random variables representing the available data for the analysis. The block maxima method, commonly referred to as the *Block Maxima* approach (denoted **BM**), consists in dividing the initial sample into several blocks of equal size and extracting the maximum of each block for further study. Assume that the data sequence is divided into n blocks of size m, denoted $B_{j,m} = \{X_{(j-1)m+1}, \ldots, X_{jm}\}, \quad j = 1, \ldots, n$. The maximum of the j^{th} block is then defined by

$$Z_{j,m} = \max\{B_{j,m}\}.$$

Under the assumption that $m \to +\infty$, Theorem 1.1 implies that these maxima follow asymptotically a Generalized Extreme Value distribution given by equation (5.2). The BM approach also assumes that the sequence of block maxima, $Z_{1,m}, \ldots, Z_{n,m}$, consists of i.i.d. random variables following a GEV distribution (see (Fisher and Tippett, 1928) and (Gnedenko, 1943) for further details). Based on these data, several inference methods can be used to estimate the parameters of the corresponding GEV distribution, including the method of moments and maximum likelihood estimation.

To illustrate the BM approach, suppose that the available data are represented in Figure 1.1. The BM method then consists in retaining, for inference, only the red observations shown in Figure 1.2.

Quantile Obtained via the BM Approach

Once the GEV distribution is fitted, one of the main objectives is the estimation of the quantile associated with a probability level $\tau \in (0,1)$ with $\tau \to 1$, defined by

$$Q(\tau) = \inf\{y : G_{\mu,\sigma,\xi}(y) \geq \tau\}.$$

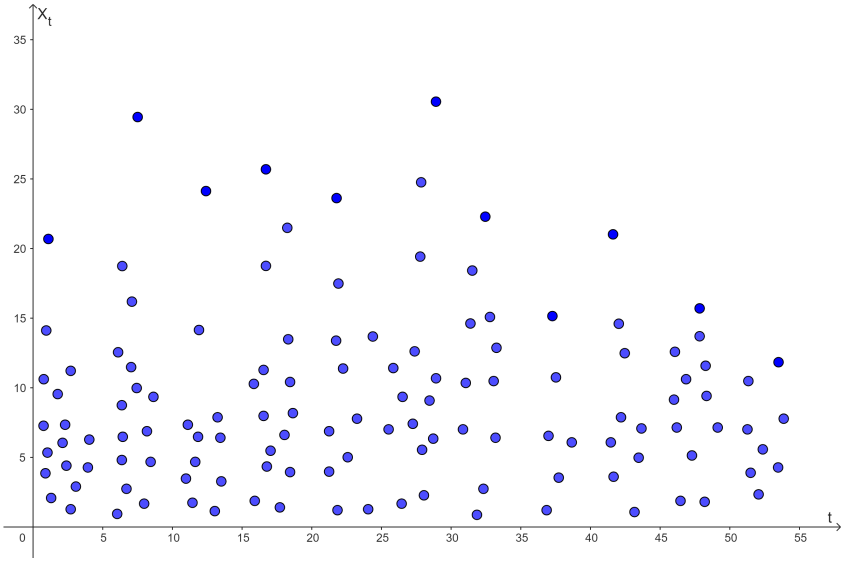

Figure 1.1: Data example

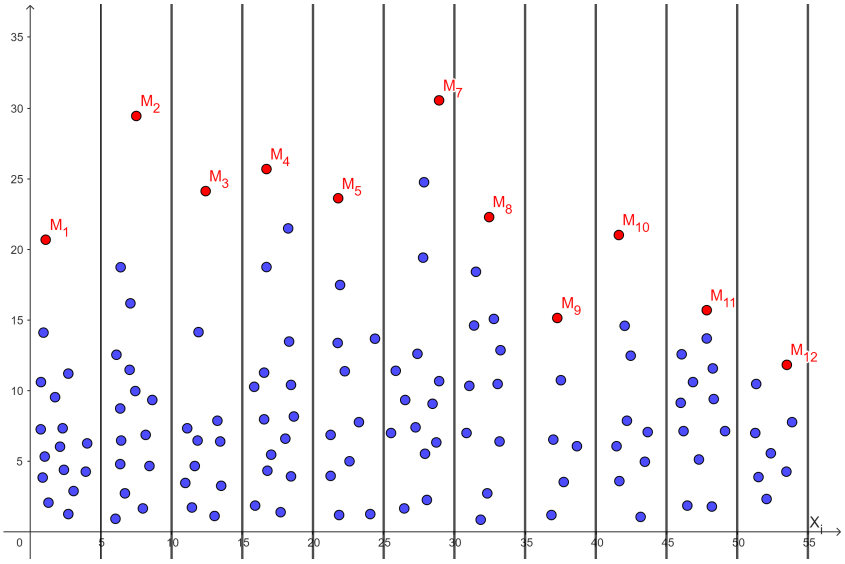

Figure 1.2: Block maxima

The quantile of order τ under the BM approach corresponds to the generalized inverse of the GEV distribution and is given by

$$Q(\tau) = \begin{cases} \mu + \dfrac{\sigma}{\xi}\left[(-\ln(\tau))^{-\xi} - 1\right], & \text{if } \xi \neq 0, \\ \mu - \sigma \ln[-\ln(\tau)], & \text{if } \xi = 0. \end{cases} \tag{1.7}$$

In hydrology, this quantile is referred to as the *return level*, and it is associated with a *return period* equal to $\frac{1}{1-\tau}$. In other words, the return level represents the value expected to be reached or exceeded, on average, once every $\frac{1}{1-\tau}$ cycles (e.g., years) (Coles, 2001).

Although widely used in practice, this approach may lead to estimation inaccuracies. In-

deed, some blocks may contain several extreme events, whereas others contain none. An inappropriate choice of block size may either increase the variance (few blocks) or introduce bias (many blocks). Thus, selecting the block size requires a bias–variance trade-off, which is a central issue in the practical implementation of the BM method.

1.4.2 Peak-Over-Threshold Method and Associated Quantile

This approach, initially introduced by (Balkema and De Haan, 1974), was later formalized more rigorously by (Leadbetter, 1991). To present the method, consider again a sequence of i.i.d. random variables $X_1, \cdots, X_N$ with cumulative distribution function F. Let u be a sufficiently high threshold such that $u < x^*$. Assume that F satisfies the conditions of Theorem 1.1.

Define the excess random variable $Y_u = X - u \mid X > u$, representing the exceedance above the threshold u. The distribution function H_u of Y_u is given by the Generalized Pareto Distribution (GPD) with parameters ξ and $\bar{\sigma}$

$$F_{Y_u}(t) = H_u(t) = 1 - \left(1 + \frac{\xi t}{\bar{\sigma}}\right)^{-\frac{1}{\xi}}, \quad \text{where } \bar{\sigma} = \sigma + \xi(u - \mu). \tag{1.8}$$

Indeed, one has

$$\begin{aligned} H_u(t) &= \mathbb{P}(X - u \leq t \mid X > u) \\ &= \frac{\mathbb{P}(u < X \leq u + t)}{\mathbb{P}(X > u)} \\ &= 1 - \frac{1 - F(u+t)}{1 - F(u)}. \end{aligned} \tag{1.9}$$

Since F satisfies the assumptions of Theorem 1.1, we have

$$F^n(t) = \exp\left[-\left(1 + \xi\frac{t-\mu}{\sigma}\right)^{-\frac{1}{\xi}}\right].$$

Thus,

$$n \log F(t) = -\left(1 + \xi\frac{t-\mu}{\sigma}\right)^{-\frac{1}{\xi}}. \tag{1.10}$$

For large values of t, one may approximate $\log F(t) \approx F(t) - 1$, which yields

$$1 - F(u) \approx \frac{1}{n}\left(1 + \xi\frac{u-\mu}{\sigma}\right)^{-\frac{1}{\xi}},$$

for sufficiently large u. Furthermore, for any $t > 0$,

$$1 - F(u+t) \approx \frac{1}{n}\left(1+\xi\frac{u+t-\mu}{\sigma}\right)^{-\frac{1}{\xi}}.$$

Substituting these approximations into (1.9), we obtain

$$\begin{aligned} H_u(t) &= 1 - \frac{\left(1+\xi\frac{u+t-\mu}{\sigma}\right)^{-\frac{1}{\xi}}}{\left(1+\xi\frac{u-\mu}{\sigma}\right)^{-\frac{1}{\xi}}} \\ &= 1 - \left(1+\frac{\xi t}{\sigma+\xi(u-\mu)}\right)^{-\frac{1}{\xi}}, \end{aligned}$$

which corresponds to the GPD distribution in (1.8).

As an illustration, the *Peaks Over Threshold* approach consists in considering only observations exceeding a fixed threshold u, and approximating the distribution of these exceedances by the GPD. In other words, this method focuses exclusively on the extreme observations (shown in red in Figure 1.3) which correspond to the highest values in Figure 1.1, i.e., those exceeding the chosen threshold u.

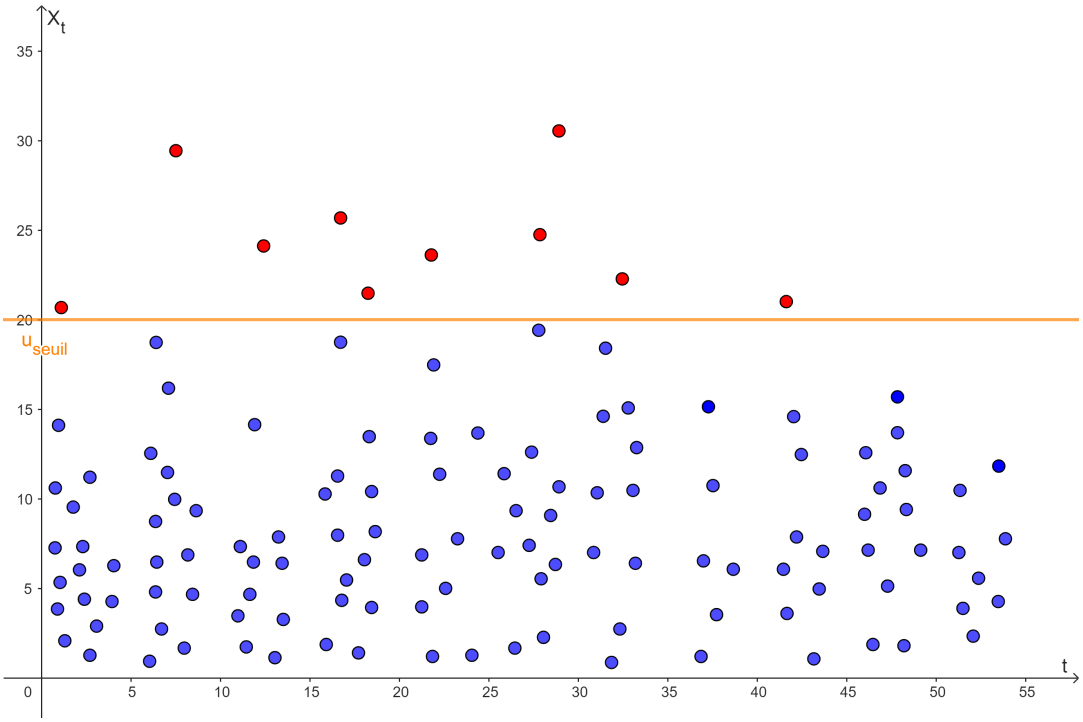

Figure 1.3: POT approach

Quantile obtained via the POT Approach

Once the Generalized Pareto Distribution (GPD) is fitted to exceedances above a threshold u, it is possible to derive the quantile associated with a probability level $\tau \in (0,1)$ with $\tau \to 1$. From (1.9), for any $x > u$ we have

$$1 - F(x) = (1 - F(u))\,(1 - H_u(x-u))\,,$$

where H_u is given by (1.8). By substituting the expression of H_u, we obtain

$$1 - F(x) = (1 - F(u)) \left(1 + \frac{\xi(x-u)}{\bar{\sigma}}\right)^{-\frac{1}{\xi}}, \quad \text{for } 1 + \frac{\xi(x-u)}{\bar{\sigma}} > 0.$$

Assuming that the threshold u corresponds to a quantile denoted Q_{τ_0} at probability level τ_0, i.e., $u = Q(\tau_0)$ and $F(u) = \tau_0$, the quantile of order $\tau \geq \tau_0$ is given by

$$Q(\tau) = \begin{cases} Q(\tau_0) + \dfrac{\bar{\sigma}}{\xi}\left[\left(\dfrac{1-\tau}{1-\tau_0}\right)^{-\xi} - 1\right], & \text{if } \xi \neq 0, \\ Q(\tau_0) + \bar{\sigma} \ln\left(\dfrac{1-\tau_0}{1-\tau}\right), & \text{if } \xi = 0. \end{cases} \tag{1.11}$$

The quantity τ_0 is often referred to as the *intermediate probability level*.

The POT approach, although widely used in practice for modeling extreme events, also presents difficulties related to the choice of the threshold, much like the choice of block size in the BM approach. Indeed, selecting the threshold is delicate: a low threshold introduces bias, whereas a high threshold increases the variance of the estimates. Thus, the bias–variance trade-off remains a central issue in the practical implementation of the POT method. The choice between these two approaches depends primarily on the nature and availability of the data used for extreme value analysis, each offering both advantages and limitations.

1.5 Domains of Attraction

Depending on the sign of the extreme–value index ξ, three domains of attraction for extreme–value distributions are distinguished.

Case $\xi = 0$

When $\xi = 0$, we have $x^* = +\infty$. The survival function $1 - G_0(x)$ decreases exponentially as $x \to +\infty$, i.e., $1 - G_0(x) \sim e^{-x}$ (light–tailed distributions). In this case, the distribution F belongs to the **Gumbel domain of attraction**, denoted $\mathcal{D}(\Lambda)$. This includes, for example, the normal, exponential, gamma, and the Gumbel distributions.

Case $\xi > 0$

This case corresponds to the Fréchet distribution with parameter $\alpha = 1/\xi$. The survival function decreases as a power function, that is, $1 - G_\xi(x) \sim \xi^{-1/\xi} x^{-1/\xi}$ as $x \to \infty$ (heavy–tailed, Pareto–type distributions). The distribution F then belongs to the **Fréchet domain of attraction**, denoted $\mathcal{D}(\Phi_{\frac{1}{\xi}})$. Here we still have $x^* = \infty$. This domain includes, in particular, the Pareto, Cauchy, Burr, and Student distributions (with small degrees of freedom).

Case $\xi < 0$

This case corresponds to the Weibull distribution with parameter $\alpha = -1/\xi$. The distribution F has a finite upper endpoint $x^* = -\frac{1}{\xi}$, beyond which $F(x) = 1$. We say that F belongs to the **Weibull domain of attraction**, denoted $\mathcal{D}(\Psi_{-\frac{1}{\xi}})$. Typical examples include the uniform, beta, and Reverse Burr distributions.

Figure 1.4 illustrates the cumulative distribution function of the extreme–value distribution for the different domains of attraction.

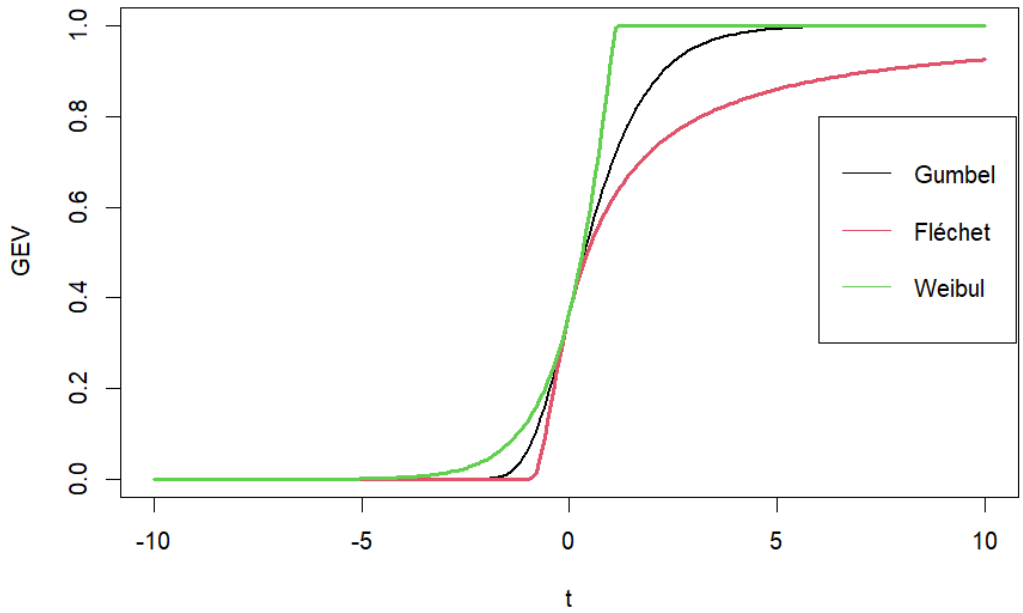

Figure 1.4: Cumulative distribution function of the GEV distribution for $\mu = 0$, $\sigma = 1$ and $\xi \in \{-0.9, 0, 0.9\}$.

We now describe in more detail the characteristics of these domains of attraction using the notions of regular variation. To do so, we introduce a few definitions and fundamental properties.

Definition 1.3. *A measurable function $H : \mathbb{R}_+ \to \mathbb{R}_+$ is said to be regularly varying at infinity with index $a \in \mathbb{R}$ (denoted $H \in \mathcal{RV}_a$) if, for every $t > 0$, we have*

$$\lim_{x\to+\infty} \frac{H(tx)}{H(x)} = t^a.$$

Intuitively, a regularly varying function behaves asymptotically like a power function. When $a = 0$, the function H is said to be slowly varying (denoted $H \in \mathcal{RV}_0$) and is often written as $L(x)$.

Propriété 1. *Every regularly varying function H with index $a \in \mathbb{R}$ at infinity can be written as*

$$H(x) = x^a L(x), \quad \textit{with } L \in \mathcal{RV}_0.$$

Proof. Let H be a regularly varying function at infinity with index a, and set $L(x) = \frac{H(x)}{x^a}$. If $H \in \mathcal{RV}_a$, then $L \in \mathcal{RV}_0$. Indeed,

$$\begin{aligned}
\lim_{x\to\infty} \frac{L(tx)}{L(x)} &= \lim_{x\to+\infty} \frac{\frac{H(tx)}{(tx)^a}}{\frac{H(x)}{x^a}} \\
&= \lim_{x\to+\infty} \frac{H(tx)}{t^a H(x)} \\
&= \frac{1}{t^a} \lim_{x\to+\infty} \frac{H(tx)}{H(x)} \\
&= 1, \quad \text{because } H \in \mathcal{RV}_a.
\end{aligned}$$

Thus $L \in \mathcal{RV}_0$. □

This result reduces the study of regularly varying functions to that of slowly varying functions at infinity.

Theorem 1.3. ***(Karamata Representation)***
Every slowly varying function L can be written as

$$L(t) = c(t) \exp\left[\int_1^t \frac{\delta(u)}{u}\, du\right]$$

where $c(\cdot)$ and $\delta(\cdot)$ are functions from $\mathbb{R}_+$ to $\mathbb{R}_+$ such that

$$\lim_{t\to+\infty} c(t) = c \quad (c > 0), \qquad \lim_{t\to+\infty} \delta(t) = 0.$$

The proof of this theorem can be found in (Resnick, 2007) and (Resnick, 1987). The following proposition will be particularly useful later.

Propriété 2. *1. If $H \in \mathcal{RV}_\xi$ with $\xi \in \bar{\mathbb{R}}$, then*

$$\lim_{x\to+\infty} \frac{\log H(x)}{\log x} = \xi, \quad \textit{and} \quad \lim_{x\to+\infty} H(x) = \begin{cases} 0, & \textit{if } \xi < 0, \\ \infty, & \textit{if } \xi > 0. \end{cases}$$

2. If $H \in \mathcal{RV}_\xi$ with $\xi > 0$ (resp. $\xi < 0$), then $H^{\leftarrow} \in \mathcal{RV}_{1/\xi}$ (resp. $H^{\leftarrow}(1/x) \in \mathcal{RV}_{-1/\xi}$).

3. If $H \in \mathcal{RV}_\xi$ with $\xi \in \mathbb{R}$ and $\sigma > 0$, then there exists t_0 such that for all $x \geq 1$ and $t \geq t_0$, we have

$$(1-\sigma)x^{\xi-\sigma} \leq \frac{H(tx)}{H(t)} \leq (1+\sigma)x^{\xi+\sigma}.$$

4. If $H \in \mathcal{RV}_\xi$ with $\xi \neq 0$, then there exists a function H^* that is absolutely continuous, strictly monotone, and such that $H(x) \sim H^*(x)$ as $x \to +\infty$.

The proof of this proposition can be found in Proposition 2.6 of (Resnick, 2007). Many results in extreme–value theory rely on these theorems and properties, notably in (De Haan and Ferreira, 2006), (Resnick, 1987), and (Beirlant et al., 2006). We will use them here to characterize the different domains of attraction, that is, to determine, from the distribution function F of a random variable X, the necessary and sufficient conditions for it to belong to a given domain of attraction, as well as the associated normalizing constants.

1.6 Characterization of the Domains of Attraction and Associated Quantile

The following theorem, known as the *Von Mises condition*, provides a sufficient condition for a probability distribution to belong to a given domain of attraction.

Theorem 1.4. ***Von Mises Condition*** *(De Haan and Ferreira, 2006)*
Let F be a distribution function defined on $\mathbb{R}$, with right endpoint x^*. Assume that F'' exists and that $F'(x) > 0$ for all x in a neighborhood of x^*. If

$$\lim_{x\to x^*} \left(\frac{1-F}{F'}\right)'(x) = \xi, \tag{1.12}$$

or equivalently,

$$\lim_{x\to x^*} \frac{(1-F(x))F''(x)}{(F'(x))^2} = -\xi - 1,$$

then F belongs to the domain of attraction of G_ξ, denoted $F \in \mathcal{D}(G_\xi)$.

In terms of the previously defined function U, condition (1.12) is equivalent to

$$\lim_{x\to\infty} \frac{xU''(x)}{U'(x)} = \xi - 1.$$

This also implies that

$$\lim_{x\to\infty} \frac{U'(tx)}{U'(x)} = t^{\xi-1}, \qquad \forall t \in (0,+\infty),$$

For more detail, see Corollary 1.1.10 in (De Haan and Ferreira, 2006). A simpler characterization can be obtained when $\xi \neq 0$.

Theorem 1.5. *(De Haan and Ferreira, 2006)*

1. *Suppose that $x^* = +\infty$ and that F' exists. If*

$$\lim_{x\to+\infty} \frac{xF'(x)}{1-F(x)} = \frac{1}{\xi}, \quad \textit{for some } \xi > 0,$$

then F belongs to the domain of attraction of the Fréchet distribution, denoted $F \in \mathcal{D}(\Phi_{\frac{1}{\xi}})$.

2. *Suppose that $x^* < +\infty$ and that F' exists for all $x < x^*$. If*

$$\lim_{x\to x^*} \frac{(x^*-x)F'(x)}{1-F(x)} = -\frac{1}{\xi}, \quad \textit{for some } \xi < 0,$$

then F belongs to the domain of attraction of the Weibull distribution, denoted $F \in \mathcal{D}(\Psi_{-\frac{1}{\xi}})$.

The interested reader may consult (De Haan and Ferreira, 2006) or (Resnick, 1987) for full proofs. We now present the necessary and sufficient conditions allowing one to determine whether a distribution function F belongs to a particular domain of attraction.

1.6.1 Fréchet Domain of Attraction and Associated Quantile

Theorem 1.6. *(De Haan and Ferreira, 2006)*
A distribution function F belongs to the Fréchet domain of attraction, denoted $F \in \mathcal{D}(\Phi_{\frac{1}{\xi}})$, if and only if $x^* = +\infty$ and

$$\lim_{x\to+\infty} \frac{1-F(tx)}{1-F(x)} = t^{-\frac{1}{\xi}}, \quad \textit{for all } t > 0, \quad \textit{with } \xi > 0.$$

Equivalently, $F \in \mathcal{D}(\Phi_{\frac{1}{\xi}})$ if and only if the survival function $\bar{F} = 1-F$ is regularly varying at infinity with index $-\frac{1}{\xi}$. The corresponding normalizing sequences are

$$a_n = (\tfrac{1}{F})^{\leftarrow}(n) \quad \textit{and} \quad b_n = 0.$$

Thus, any distribution F in the Fréchet domain of attraction can be written as

$$F(x) = 1 - x^{-\frac{1}{\xi}} L(x), \quad \text{where } L(\cdot) \in \mathcal{RV}_0. \tag{1.13}$$

This result, due to (Gnedenko, 1943), shows that a distribution function belongs to Fréchet's domain of attraction if and only if its survival function is regularly varying. The complete proof of this theorem is available in (Resnick, 1987) and (De Haan and Ferreira, 2006).

From (1.13), the associated quantile of order τ is obtained by inverting the probability distribution function. This gives

$$Q(\tau) = (1-\tau)^{-\xi} \mathcal{L}\left(\frac{1}{1-\tau}\right), \quad \mathcal{L} \in \mathcal{RV}_0, \quad \tau \in (0,1).$$

This expression has led to many estimators of extreme quantiles and tail index for heavy-tailed distributions, including Hill's estimator. Thanks to Karamata's representation, one also has the following characterization.

Corollaire 1.1. *(Corollary 1.12 in (Resnick, 1987))*
$F \in \mathcal{D}(\Phi_{\frac{1}{\xi}})$ if and only if there exist measurable functions $c(\cdot)$ and $d(\cdot)$ defined on $(1,+\infty)$ such that

$$\lim_{x\to+\infty} c(x) = C > 0, \qquad \lim_{x\to+\infty} d(x) = \frac{1}{\xi} > 0,$$

and

$$\bar{F}(x) = c(x) \exp\left\{ -\int_1^x t^{-1} d(t)\, dt \right\}, \qquad \forall x \geq 1.$$

1.6.2 Weibull Domain of Attraction and Associated Quantile

Theorem 1.7. *(Resnick, 1987)*
A distribution function F belongs to the Weibull domain of attraction, that is $F \in \mathcal{D}(\Psi_{-\frac{1}{\xi}})$, if and only if $x^* < +\infty$ and

$$1 - F\left(x^* - \frac{1}{x}\right) \in \mathcal{RV}_{\frac{1}{\xi}}, \qquad x \to \infty, \quad (\xi < 0).$$

Similarly, the following results show that it is possible to establish a direct correspondence between the domain of attraction of Weibull and that of Fréchet using a simple change of variable in the distribution function (see (Resnick, 1987)).

Theorem 1.8. *A distribution function F belongs to the Weibull domain of attraction with ex-*

treme value index $\xi < 0$ *if and only if* $x^* < +\infty$ *and the function*

$$H(x) = \begin{cases} 0, & \text{if } x < 0, \\ F(x^* - x^{-1}), & \text{if } x \geq 0, \end{cases}$$

belongs to the Fréchet domain of attraction with index $-\xi$.

Thus any F in the Weibull domain can be written as

$$F(x) = 1 - (x^* - x)^{-\frac{1}{\xi}} L\big((x^* - x)^{-1}\big), \quad \text{for all} \quad x < x^*, \quad \text{where} \quad L \in \mathcal{RV}_0. \tag{1.14}$$

The associated quantile of order τ is

$$Q(\tau) = x^* - (1-\tau)^{-\xi} \mathcal{L}\left(\frac{1}{1-\tau}\right), \quad \mathcal{L} \in \mathcal{RV}_0.$$

The normalizing sequences are:

$$a_n = x^* - \bar{F}^{\leftarrow}(1/n), \qquad b_n = x^*.$$

Another characterization based on Karamata's representation is given by the following corollary.

Corollaire 1.2. *(Corollary 1.14 in (Resnick, 1987))*
$F \in \mathcal{D}(\Psi_{-\frac{1}{\xi}})$ *if and only if* $x^* < \infty$ *and there exist functions* $c : \mathbb{R}_+ \to \mathbb{R}_+$, $d : \mathbb{R}_+ \to \mathbb{R}_+$ *and a constant* $c_0 > 0$ *such that:*

$$\lim_{x \to x^*} c(x) = -\frac{1}{\xi} > 0, \qquad \lim_{x \to x^*} d(x) = c_0,$$

and

$$\bar{F}(x) = d(x) \exp\left\{ - \int_{x^*-1}^{x} \frac{c(t)}{x^* - t}\, dt \right\}, \qquad x < x^*.$$

1.6.3 Gumbel Domain of Attraction and Associated Quantile

The Gumbel domain of attraction is generally harder to characterize than the Fréchet or Weibull domains. We present here a classical characterization from Proposition 1.4 of (Resnick, 1987).

A distribution F^* with endpoint x^* is called a *Von Mises function* if there exist $z_0 < x^*$ and $c > 0$ such that for all $x \in (z_0, x^*)$:

$$1 - F^*(x) = c \exp\left\{ - \int_{z_0}^{x} \frac{1}{k(u)}\, du \right\}, \tag{1.15}$$

where k is positive, absolutely continuous on (z_0, x^*), with derivative $k'(u)$ satisfying

$$\lim_{u \to x^*} k'(u) = 0.$$

The function k is called the **auxiliary function**. The following characterization establishes the link between Von Mises functions and the Gumbel attraction domain.

Theorem 1.9. *A distribution function F belongs to the Gumbel domain of attraction, $F \in \mathcal{D}(\Lambda)$, if and only if there exists a Von Mises function F^* such that for all $x \in (z_0, x^*)$,*

$$\bar{F}(x) = c(x)\,(1 - F^*(x)) = c(x)\,\exp\left\{ -\int_{z_0}^{x} \frac{1}{k(u)}\,du \right\}, \tag{1.16}$$

with

$$\lim_{x \to x^*} c(x) = c > 0.$$

and where the function $k(.)$ and the constant z_O are defined in equation 1.16.

The endpoint x^* may be finite or infinite in this domain of attraction. The corresponding normalizing sequences may be chosen as

$$a_n = Q\left(\frac{1}{n}\right), \qquad b_n = \frac{1}{\bar{F}(a_n)} \int_{a_n}^{x^*} (1 - F)(t)\,dt.$$

The explicit characterization of the quantile function in this domain of attraction is more complex than in the Fréchet or Weibull cases. However, it is possible to obtain a specific expression for a subfamily of distributions known as **Weibull-type tail laws**. We first define this family of distributions before providing a characterization of the corresponding quantile function.

Definition 1.4. *A distribution F is said to have a* Weibull-type tail *if there exists $\beta > 0$ such that the survival function satisfies*

$$-\log(\bar{F}(x)) \in \mathcal{RV}_{\frac{1}{\beta}}.$$

The index β is called the Weibull tail index and describes the rate of decay of the tail of the distribution. If we assume that the auxiliary function $k(\cdot)$ used in equation (1.16) satisfies $k' \in \mathcal{RV}_{-\frac{1}{\beta}}$, then the survival function can be written as

$$\bar{F}(x) = \exp\left(-x^{-\frac{1}{\beta}} L(x)\right), \qquad \text{where} \quad L \in \mathcal{RV}_0.$$

Basis on this forms, the quantile of order $\tau \in (0,1)$ is then

$$Q(\tau) = (-\log(1-\tau))^{\beta}\, \mathcal{L}(-\log(1-\tau)), \qquad \mathcal{L} \in \mathcal{RV}_0.$$

This expression shows that, unlike the classical Fréchet and Weibull domains, distributions in the Gumbel domain exhibit logarithmic growth of the extreme quantile, governed by the slow regularity of the function $\mathcal{L}(\cdot)$. They include, in particular, common distributions such as the normal, exponential, and Gamma distributions, all of which have a Weibull-type tail. Knowledge of the structure of these tails and the shape of the associated quantile plays an essential role in modeling extreme values in rapidly decreasing distributions.

We now introduce the notion of the extreme conditional quantile, before reviewing the main estimation methods proposed in the literature and analyzing their limitations. These issues will form the basis of the methodological developments presented in this thesis.

1.7 Extreme Conditional Quantiles

Let Y be the univariate response variable of interest and X a p-dimensional covariate vector. We observe a random sample $\{(y_i, x_i), i = 1, \ldots, n\}$ drawn from the joint distribution of (Y, X). For a probability level $\tau_n \in (0,1)$, the conditional τ_n-quantile of Y given $X = x$, denoted by $Q_Y(\tau_n \mid x)$, is defined as the solution to

$$F_{Y|X=x}(Q_Y(\tau_n \mid x)) = \tau_n, \tag{1.17}$$

where $F_{Y|X=x}$ denotes the conditional distribution function of Y given $X = x$. We refer to *extreme conditional quantiles* when $\tau_n \to 1$ as $n \to +\infty$. The estimation of lower-tail quantiles ($\tau_n \to 0$) can be handled similarly by working with the transformed variable $-Y$.

Estimating extreme conditional quantiles is central in the analysis of rare events, particularly in finance, insurance, hydrology, and climatology. The literature offers several classes of methods for estimating quantiles when τ approaches 1, differing in terms of structural assumptions on the conditional distribution, degree of parametrization, and the way dependence between Y and the covariates is modeled. In this section, we provide a unified review of the main parametric, semi-parametric, non-parametric approaches, as well as extreme quantile regression methods.

1.8 Different Methods for Estimating Extreme Conditional Quantiles

1.8.1 Parametric Approaches

Parametric approaches assume that the conditional distribution of Y given $X = x$ belongs to a known family of distributions. In the extreme-value framework, two models dominate: the generalized extreme value distribution (GEV), derived from the block-maxima approach, and

the generalized Pareto distribution (GPD), used for excesses over a high threshold ((Coles, 2001), (Embrechts et al., 1997)).

For a level $\tau_n \to 1$, the conditional quantile can be approximated by replacing $F_{Y|X=x}$ in (1.17) with a GEV distribution with parameters $\mu(x)$, $\sigma(x)$, and $\xi(x)$, or with a GPD having parameters $\sigma(x)$ and $\xi(x)$. Dependence on x is incorporated by assuming that the parameters follow a predefined parametric structure, such as linear or polynomial forms. The framework proposed by (Davison and Smith, 1990), where

$$\sigma(x) = \exp(\beta_0 + \beta_1 x), \qquad \xi(x) = \gamma_0 + \gamma_1 x,$$

remains one of the most widely used models. Significant extensions were later proposed, notably by (Wang et al., 2012; Wang and Li, 2013). Although efficient when the model is correctly specified, these approaches are sensitive to misspecification, which may induce substantial bias in extreme quantile estimation.

1.8.2 Semi-Parametric Approaches

Semi-parametric approaches aim to reduce the impact of strong structural assumptions on the conditional distribution while relying on extreme-value theory to extrapolate into the tail. They typically rely on estimating an intermediate quantile $Q_x(\tau_0)$, for some moderate probability level τ_0, using classical quantile regression (Koenker and Bassett, 1978), followed by an extrapolation toward $\tau_n \to 1$ based on the approximation

$$Q_x(\tau_n) \approx Q_x(\tau_0) + \frac{\left(\frac{1-\tau_0}{1-\tau_n}\right)^{-\xi(x)} - 1}{\xi(x)}\,\sigma(x),$$

where $\xi(x)$ is the local tail index and $\sigma(x)$ a local scale parameter.

The tail index may be estimated using conditional adaptations of Hill's estimator (Hill, 1975), Pickands' estimator (Pickands III, 1975), or kernel-based methods (Wang et al., 2012), (Daouia et al., 2013). These approaches now form the conceptual foundation of many modern extreme quantile regression techniques.

1.8.3 Non-Parametric Approaches

Non-parametric methods aim to directly estimate the extreme conditional quantile without imposing a specific parametric form on the conditional distribution. They rely on kernel estimators of the conditional distribution function or on resampling techniques adapted to extreme tails. Recent advances include the works of (Daouia et al., 2013), (Gardes et al., 2020), (Allouche et al., 2024), which show that estimation remains feasible even in contexts where

standard parametric assumptions fail. However, these approaches typically suffer from high variance, especially when the covariate dimension is large.

1.8.4 Extreme Quantile Regression

Quantile regression, introduced by (Koenker and Bassett, 1978), provides a flexible alternative for modeling the dependence between Y and X in a nonlinear and heteroscedastic manner. However, when the level τ is very close to 1, the number of available observations in this region becomes extremely small, leading to numerical instability and large variance. We return to these methods in Chapter 2, after introducing the framework of statistical learning.

Overall, existing approaches suffer from structural limitations: sensitivity to the threshold choice in POT methods and Block size in BM methods, dependence on the intermediate level in semi-parametric approaches, difficulty in estimating the tail index under heterogeneity or dependence, and high variance in non-parametric and direct quantile regression approaches. These limitations motivate the development of new methods, based on statistical learning, capable of more accurately modeling the conditional tail, capturing the dependence between covariates and tail parameters, and providing robust estimators for extreme probability levels even when the covariate dimension is large.

CHAPTER DEUX

GENERALITIES ON STATISTICAL LEARNING AND APPLICATION TO QUANTILE REGRESSION

Contents

This chapter relies primarily on standard references in statistical learning, including (Vapnik, 2000), (Hastie, 2017), as well as additional works such as (Azencott, 2022) and (Arlot, 2018).

2.1 Introduction

Over the past few decades, statistical learning has become an important tool for analyzing complex data, especially when the relationships between the variable of interest and the covariates are nonlinear, exhibit heteroscedasticity, or involve a large number of variables. In many fields, such as financial risk management, hydrology, or engineering, these difficulties become even more significant due to a major problem: the scarcity of extreme observations. Classical tools from extreme value theory (EVT), presented in Chapter 1, provide a rigorous description of asymptotic tail behavior, but they may become limited when one seeks to fully incorporate covariate information or to model highly multidimensional phenomena. Statistical learning offers a flexible and powerful collection of methods for exploiting such information, particularly through non-parametric or semi-parametric regression approaches. It provides a general framework for estimating an unknown function linking a response variable Y to a covariate vector X, whether in regression, classification, or unsupervised learning. In a supervised setting, the goal is to estimate an unknown function f relating Y to X, based on a training sample. This estimation relies on the minimization of a theoretical risk, which is approximated in practice by its empirical counterpart. Controlling model complexity, at the heart of the theory introduced by Vapnik (Vapnik, 2000), prevents overfitting and ensures good generalization properties on new data. This theory was introduced by (Vapnik, 2000) and later popularized by Hastie in his book (Hastie, 2017). The objective of this chapter is therefore twofold. On the one hand, we present the theoretical foundations of statistical learning necessary for understanding the regression methods that we will use later. On the other hand, we show how these tools relate to issues specific to extremes, and how they motivate the extreme quantile regression methods developed later in this manuscript.

2.2 General Principle of Supervised Learning

Consider a training dataset

$$\mathcal{B}_n = \{(X_i, Y_i)\}_{i=1}^n,$$

where $X_i \in \mathcal{X} \subset \mathbb{R}^p$ denotes a vector of covariates and $Y_i \in \mathcal{Y} \subset \mathbb{R}$ the associated output. The pairs (X_i, Y_i) are assumed to be independent and identically distributed according to an unknown distribution $\mathbb{P}$ associated with the random couple (X, Y). The following terminology is commonly used

- (X_i, Y_i): example or observation,
- X_i: input variable, explanatory variable, covariate, or feature,
- Y_i: output variable, response, or target.

In supervised learning, the aim is to construct a prediction function which, when applied to a new observation x_0 outside the training sample, provides an accurate estimate of the corresponding output variable y_0. To formalize this, we introduce the following notions.

Definition 2.1. *A **prediction rule** is any measurable function $f : \mathcal{X} \to \mathcal{Y}$ that assigns a predicted value $f(x)$ to any input $x \in \mathcal{X}$.*

To evaluate the quality of a prediction rule, we use a loss function that measures the discrepancy between the real value y and the prediction $f(x)$.

Definition 2.2. *A **loss function** is a measurable mapping $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ satisfying*

1. $L(y,y) = 0$ *for all* $y \in \mathcal{Y}$,
2. $L(y,y') > 0$ *for all* $y \neq y'$.

Example 2.1. *Among the most commonly used loss functions, we mention*

- *the absolute loss:* $L(y,y') = |y-y'|$,
- *the L^p loss:* $L(y,y') = |y-y'|^p$ *for* $p \geq 1$ *(with* $p = 2$ *in the quadratic case),*
- *the quantile loss*

$$\rho_\tau(y,q) = \begin{cases} \tau(y-q), & y \geq q, \\ (\tau-1)(y-q), & y < q. \end{cases}$$

The performance of a prediction rule f is measured through its theoretical risk, also called the *generalization error* (Vapnik, 2000), defined as

$$R_{\mathbb{P}}(f) = \mathbb{E}_{\mathbb{P}}\left[L(Y, f(X))\right]. \tag{2.1}$$

If $\mathcal{F}$ denotes the considered model class, that is, the set of admissible prediction rules, the best rule (also called the *oracle* or *Bayes predictor*) is defined as

$$f^* \in \arg\min_{f \in \mathcal{F}} R_{\mathbb{P}}(f). \tag{2.2}$$

This function depends on the joint distribution of (X,Y), which is unknown in practice, making its use difficult. Note also that this target function is not necessarily unique. The case where we consider the quadratic loss provides an important example of a target function.

Remark 2.1. *For the quadratic loss, the target function is given by the conditional expectation:*

$$f^*(x) = \mathbb{E}[Y \mid X = x].$$

In practice, since f^* is generally inaccessible, it must be approximated from the training sample using an appropriate estimation algorithm.

2.3 Learning Algorithms

A learning algorithm, also called a prediction algorithm, is a measurable mapping

$$\widehat{f} : (\mathcal{X} \times \mathcal{Y})^n \longrightarrow \mathcal{F}$$

that associates to any training set $\mathcal{B}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ a trained prediction rule, denoted $\widehat{f}(\cdot\,; \mathcal{B}_n)$. In other words, a prediction algorithm transforms a sample into a function meant to predict new observations.

The quality of a learning algorithm is naturally evaluated using the theoretical risk of the estimated function. We consider the difference

$$R_{\mathbb{P}}\Big(\widehat{f}(\cdot\,; \mathcal{B}_n)\Big) - R_{\mathbb{P}}(f^*),$$

called the *excess risk*. Since this quantity is random (as it depends on $\mathcal{B}_n$), several measures are used to assess the performance of the learning algorithm, notably

1. the *expected risk*:
$$\mathbb{E}\big[R_{\mathbb{P}}(\widehat{f}(\cdot\,; \mathcal{B}_n))\big] - R_{\mathbb{P}}(f^*),$$
where the expectation is taken with respect to the distribution generating the sample;

2. the *excess risk probability*:
$$\mathbb{P}\Big(R_{\mathbb{P}}(\widehat{f}(\cdot\,; \mathcal{B}_n)) - R_{\mathbb{P}}(f^*) > \varepsilon\Big), \qquad \varepsilon > 0.$$

For notational simplicity, we will simply write $\widehat{f}$ (resp. $\widehat{f}(x)$) instead of $\widehat{f}(\cdot\,; \mathcal{B}_n)$ (resp. $\widehat{f}(x; \mathcal{B}_n)$).

A fundamental criterion for assessing the quality of an algorithm is its ability to approximate the target function f^* as the sample size tends to infinity. This property, which ensures good generalization, is formalized by the notion of consistency.

Definition 2.3. *(Arlot, 2018)*
Let $(\widehat{f}_n)_n$ be a sequence of learning algorithms.

1. *We say that $\widehat{f}_n$ is **weakly consistent** for the distribution $\mathbb{P}$ if*
$$\mathbb{E}\Big[R_{\mathbb{P}}(\widehat{f}_n)\Big] \longrightarrow R_{\mathbb{P}}(f^*) \quad \text{as } n \to +\infty.$$

2. *We say that $\widehat{f}_n$ is **strongly consistent** for the distribution $\mathbb{P}$ if*
$$R_{\mathbb{P}}(\widehat{f}_n) \xrightarrow{a.s.} R_{\mathbb{P}}(f^*),$$

where f^* denotes a target function.

Definition 2.4. *An algorithm is said to be consistent for a family of distributions $\mathcal{P}$ if it is consistent for all $\mathbb{P} \in \mathcal{P}$.*

Definition 2.5. *An algorithm is universally consistent if it is consistent for every probability distribution on $\mathcal{X} \times \mathcal{Y}$.*

The ultimate goal is to construct an algorithm that minimizes the generalization error (2.1), which depends on the unknown distribution $\mathbb{P}$. Consistency, especially universality, provides a strong theoretical guarantee: it ensures that the algorithm will converge to the optimal rule regardless of the mechanisms generating the data. When $\mathbb{P}$ is unknown, it is natural to replace it with its empirical counterpart based on the sample $\mathcal{B}_n$, which leads to the minimization of the empirical risk. The next section describes the fundamental properties of algorithms built according to this principle.

2.4 Empirical Risk Minimization Algorithms

Throughout the framework of statistical learning, the fundamental objective is to approximate as closely as possible the optimal prediction function, denoted f^*, defined as the minimizer of the generalization error

$$R_{\mathbb{P}}(f) = \mathbb{E}_{(X,Y)\sim\mathbb{P}}[L(Y, f(X))].$$

When the distribution $\mathbb{P}$ is unknown, which is the typical situation in practice, one only has access to a training sample $\mathcal{B}_n = \{(X_i, Y_i)\}_{i=1}^n$. The risk is then estimated using the empirical risk

$$\hat{R}_n(f; \mathcal{B}_n) = \frac{1}{n}\sum_{i=1}^{n} L(Y_i, f(X_i)). \tag{2.3}$$

2.4.1 Inductive Principle

To approximate the generalization error $R_{\mathbb{P}}$, statistical learning theory formalizes a fundamental *inductive principle* composed of two essential ideas:

- replacing the ideal objective $R_{\mathbb{P}}(f)$ by its observable approximation, the empirical risk $\hat{R}_n(f)$;
- approximating the minimizer of $R_{\mathbb{P}}$ by the minimizer of $\hat{R}_n$ over a constrained function class.

This paradigm forms the basis of a wide range of supervised learning methods: least squares regression, maximum likelihood, logistic regression, support vector machines, among others

(Vapnik, 2000). The empirical risk is indeed an unbiased estimator of the true risk, making it natural to choose the prediction rule that minimizes $\hat{R}_n$.

Definition 2.6. *Let $\mathcal{H} \subset \mathcal{F}$ be a class of predictors (also called a* model*) and $\mathcal{B}_n$ a training sample. An empirical risk minimization (ERM) algorithm over $\mathcal{H}$ is defined by*

$$\hat{f}_{\mathcal{H}} \in \arg\min_{f \in \mathcal{H}} \hat{R}_n(f; \mathcal{B}_n).$$

By abuse of notation, we will denote this minimizer by $\hat{f}_{\mathcal{H}}$ when the sample is clear from context.

The choice of the class $\mathcal{H}$ is crucial: a class that is too restrictive leads to a high approximation bias, whereas an overly rich class results in overfitting. The optimal trade-off is classically interpreted through the bias–variance decomposition.

2.4.2 Consistency and Excess Risk

The property of *consistency* characterizes the ability of an algorithm to learn the best function in the model $\mathcal{H}$ as the sample size grows.

Definition 2.7. *(Vapnik, 2000) An ERM algorithm is **consistent** for a model $\mathcal{H}$ and a distribution $\mathbb{P}$ if*

$$R_{\mathbb{P}}(\hat{f}_{\mathcal{H}}) \xrightarrow{\mathbb{P}} \inf_{f \in \mathcal{H}} R_{\mathbb{P}}(f) \quad \text{and} \quad \hat{R}_n(\hat{f}_{\mathcal{H}}) \xrightarrow{\mathbb{P}} \inf_{f \in \mathcal{H}} R_{\mathbb{P}}(f).$$

In other words, the empirical risk and the true risk converge to the same limit as $n \to +\infty$.

The excess risk is classically decomposed into two components

$$\underbrace{R_{\mathbb{P}}(\hat{f}_{\mathcal{H}}) - R_{\mathbb{P}}(f^*)}_{\text{Excess risk}} = \underbrace{R_{\mathbb{P}}(\hat{f}_{\mathcal{H}}) - \inf_{f \in \mathcal{H}} R_{\mathbb{P}}(f)}_{\text{Estimation error}} + \underbrace{\inf_{f \in \mathcal{H}} R_{\mathbb{P}}(f) - R_{\mathbb{P}}(f^*)}_{\text{Approximation error}}.$$

The approximation error decreases when the class $\mathcal{H}$ is enriched, while the estimation error tends to increase, this is the fundamental bias–variance dilemma.

Upper Bound on the Estimation Error

A first general bound is given by the following proposition.

Propriété 3. *(Arlot, 2018)*
For any model $\mathcal{H} \subset \mathcal{F}$ and any empirical minimizer $\hat{f}_{\mathcal{H}}$, we have

$$R_{\mathbb{P}}(\hat{f}_{\mathcal{H}}) - \inf_{f \in \mathcal{H}} R_{\mathbb{P}}(f) \leq 2 \sup_{f \in \mathcal{H}} |R_{\mathbb{P}}(f) - \hat{R}_n(f)|.$$

This bound expresses that the generalization ability is entirely controlled by the uniform deviation between the true risk and its empirical estimate.

2.4.3 Concentration Tools

To bound this deviation, concentration inequalities are used. We recall here Hoeffding's inequality, derived from the following lemma.

Lemma 2.1. *Hoeffding's Lemma*
Let Z be a bounded random variable in $[a,b]$. Then, for any α,

$$\ln \mathbb{E}[e^{\alpha Z}] \leq \alpha \mathbb{E}[Z] + \frac{\alpha^2 (b-a)^2}{8}.$$

Propriété 4 (Hoeffding's Inequalities)**.** *Let $Z_1, \ldots, Z_n$ be independent bounded random variables. Suppose that for each $i \in \{1, \ldots, n\}$, there exist constants $a_i < b_i$ such that*

$$a_i \leq Z_i \leq b_i \quad \textit{almost surely}.$$

Define

$$\mathcal{S}_n = \sum_{i=1}^{n} (Z_i - \mathbb{E}[Z_i]).$$

Then for every $t > 0$,

$$\mathbb{P}(\mathcal{S}_n \geq t) \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^{n} (b_i - a_i)^2} \right)$$

Complete proofs can be found in (Boucheron et al., 2013).

2.4.4 Case of a Finite Model $\mathcal{H}$

Consider a finite model

$$\mathcal{H} = \{f_1, \ldots, f_M\},$$

where f_i represents a prediction rule for all $i \in \{1, \ldots, M\}$.

Propriété 5. *Assume that the loss function L is bounded in $[a,b]$. Then, for all $x \geq 0$,*

$$\mathbb{P}\left(R_{\mathbb{P}}(\hat{f}_{\mathcal{H}}) - \inf_{f \in \mathcal{H}} R_{\mathbb{P}}(f) < (b-a) \sqrt{\frac{2(x + \ln(2M))}{n}} \right) \geq 1 - e^{-x}.$$

This bound describes the generalization capacity of a finite model: the larger M is, the more flexible the model becomes, but the logarithmic penalty $\ln M$ increases accordingly.

Model Selection via Penalized Criteria

Given a collection of models $\mathcal{C}$, the goal is to identify the model that offers the best trade-off between goodness of fit and complexity. A classical approach consists in selecting the model that minimizes a penalized criterion

$$\hat{\mathcal{H}} = \arg\min_{\mathcal{H}\in\mathcal{C}} \left\{ \hat{R}_n\big(\hat{f}_{\mathcal{H}}\big) + \text{pen}(\mathcal{H}) \right\}.$$

The penalty term $\text{pen}(\mathcal{H})$ controls the effective complexity of the model, thereby limiting the risk of overfitting by counterbalancing an overly precise fit of the training data. Designing a relevant penalty is a central issue in model selection theory. Many classical criteria fall within this framework, such as the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), cross-validation, or approaches stemming from the principle of structural risk minimization. Regularization methods also play a major role: notable examples include Ridge regression (Hoerl and Kennard, 1970), the Lasso (Tibshirani, 1996), the Smoothly Clipped Absolute Deviation (SCAD) penalty (Fan and Li, 2001), adaptive Lasso (Zou and Hastie, 2005), and the Minimax Concave Penalty (MCP) (Zhang, 2010). These works have significantly contributed to the development of effective penalties for selection and sparsity in modern statistical models. For more details, see the book (Hastie, 2017).

Model selection fits into a broader perspective aimed at controlling the generalization ability of statistical learning methods. Once a model is chosen, it becomes essential to assess the quality of the associated estimation procedure, particularly in terms of stability, approximation capacity of the target function, and asymptotic behavior as the sample size grows. These considerations motivate the study of consistency properties, approximation guarantees, and empirical criteria for rigorously comparing algorithms. The next section presents some statistical learning methods commonly used in practice, with a focus on those that will serve as the foundation for our approach to extreme quantile regression.

2.5 Some Statistical Learning Methods

Suppose we observe a dataset

$$\{(X_i, Y_i)\}_{i=1}^n,$$

where $X_i \in \mathcal{X}$ is a covariate vector and $Y_i \in \mathcal{Y}$ is the response variable. The goal of regression is to find a function representing the target function $\eta : \mathbb{R}^p \to \mathbb{R}$ such that

$$\eta(x) \approx \mathbb{E}[Y \mid X = x].$$

In other words, we aim to approximate the regression function $x \longmapsto \mathbb{E}[Y \mid X = x]$. In the following, we discuss several learning methods used to estimate the regression function.

2.5.1 Regression Trees

Regression trees form a widely used family of supervised learning methods, for both regression and classification. In the regression setting, the objective is to estimate the function η defined above. The principle of a regression tree is to partition the feature space $\mathcal{X}$ into homogeneous regions (typically hyperrectangles) and to assign a constant prediction to each region, usually the local average of the target variable. Among the most commonly used algorithms is the CART algorithm (Classification and Regression Trees), introduced by (Breiman et al., 1984). Consider a training dataset $\{(x_i, y_i)\}_{i=1}^n$ with $x_i = (x_i^1, \ldots, x_i^p) \in \mathcal{X}$. The main steps in building a regression tree to approximate η can be summarized as follows.

1. **Selection of the variable and split threshold.**
   The algorithm starts by selecting the best feature along which the data can be divided into two subsets. In other words, it selects the best split of the space, represented by a pair (j,t) consisting of a component j of the covariate vector and a threshold t used to divide the data into two regions. The notion of "best" is generally linked to the reduction in the variance of the target variable within the resulting subsets. Variance is used because the goal is to minimize the dispersion of the target variable within each region. The best feature is thus a pair (j,t), where j is the index of the selected variable (i.e., the j^{th} component of x_i) and t is the threshold value of x_i^j used to split the data while reducing the variance of the target variable in the two resulting regions. If we consider the j^{th} coordinate and t as the splitting point (with t belonging to the support of x_i^j), the two regions are defined as

   $$R_1(j,t) = \{x_i \mid x_i^j \leq t\} \quad \text{and} \quad R_2(j,t) = \{x_i \mid x_i^j > t\}.$$

   To determine the optimal pair (j,t), the first step is to fixes a coordinate j and then varies t over the support of x_i^j. For each value of t, the variance of the y_i within the two regions $R_1(j,t)$ and $R_2(j,t)$ is evaluated. This procedure is repeated for all coordinates $j = 1, \ldots, p$, and the pair minimizing the total variance in both regions is retained. This corresponds to solving the optimization problem

   $$\arg\min_{j,t} \left[\sum_{x_i \in R_1(j,t)} (y_i - \text{mean}\{y_i | x_i \in R_1(j,t)\})^2 + \sum_{x_i \in R_2(j,t)} (y_i - \text{mean}\{y_i | x_i \in R_2(j,t)\})^2 \right]. \tag{2.4}$$

2. **Data splitting.**

Once the pair (j,t) is determined, the data are split into two subsets corresponding to $R_1(j,t)$ and $R_2(j,t)$, forming two child nodes of the tree.

3. **Recursive procedure.**
   The previous steps are then recursively repeated for each node. The process stops when a stopping condition is reached, such as a maximum depth or a minimum number of observations per leaf (i.e., a terminal node that will not be split further).

4. **Prediction.**
   Once the tree is built, to predict the value y for a new observation x, it is passed through the tree following the appropriate branches until it reaches a leaf. The prediction is the average of the responses of the observations contained in that leaf. Denoting by $R(x) \subset \mathcal{X}$ the leaf containing x, the prediction is

$$\hat{y} = \hat{\eta}(x) = \frac{1}{|R(x)|} \sum_{\{i: x_i \in R(x)\}} y_i,$$

   where $|E|$ denotes the cardinality of E.

Example 2.2. *Figure 2.1 illustrates an example of partitioning in the case $p = 2$, resulting in five regions. The resulting model predicts a constant value C_m on each region R_m, i.e.,*

$$\hat{\eta}(x) = \sum_{m=1}^{5} C_m \mathbb{1}_{\{(x_1,x_2) \in R_m\}}.$$

Regression trees offer several advantages: simplicity, interpretability, and the ability to capture nonlinear relationships without heavy preprocessing. However, they suffer from high instability, small changes in the data can drastically alter the tree structure, and their predictive performance is often limited. They are therefore considered *weak learners*. These limitations motivate the use of ensemble methods, which stabilize and significantly improve performance.

2.5.2 Ensemble learning Methods

Ensemble methods aim to improve predictive accuracy by combining multiple models, based on the idea that an aggregate of weak learners can produce a robust, accurate, and generalizable model. They consist in training several regression models and aggregating their predictions. Ensemble methods have become essential in statistical learning due to their ability to reduce overfitting, improve stability, and adapt to various data structures (Hastie, 2017).

Among the most widely used techniques are *Boosting* and *Bagging* (Bootstrap Aggregation). In what follows, we focus on Bagging, as it is the building block of random forests, which play a central role later in this chapter.

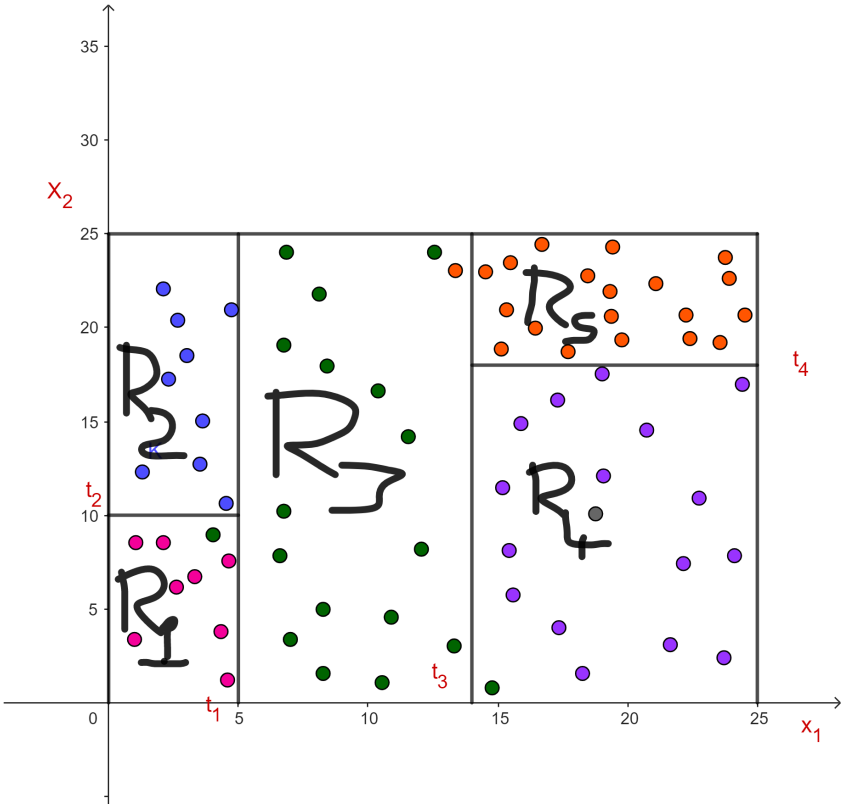

Figure 2.1: Partition of a sample with predictor dimension $p = 2$ into five regions: t_1, t_3 split X_1, and t_2, t_4 split X_2.

2.5.3 Bagging predictors

The bagging method introduce by (Breiman, 1996) is an ensemble learning method consisting of training multiple learning models in parallel, each on a different bootstrap sample[1] drawn from the training set. The final prediction is obtained by averaging the predictions of the models (in regression) or by majority vote (in classification). The key idea of bagging (bootstrap aggregating) is to generate multiple independent learning models and to combine their predictions in order to obtain a more stable and better–generalizing estimator. As an illustration, consider a regression setting with a training sample $\mathcal{D} = \{(x_i, y_i)\}_{i=1,\ldots,n}$. The goal is to estimate a predictive function able to predict, for any new observation x independent of the training data, an accurate estimate of the associated output y, denoted $\hat{y} = \hat{\eta}(x)$. Bagging proceeds by constructing B bootstrap samples $\mathcal{D}^{*b}$, $b = 1, \ldots, B$, drawn with replacement from $\mathcal{D}$, and by fitting the learning algorithm separately on each bootstrap sample (decision trees are typically used as base learners). Let $\hat{\eta}^{*b}(x)$ denote the prediction produced for an input x by the b-th

[1] **Bootstrap sample**: a sample of size n obtained by randomly drawing with replacement n observations from the original dataset of size n.

model in the ensemble. The bagging predictor is then defined as

$$\hat{\eta}_{\text{bag}}(x) = \frac{1}{B}\sum_{b=1}^{B} \hat{\eta}^{*b}(x).$$

The main effect of bagging is a reduction in the variance of the individual estimators. This variance reduction leads to improved stability and predictive performance compared with the base models taken individually. Bagging is therefore particularly effective for learning algorithms characterized by high variance and low bias, such as decision trees.

2.5.4 Random Forests

Random forests are a type of learning method used for both classification and regression (Breiman, 2001). They belong to the family of ensemble learning methods and provide a non-parametric estimator of the conditional mean, i.e., the regression function η. Given $(X_i, Y_i) \in \mathcal{X} \times \mathbb{R}$, with $\mathcal{X} \subset \mathbb{R}^p$, random forests estimate $\eta(x) = \mathbb{E}(Y_i|X_i = x)$. A random forest consists of an ensemble of decision trees, each constructed independently, similar to bagging. A key difference between random forests and standard bagging lies in how the predictor variables are used to split the nodes of each tree. Rather than considering all variables at each split, a random subset of variables is selected. This additional randomization increases the diversity among trees, enhancing the model's generalization performance. Specifically, a random forest aggregates B trees built in parallel from bootstrap samples drawn from the original training set. Unlike standard CART trees, at each node, $q < p$ variables are randomly selected from which the splitting variable is chosen.

Let $\eta_b(x)$ denote the prediction of the b-th tree for a data point $x \in \mathcal{X}$. In regression, this prediction can be written as

$$\hat{\eta}_b(x) = \sum_{i=1}^{n} \frac{\mathbb{1}_{\{X_i \in R_b(x)\}} Y_i}{|\{i : X_i \in R_b(x)\}|}, \qquad b = 1, \ldots, B,$$

where $R_b(x) \subset \mathbb{R}^p$ denotes the region of tree b containing x. The final prediction of the random forest is given by

$$\begin{aligned} \hat{\eta}(x) &= \frac{1}{B}\sum_{b=1}^{B} \eta_b(x) \\ &= \sum_{i=1}^{n} w_n(x, X_i) Y_i \end{aligned}$$

with

$$w_n(x, X_i) = \frac{1}{B}\sum_{b=1}^{B} \frac{\mathbb{1}_{\{X_i \in R_b(x)\}}}{|\{i : X_i \in R_b(x)\}|}. \tag{2.5}$$

The weights $w_n(x, X_i)$ act as similarity measures and satisfy $\sum_{i=1}^{n} w_n(x, X_i) = 1$.

It is important to note that this regression function is estimated by minimizing the squared error loss during the construction of each tree. Random forests are widely recognized for their ability to produce robust models that generalize well to datasets independent of the training sample. The diversity introduced by random variable selection and the generation of bootstrap samples reduces the risk of overfitting. These models achieve competitive performance compared to more complex methods, while being less sensitive to hyperparameter settings. Moreover, they remain relatively easy to use, requiring fewer hyperparameters to tune. Owing to their robustness, reliable predictive performance, capacity to handle diverse data types, and ease of use, random forests have become a popular choice in a wide range of statistical learning applications.

Example 2.3. *Figure 2.2 illustrates the prediction process of a random forest composed of three trees.*

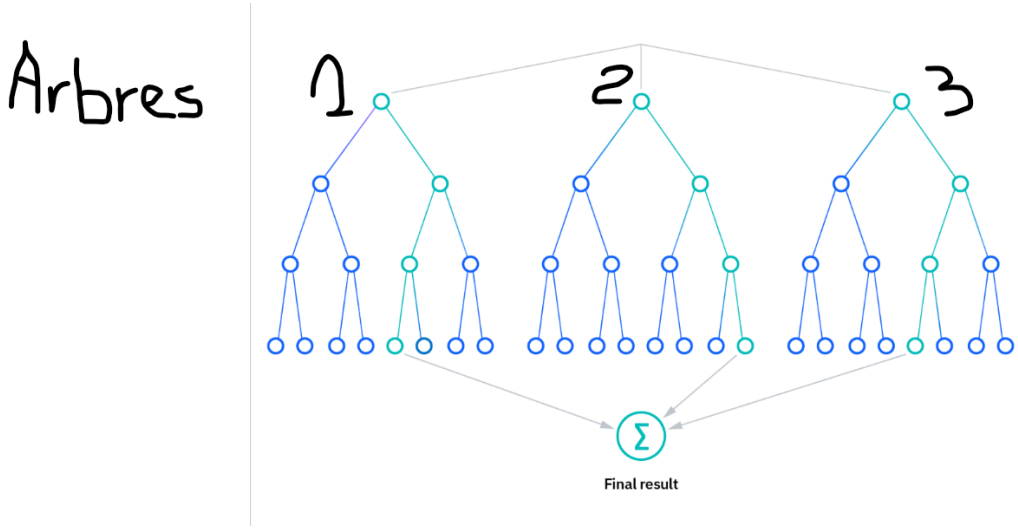

Figure 2.2: Example of a random forest consisting of 3 trees

The following subsection presents an adaptation of this learning method to the quantile regression setting, which will be employed in this thesis to address the challenges of extreme quantile regression.

2.5.5 Generalized Random Forests

Objective

The Generalized random forests, introduced by (Athey et al., 2019), extend the scope of classical random forests by enabling the estimation of statistical parameters defined as solutions to local estimating equations. Unlike standard regression forests (Breiman, 2001), which aim

to estimate the conditional mean $\mathbb{E}[Y \mid X = x]$, generalized random forests (grf) enable the estimation of a potentially vector-valued parameter $\theta(x)$ defined as the solution to

$$\mathbb{E}\left[\psi\big(Y, X; \theta(x)\big) \,\big|\, X = x\right] = 0,$$

where $\psi(\cdot)$ is an estimating function chosen according to the task. The key contribution of generalized random forests lies in the use of similarity weights $w_i(x)$ derived from the forest structure to build a localized empirical version of the above equation. Thus, the estimator $\hat{\theta}(x)$ is obtained as the solution to

$$\sum_{i=1}^{n} w_i(x)\, \psi\big(Y_i, X_i; \theta(x)\big) = 0.$$

This unified formulation allows the estimation of a wide range of statistical quantities, including

- conditional quantiles (quantile forests);
- heterogeneous treatment effects (causal forests);
- local least squares and various semiparametric models;
- more generally, any quantity that can be defined by a local estimation equation.

This approach combines the flexibility of nonparametric methods with the statistical stability provided by honesty and controlled subsampling of forests, thus providing a robust framework for estimating complex parameters that are locally dependent on x.

In the following, we present in detail the functioning of these so-called *honest* random forests, as introduced by (Athey et al., 2019), emphasizing the mechanisms of tree construction, the role of subsampling, and the definition of similarity weights, which play a central role in many estimation procedures based on this learning method.

Subsampling and Tree Honesty

The random forests considered here rely on systematic subsampling without replacement. From an initial sample of size n, each tree is built from a subsample

$$S_b \subset \{1, \ldots, n\}, \qquad |S_b| = s < n,$$

where the subsample size s satisfies:

$$s \to \infty, \qquad \frac{s}{n} \to 0 \quad \text{as } n \to \infty.$$

This condition ensures that each tree uses an increasing amount of information while remaining weakly correlated with the others. To ensure honesty, each subsample S_b is split into two disjoint sets

- a set R_1^b used to determine the structure of the tree (split selection);
- a set R_2^b used only for estimation within the leaves.

Most often, these two sets have equal size, providing a balance between tree depth and estimation stability.

Split Properties and Tree Depth

Tree construction relies on a randomized procedure for partitioning the covariate space. Splits satisfy two essential properties

1. **Symmetry**: the procedure is invariant under permutations of the observations.
2. **Balancing**: at each split, a minimum proportion $\omega > 0$ of observations is sent to each child node.

Candidate splitting variables are selected randomly, each variable having a probability lower bounded by a parameter $\pi > 0$. This randomization prevents domination of the tree structure by a small number of variables.

In theory, the full set of $\binom{n}{s}$ possible trees could be exploited; in practice, a finite number B is used, chosen large enough so that Monte Carlo error becomes negligible.

Similarity Weights $w_i(x)$

A fundamental feature of generalized random forests is the interpretation of predictions as weighted linear combinations of the training observations. For a new observation x, the similarity weights are defined by

$$w_i(x) = \frac{1}{B} \sum_{b=1}^{B} w_{i,b}(x), \tag{2.6}$$

where $w_{i,b}(x)$ denotes the contribution of observation i in tree b. For a given tree, observation i contributes to the prediction if it belongs to the leaf $L_b(x)$ containing x

$$w_{i,b}(x) = \frac{\mathbb{1}_{\{X_i \in L_b(x),\, i \in R_2^b\}}}{|L_b(x)|},$$

where

$$|L_b(x)| = \sum_{i=1}^{n} \mathbb{1}_{\{X_i \in L_b(x),\, i \in R_2^b\}}.$$

By construction, leaf sizes are controlled by the allowable depths and the balancing parameter, preventing degeneracy of the weights. The diameter of a leaf, defined by

$$\operatorname{diam}(L_b(x)) = \sup_{y \in L_b(x)} \|y - x\|_2,$$

plays a key role in consistency analysis. It has been shown in (Wager and Athey, 2018) that this diameter decreases sufficiently fast to obtain asymptotic guarantees for honest forests. These properties are essential for establishing the asymptotic validity of estimators based on this learning method, especially in the presence of complex nonlinear relationships or high-dimensional data structures.

We now introduce quantile regression methods, before reviewing estimation procedures based on statistical learning and extreme value theory.

2.6 Quantile Regression

Classical regression, introduced by (Galton, 1889), aims to model the relationship between an explanatory variable X and a response variable Y through a deterministic function f such that

$$Y = f(X) + \varepsilon,$$

where ε denotes a random noise term independent of X. The main objective is then to estimate the conditional mean of Y given $X = x$. Numerous parametric methods (linear regression, polynomial regression, etc.) and nonparametric methods have been developed within this framework. However, these approaches exhibit limitations in the presence of extreme values or pronounced heteroscedasticity. The classical solution consisting in removing outlying observations often leads to a significant loss of information. To obtain tools that are more robust and more informative regarding the conditional structure of the distribution of Y, (Koenker and Bassett, 1978) proposed quantile regression (QR). This method relies on estimating the conditional quantiles of Y given $X = x$ and enables a richer analysis than the sole conditional mean. QR is now widely used, notably in the study of wage inequality (Angrist et al., 2006), insurance pricing (Abad et al., 2014), and precipitation modeling (Bagirov et al., 2017). This model is particularly useful and powerful for analyzing data with asymmetric distributions or containing extreme values.

For $\tau \in (0,1)$, the conditional quantile of order τ is defined by

$$Q_{Y|X=x}(\tau) = \inf\{y : F_{Y|X=x}(y) \geq \tau\},$$

and, when $F_{Y|X=x}$ is continuous,

$$Q_{Y|X=x}(\tau) = F^{-1}_{Y|X=x}(\tau). \tag{2.7}$$

The quantile regression model can thus be written as

$$Y = Q_{Y|X=x}(\tau) + \varepsilon,$$

where ε is an error term. Another characterization of the conditional quantile is given by:

$$Q_Y(\tau \mid X = x) = \arg\min_{q \in \mathcal{Y}} \mathbb{E}\big[\rho_\tau(Y - q) \,\big|\, X = x\big], \tag{2.8}$$

with $\rho_\tau(c) = c(\tau - \mathbb{1}_{c<0})$ the asymmetric loss function.

Many contributions have enriched the literature on conditional quantile estimation, both parametric (Wang et al., 2012), (Chernozhukov, 2005), (Koenker and Hallock, 2001), (Angrist et al., 2006) and nonparametric (Allouche et al., 2024), (Daouia et al., 2013), (Benziadi et al., 2016), (Meinshausen and Ridgeway, 2006), (Takeuchi et al., 2006), (Ye and Padilla, 2020), (Dabrowska, 1992). For example, (Koenker and Hallock, 2001) analyzed the evolution of wages in the United States using QR, revealing an increase in inequality in the upper part of the distribution. (Yu and Moyeed, 2001) introduced a Bayesian approach to quantile regression, allowing the treatment of missing data and variable selection. Extensions to the spatial context have also been proposed, for instance in (Laksaci and Maref, 2009) and (Dabo-Niang and Laksaci, 2012).

The previously mentioned methods allow efficient estimation of conditional quantiles for intermediate probability levels. However, when dealing with quantiles located in the tail of the distribution, i.e., when τ is close to 1, these approaches quickly reach their limits. In particular, classical quantile regression models—parametric, nonparametric, or based on statistical learning methods—suffer from high variability in the extreme region, instability due to the scarcity of relevant observations, and a lack of theoretical guarantees adapted to the asymptotic behavior of conditional tails. To overcome these difficulties, it is natural to rely on Extreme Value Theory (EVT), which provides a rigorous framework for modeling distributions in regions of scarcity. Integrating EVT into quantile regression models thus enables the construction of estimators that are better suited for the case where $\tau \to 1$, more robust, and endowed with well-established asymptotic properties. The next section therefore presents extreme quantile regression methods based on EVT and statistical learning techniques. These approaches combine the flexibility of nonparametric models and the theoretical strength of extreme value results to deliver effective estimators in the conditional tails.

2.7 Extreme Quantile Regression Based on EVT and Statistical Learning Methods

Let $Y \in \mathcal{Y} \subset \mathbb{R}$ be a response variable and $X \in \mathcal{X} \subset \mathbb{R}^p$ a vector of covariates. Extreme quantile regression focuses on estimating the conditional quantile $Q_Y(\tau \mid X = x)$ when the probability level τ tends to 1, i.e., when studying the behavior of the upper tail of the conditional distribution. Considering the characterization (2.8) and a sample $\mathcal{D}_n = \{(X_i, Y_i)\}_{i=1}^n$, a natural estimator of the conditional quantile is given by

$$\hat{Q}_{Y|X=x}(\tau) = \arg\min_{q \in \mathcal{Y}} \sum_{k=1}^{n} w_n(x, X_k)\, \rho_\tau(Y_k - q). \tag{2.9}$$

Existing methods differ mainly in the construction of the weights $w_n(x, X_k)$, which measure the similarity between x and the observations X_k. However, when the dimension p is large, these methods encounter the curse of dimensionality. Statistical learning approaches then offer suitable solutions.

To our knowledge, the first occurrences of quantile regression in the statistical learning literature appear in (Le et al., 2005) and (Meinshausen and Ridgeway, 2006). The latter introduces *Quantile Regression Forests* (QRF), an extension of random forests for estimating conditional quantiles. Since

$$F_{Y|X=x}(y) = \mathbb{E}\big[\mathbf{1}_{\{Y \le y\}} \mid X = x\big],$$

the estimator of the conditional distribution proposed in (Meinshausen and Ridgeway, 2006) is given by

$$\hat{F}(y \mid X = x) = \sum_{i=1}^{n} w_n(x, X_i)\, \mathbf{1}_{\{Y_i \le y\}},$$

where the weights $w_n(x, X_i)$ are those defined by random forests (see (2.5)). The estimator of the conditional quantile then directly follows from relation (2.7). Another approach based on random forests is the *Generalized Random Forests* (GRF) method of (Athey et al., 2019), implemented via the R package `grf`. This method optimizes tree construction according to a loss function adapted to the task under consideration. In the context of quantile regression, the loss function ρ_τ is used to build weights $w_n(x, X_i)$ that accurately capture the local heterogeneity of the quantile. Other contributions based on decision trees have also been proposed, such as (Chaudhuri and Loh, 2002).

Although these approaches are effective for estimating non-extreme quantiles and allow handling high-dimensional data, they face difficulties when the quantile of interest becomes extreme, i.e., when $\tau \to 1$. This limitation motivates the use of approaches combining extreme value theory and learning methods.

2.8 Motivations for the Work in This Thesis

Before presenting the contributions developed in this manuscript, it is necessary to highlight the theoretical and methodological challenges encountered when estimating extreme conditional quantiles. The issues described below directly motivate the approaches we propose. A first difficulty associated with quantile regression models arises when the dimension p of the covariate space $\mathcal{X}$ is high and when the relationship between the variable of interest Y and the characteristics X is potentially complex. To address this situation, various approaches from statistical learning have been developed, as discussed in the previous section (see, for example, (Meinshausen and Ridgeway, 2006), (Athey et al., 2019), (Chaudhuri and Loh, 2002)).

The second difficulty arises when estimating quantiles associated with a high probability level τ_n, which requires extrapolation into the distribution tail, as illustrated in Figure 2.3. This extrapolation is delicate because of the scarcity of extreme observations. To illustrate this

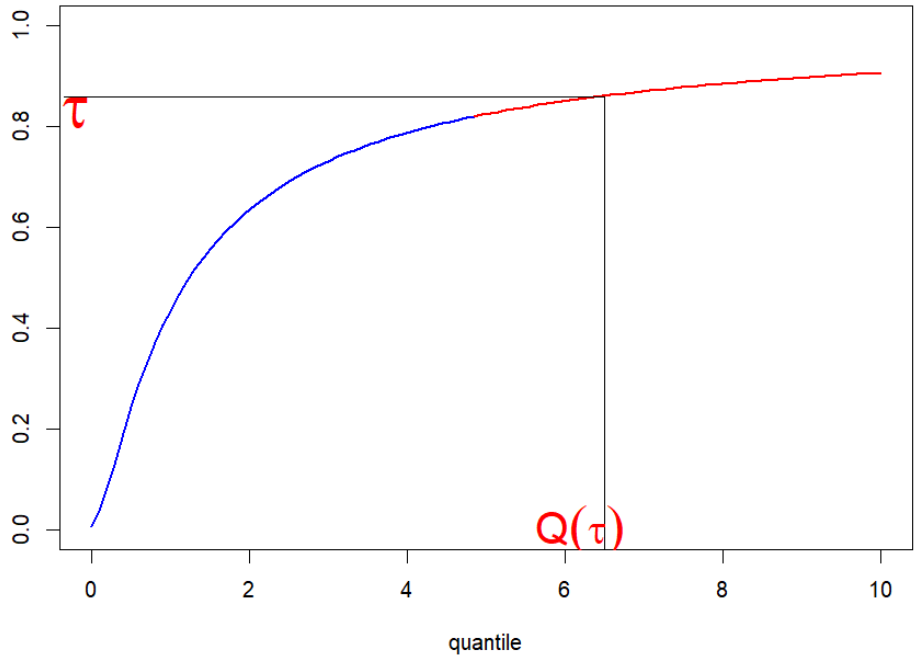

Figure 2.3: Illustration of the probability level as a function of the associated quantile.

issue theoretically in the unconditional framework, consider a sequence of independent and identically distributed random variables $Y_1, \ldots, Y_n$ with an unknown distribution function F. The distribution function of the maximum $Y_{(n)} = \max_i\{Y_i\}$ is given by

$$F_{Y_{(n)}}(y) = F^n(y).$$

The quantile q_{τ_n} of order τ_n, assumed to depend on the sample size, associated with the maximum $Y_{(n)}$ satisfies

$$P(Y_{(n)} \leq q_{\tau_n}) = \tau_n.$$

When $\tau_n \to 1$, we have

$$\begin{aligned} F_{Y_{(n)}}(q_{\tau_n}) &= P\left(Y_{(n)} \leq q_{\tau_n}\right) \\ &= [F(q_{\tau_n})]^n \\ &= (\tau_n)^n \\ &= \exp\left(n \ln(\tau_n)\right). \end{aligned}$$

Near 1, we have $\ln(y) = (y-1) + (y-1)\varepsilon(y-1)$ with $\lim_{h\to 0} \varepsilon(h) = 0$. Thus, when $\tau_n \to 1$, we obtain

$$P(Y_{(n)} \leq q_{\tau_n}) = \exp\left[-n(1-\tau_n)(1+\varepsilon(\tau_n - 1))\right] = \exp\left[-n(1-\tau_n)(1+o(1))\right].$$

We see that the probability that the quantile exceeds the sample maximum depends on the behavior of $n(1-\tau_n)$. Thus, to estimate the extreme quantile, it is important to distinguish between two cases:

- If $n(1-\tau_n) \to +\infty$ as $\tau_n \to 1$, then $P(Y_{(n)} \leq q_{\tau_n}) = 0$.

In this case, we estimate a quantile lying inside the sample. The number of sample elements exceeding q_{τ_n} is $n(1-\tau_n)$, and a natural estimator of q_{τ_n} would be $Y_{(n\tau_n)}$ if $n\tau_n$ is an integer, or $X_{(\lfloor n\tau_n \rfloor + 1)}$ otherwise (where $\lfloor \cdot \rfloor$ denotes the integer part). Here, quantile estimation poses no difficulty. We call an intermediate quantile the largest quantile for which the condition $n(1-\tau_n) \to \infty$ is satisfied when $\tau_n \to 1$. Let the order of this quantile be τ_0, and recall that existing methods for extreme quantile estimation work well for quantiles of order τ_n satisfying $\tau_n \leq \tau_0$.

- *If* $n(1-\tau_n) \in [0, +\infty)$ as $\tau_n \to 1$, then $P(Y_{(n)} \leq q_{\tau_n}) \in (0, 1]$.

The quantile to be estimated lies beyond the sample maximum, or only a few observations exceed it. In this case, estimating q_{τ_n} is no longer possible by simply inverting the empirical distribution function (for example, the empirical cdf satisfies $\hat{F}_n(y) = 1$ as soon as $y \geq Y_{(n)}$). In our context, this corresponds to the extreme conditional quantile $Q_{Y|X=x}(\tau_n)$. It suffices to replace the probabilities in the two cases above by those of the conditional distribution of $Y|X = x$. To address this issue, an estimation method capable of extrapolating beyond the data range is needed. Models based on extreme value theory have been proposed in the literature to tackle this problem; see (Chernozhukov et al., 2017) for a detailed bibliography. The main objective of this thesis is to propose models that address simultaneously the two issues mentioned above.

Several recent works have combined the Peak-Over-Threshold (POT) approach with statistical learning methods. For instance, (Velthoen et al., 2023) propose an estimator based on gradient boosting, while (Pasche and Engelke, 2024) use neural networks to model the parameters of the generalized Pareto distribution. More recently, (Gnecco et al., 2024) develop an approach combining POT estimation and generalized random forests (GRF). These authors approximate the conditional distribution $F_{Y|X=x}$ by a conditional generalized Pareto distribution, leading to the quantile

$$Q_x(\tau) = Q_x(\tau_0) + \sigma(x)\frac{\left(\frac{1-\tau}{1-\tau_0}\right)^{-\xi(x)} - 1}{\xi(x)}.$$

The estimation of the conditional quantile therefore relies on estimating the conditional parameters $\sigma(x), \xi(x)$, and $Q_x(\tau_0)$, obtained by weighted maximum likelihood using GRF weights. Simulations show that this approach outperforms classical quantile regression methods as well as standard models from extreme value theory.

However, all these methods rely on the POT approach. In many practical situations, only the block maxima are available, which motivates the use of the Block Maxima approach from extreme value theory. To address this constraint, we propose in this thesis to integrate the BM framework into statistical learning methods, thereby broadening the set of available tools for tackling extreme quantile regression problems. In this direction, we develop new methodologies that combine the generalized extreme value (GEV) distribution with generalized random forests (GRF). The parameters of the GEV distribution, assumed to depend on the covariates, are estimated through a weighted maximum likelihood estimator, where the weights are provided by the GRF procedure. To mitigate overfitting and improve the estimation of the extreme value index, we first introduce an L2-type penalization, followed by a specific penalty on the extreme value index, as proposed by (Coles and Dixon, 1999), in order to stabilize estimation even in small samples. The performance of the proposed methods is assessed through extensive simulation studies, and applications to real datasets are also conducted. The results demonstrate a substantial improvement in robustness and accuracy for estimating extreme quantiles, establishing these methods as credible alternatives to existing approaches. Moreover, we establish the theoretical convergence of the weighted maximum likelihood estimator constructed using GRF weights. The following chapters provide a detailed presentation of the extreme quantile regression methods developed in this thesis.

CHAPTER TROIS

CONSISTENCY OF WEIGHTED MAXIMUM LIKELIHOOD ESTIMATOR FOR EXTREME QUANTILE REGRESSION.

Les résultats présentés dans ce chapitre ont fait l'objet d'un article de recherche qui est encours de soumission:
Vidagbandji *et al.* (2026). Consistency of Weighted maximum likelihood estimator for extreme quantile regression. To be submitted.

Contents

3.1 Introduction

Quantile regression is a statistical method based on the estimation of conditional quantiles. More precisely, given a response variable $Y \in \mathcal{Y} \subset \mathbb{R}$ depending on a covariate $X \in \mathcal{X} \subset \mathbb{R}^p$, the objective of quantile regression is to estimate the conditional quantile of the response variable Y given $X = x$, for any $\tau \in (0,1)$ (Koenker and Bassett, 1978). Although this method provides more comprehensive information on the distribution of Y as a function of the predictor variable X than the conditional mean alone, it faces two main challenges. The first arises when the relationship between the quantile function $Q_\tau(Y \mid X = x)$ and the covariate x is complex, highly nonlinear, and when the dimension p of the predictor space $\mathcal{X}$ is large. The second difficulty arises when we want to estimate a conditional quantile corresponding to an extreme probability level, that is, a quantile of level $\tau_n \to 1$ such that $n(1-\tau_n)$ remains finite when $n \to \infty$, where n denotes the size of the available sample. This situation poses a real challenge, as the estimation requires extrapolation in the tail of the distribution, which is often problematic due to the scarcity of data in this region. In this paper, we propose a method for estimating the conditional quantile that simultaneously addresses these two issues.

It should be noted that the methods for estimating extreme conditional quantiles proposed in the literature can be grouped into four main classes (Wang and Li, 2013). The first consists of directly applying quantile regression to model extreme conditional quantiles, without imposing any assumptions on the conditional distribution $Y \mid X = x$. However, this approach suffers from a lack of reliability for extrapolation in the tail of the distribution, which corresponds to an area of low data density ((Chernozhukov, 2005), (Bremnes, 2004), (Hallock and Koenker, 2001)). The second extends extreme value theory to the regression framework by adopting a local estimation method, based on neighbourhoods of x, for the conditional quantile of Y given $X = x$. The performance of this approach therefore depends heavily on the density of the data around the considered point x ((Gardes and Stupfler, 2015), (Daouia et al., 2013), (Gardes and Stupfler, 2019)). The third combines quantile regression and extreme value theory, making certain assumptions about the behaviour of the tails. These assumptions, which are often restrictive, such as linearity ((Chernozhukov and Fernandez-Val, 2011), (Wang et al., 2012)) or tail equivalence (Wang and Li, 2013), may fail in real-world applications, particularly when the shape of the distribution varies according to the covariates. Finally, the fourth class is based on parametric models, under the assumption that the distribution tail follows either a generalised Pareto distribution (GPD) or a generalised extreme value distribution (GEV), with parameters depending on covariates, either parametrically ((Wang and Li, 2013), (Wang and Tsai, 2009)) or nonparametrically ((Gnecco et al., 2024), (Pasche and Engelke, 2024), (Velthoen et al., 2023), (Farkas et al., 2024)). However, this approach requires delicate choices, such as the selection of thresholds or block sizes. The conditional quantile of order τ is defined by $Q_{Y|X=x}(\tau) = \inf\{y : F_{Y|X=x}(y) \geq \tau\}$, and, when the conditional distribution $F_{Y|X=x}(\cdot)$ is

continuous, this amounts to :

$$Q_{Y|X=x}(\tau) = F^{-1}_{Y|X=x}(\tau), \tag{3.1}$$

for all $x \in \mathbb{R}^p$ and $\tau \in (0,1)$. For simplicity of notation, we will denote in the following $Q_x(\tau)$ to refer to the conditional quantile function $Q_{Y|X=x}(\tau)$. Extreme value theory facilitates extrapolation into the tail of the distribution and is widely used in the literature to address the second challenge of quantile regression mentioned above. For a detailed review, see for example (Beyerlein, 2014a) and (Chernozhukov et al., 2017). Regarding the first challenge, several models based on statistical learning methods have been proposed ((Meinshausen and Ridgeway, 2006), (Athey et al., 2019), (Chaudhuri and Loh, 2002)). More recent works have attempted to address both challenges simultaneously, mainly by combining the *Peak-Over-Threshold* (POT) approach from extreme value theory with different statistical learning methods. Among these are models based on gradient boosting (Velthoen et al., 2023), generalized additive models (Youngman, 2019), neural networks (Pasche and Engelke, 2024), and models using generalized random forests (Gnecco et al., 2024). The method proposed in this work is an extension of these approaches within the block maxima (BM) framework, which is preferred in many fields when modeling extremes, particularly in meteorology (Boudrissa et al., 2017), risk analysis (Calabrese and Giudici, 2015), and hydrology. More specifically, in order to facilitate extrapolation into the tail of the distribution and address the second challenge of quantile regression, we use the generalized extreme value (GEV) distribution, with parameters depending on the covariate x, to model block maxima. These parameters are then estimated via weighted likelihood, with weights obtained from the generalized random forest (grf) method introduced by (Athey et al., 2019). Thus, the distribution $F_{Y|X=x}(\cdot)$ in equation (3.1) is approximated by the GEV distribution, explicitly defined by :

$$G_{(\xi(x),\mu(x),\sigma(x))}(z) = \begin{cases} \exp\left(-(1+\xi(x)\dfrac{z-\mu(x)}{\sigma(x)})_+^{-\frac{1}{\xi(x)}}\right) & if \quad \xi(x) \neq 0, \\ \exp\left(-\exp(-\frac{z-\mu(x)}{\sigma(x)})\right) & if \quad \xi(x) = 0, \end{cases}$$

defined on $\{z \in \mathbb{R} : 1+\xi(x)\dfrac{z-\mu(x)}{\sigma(x)} > 0\}$ for all $x \in \mathbb{R}$ with $a_+ = \max\{0,a\}$. The parameters $\mu \in \mathbb{R}$, $\sigma > 0$ and $\xi \in \mathbb{R}$ are respectively the location, scale and shapes parameters. We estimate $\theta(x) = (\mu(x),\sigma(x),\xi(x))$ by

$$\hat{\theta}_n(x) \in \arg\max_{\theta\in\Theta} L_n(\theta;x)$$

where $L_n(\theta;x)$ is given by

$$L_n(\theta,x) = \sum_{i=1}^{n} w_i(x)\ell_\theta(z_i)$$

with $\ell_\theta(z_i) = \dfrac{dG_{\mu,\sigma,\xi}}{dz_i}(z_i)$, and $w_i(x)$ for all $i = 1,\cdots,n$, representing weights obtained through

the grf method. These weights are used to address the first challenge, namely capturing the complex structure in the data and facilitating estimation when the dimension p of the covariate space $\mathcal{X}$ is large. Based on this estimation, the conditional quantile is obtained by plugging the estimated GEV distribution with parameter $\hat{\theta}_n(x)$ into equation (3.1), thereby yielding the conditional quantile estimator. The main objective of this work is to establish the existence and consistency of the proposed estimator $\hat{\theta}_n(x)$.

The remainder of the paper is organized as follows. Section 3.2 introduces the notation and provides the necessary background on generalized random forests and block maxima approach from extreme value theory, which are essential for establishing the consistency result of the proposed estimator, which will be presented in section 3.3. Finally, the section 3.4 is devoted to the proof of the consistency theorem.

3.2 Proposed method

We begin by recalling some classical results from extreme value theory and the generalized random forest method, within the framework of quantile regression, which are essential for understanding the theoretical results established in this work.

3.2.1 Quantile regression and Generalised random forest

Let $(X_1,Y_1),\ldots,(X_n,Y_n)$ be an $i.i.d.$ sample drawn form the joint distribution of (X,Y), where $X_i \in \mathcal{X} \subset \mathbb{R}^d$ denotes a vector of covariates, and $Y_i \in \mathcal{Y} \subset \mathbb{R}$ the reponse variable. In quantile regression, the objectives is to estimate the conditional quantile of probability level $\tau_n \in (0,1)$, defined as

$$Q_x(\tau_n) = \inf\{y : F_{Y|X=x}(y) \geq \tau_n\} \tag{3.2}$$

where $F_{Y|X=x}(.)$ denotes the conditional distribution function of Y given $X = x$. Another characterisation of the conditional quantile is based on the following optimisation problem :

$$Q_x(\tau_n) = \arg\min_q \mathbb{E}\left[\rho_{\tau_n}(Y,q) \mid X = x\right],$$

where $\rho_{\tau_n}(.,.)$ is the asymmetric loss function defined as

$$\rho_{\tau_n}(Y,q) = \begin{cases} \tau_n(Y-q), & Y-q \geq 0 \\ (1-\tau_n)(Y-q), & Y-q < 0. \end{cases} \tag{3.3}$$

A variety of conditional quantile estimators have been proposed based on this formulation, under both parametric and nonparametric assumptions on the quantile function ((Chernozhukov, 2005), (Chernozhukov et al., 2017), (Buhai, 2005), (Wang et al., 2012), (Wang and Li, 2013),

(Fakoor et al., 2023)). Based on this characterization and without parametric assumptions on the conditional distribution, a natural weighted approximation of $Q_\tau(x)$ is

$$\hat{Q}_x(\tau_n) = \arg\min_q \sum_{i=1}^{n} w_i(x)\rho_{\tau_n}(Y_i, q),$$

where $w_i(x)$ is a localized weight function reflecting the influence of the observation i in a neighborhood of x.

Using the characterization in equation (3.1), (Meinshausen and Ridgeway, 2006) proposed a method based on the random forests introduced by (Breiman, 2001) to construct an empirical estimator of the conditional distribution function : $\hat{F}(y \mid X = x) = \sum_{i=1}^{n} w_i(x)\mathbb{1}_{\{Y_i \leq y\}}$. This estimated distribution function is then plugged into equation (3.1) to obtain an estimate of the conditional quantile. The weights $w_i(x)$ are obtained via the random forest method, which is based on conditional mean estimation ; hence $w_i(x)$ tends to be large for $i \in \{1, \cdots, n\}$ whenever $\mathbb{E}(Y|X = x) \approx \mathbb{E}(Y|X = X_i)$ ((Lin and Jeon, 2006), (Meinshausen and Ridgeway, 2006)). However, the situation differs when estimating conditional quantiles. As shown by (Athey et al., 2019) and (Gnecco et al., 2024), the weight $w_i(x)$ may not be large for $i \in \{1, \cdots, n\}$, even if $Q_x(\tau) \approx Q_{X_i}(\tau)$. Consequently, the classical random forest method of (Breiman, 2001) fails to capture the complex structure of the quantile function. The generalized random forest, introduced by (Athey et al., 2019) for quantile regression problems, overcomes this limitation and provides a more flexible and targeted estimation of heterogeneous effects, particularly for conditional quantiles.

The random forest method relies on aggregating the predictions of B decision trees to produce an estimator that is more stable, robust, and accurate than any single tree considered in isolation (Breiman, 2001). For an observation $x \in \mathbb{R}^p$, the prediction $\eta_b(x)$ from the b^{th} tree in the regression context is defined as the average of the responses Y_i corresponding to observations X_i that fall in the same terminal region $R_b(x)$ as x :

$$\eta_b(x) = \sum_{i=1}^{n} \frac{\mathbb{1}_{\{X_i \in R_b(x)\}} Y_i}{|\{i : X_i \in R_b(x)\}|}, \quad b = 1, \ldots, B,$$

where $R_b(x) \subset \mathbb{R}^p$ denotes the leaf or terminal region containing x in the tree b, and $|E|$ represents the cardinality of the set E. The final random forest estimator is obtained by averaging the predictions across all B trees :

$$\eta(x) = \sum_{i=1}^{n} w_i(x) Y_i,$$

where the similarity weights $w_i(x)$ are given by

$$w_i(x) = \frac{1}{B}\sum_{b=1}^{B} \frac{\mathbb{1}_{\{X_i \in R_b(x)\}}}{|\{i : X_i \in R_b(x)\}|}. \tag{3.4}$$

These weights $w_i(x)$ reflect the proximity between the observation x and each observation X_i, as induced by the structure of the trees in the forest. The generalized random forest approach extends the classical method by constructing trees based on a customizable loss function suited to the learning task. In the case of quantile estimation, the loss function is that given in equation (3.3). The grf similarity weights $w_i(x)$ retain the same structure as in classical random forests, with the key difference being the splitting criterion used to partition the data. We employ this learning method to capture the complex structure of the quantile function and to facilitate estimation when the dimension of the covariate space is high. The specific properties of this method, which are essential for studying the consistency of the proposed estimator, will be recalled in Section 3.3.

3.2.2 Extreme quantile regression based on BM approach

The block maxima method is a widely used approach for modeling extreme values across various fields. In this work, it will be employed to facilitate extrapolation in the tail of the distribution, thereby allowing the estimation of quantiles at probability levels $\tau_n \to 1$ such that $n(1-\tau_n)$ remains finite as $n \to +\infty$. This approach addresses the challenge of estimating conditional quantiles at extreme probability levels. For a more in-depth understanding of the extreme value theory (EVT) results presented in this section, we refer to (De Haan and Ferreira, 2006) and (Coles, 2001). We first recall some classical results in the unconditional case, before showing how this approach can be combined with the generalized random forest method to simultaneously tackle the issues identified in quantile regression.

The BM method is based on the limiting theorem describing the asymptotic behavior of the maximum of a sequence of independent and identically distributed random variables $Y_1, \ldots, Y_m$ with an unknown probability distribution F. This fundamental result originates from the pioneering work of (Fisher and Tippett, 1928) and (Gnedenko, 1943), who established conditions such that there exist normalization sequences $a_m > 0$ and $b_m \in \mathbb{R}$ such that

$$\lim_{m\to+\infty} F^m(a_m y + b_m = G_\xi(y), \quad \text{for all } y \in \mathbb{R}, \tag{3.5}$$

where G_ξ is a non-degenerate probability distribution defined by

$$G_\xi(y) = \exp\left(-(1+\xi y)^{-1/\xi}\right), \quad \text{for } 1+\xi y > 0.$$

This law is known as the *Generalized Extreme Value* (GEV) distribution. A distribution func-

tion F that satisfies relation (3.5) is said to belong to the maximum domain of attraction of G_ξ (De Haan and Ferreira, 2006), which is denoted by $F \in \mathcal{D}(G_\xi)$.

To illustrate the block maxima approach, consider a sequence of independent and identically distributed random variables $Y_1, \ldots, Y_N$, following a probability distribution F that belongs to the maximum domain of attraction $\mathcal{D}(G_{\xi_0})$. The BM method consists in partitioning these data into n disjoint blocks of approximately equal size $m > 1$, defined as:

$$B_{k,m} = \{Y_{(k-1)m+1}, \ldots, Y_{km}\}, \quad k = 1, \ldots, n.$$

For each block, we consider the maximum statistic: $Z_k = \max_{i \in B_{k,m}} Y_i$. Since the Y_i are i.i.d., the distribution of the block maximum Z_k of size m is given by F^m. According to the classical limit theorem for maxima, this distribution converges, after normalization, to the GEV distribution with parameters (a_m, b_m, ξ_0). The BM method assumes that the block maxima $Z_1, \ldots, Z_n$ follow a GEV distribution and are independent. The general form of the GEV distribution is given by:

$$G_{\mu,\sigma,\xi}(z) = \begin{cases} \exp\left(-\left(1+\xi\dfrac{z-\mu}{\sigma}\right)_+^{-1/\xi}\right), & \xi \neq 0, \\ \exp\left(-\exp\left(-\dfrac{z-\mu}{\sigma}\right)\right), & \xi = 0, \end{cases} \tag{3.6}$$

defined for $1+\xi\dfrac{z-\mu}{\sigma} > 0$, with $\mu \in \mathbb{R}$, $\sigma > 0$ and $\xi \in \mathbb{R}$. The following section outlines how this approach is applied in the context of extreme quantile regression.

3.3 Main results

We combine the BM approach, which facilitates extrapolation in the tail of the distribution, with the generalized random forest method, which captures complex structures in the data and improves estimation when the covariate space is high-dimensional. To illustrate the proposed method, consider a sequence of i.i.d. random vectors $(X_1, Y_1), \ldots, (X_N, Y_N)$ distributed as the random vector (X, Y), where $X \in \mathcal{X} \subset \mathbb{R}^p$ and $Y \in \mathcal{Y} \subset \mathbb{R}$. We assume that the sample size is $N = n \times m$, and that we form n blocks of size m, so that the k^{th} block, for $k = 1, \ldots, n$, is given by

$$B_{k,m} = \{(X_{(k-1)m+1}, Y_{(k-1)m+1}), \cdots, (X_{km}, Y_{km})\}$$

Let us define

$$Z_{k,m} = \max\{Y_i : (X_i, Y_i) \in B_{k,m}\} \tag{3.7}$$

and $X_{k,m}$ denote the covariate X_i corresponding to the maximizing Y_i in $\{Y_i : (X_i, Y_i) \in B_{k,m}\}$. Thus, $(X_{k,m}, Z_{k,m})$ for $k = 1, \ldots, n$ forms the block maxima sample with respect to the response variable Y. For all $x \in \mathcal{X}$, let $F_x(\cdot)$ denote the conditional distribution of Y given $X = x$.

Assumption 3.1 (Conditional max-domain of attraction)**.** *For every $x \in \mathcal{X}$, the conditional distribution function $F_x(.)$ belongs to the max-domain of attraction of an extreme value distribution $G_{\xi_0(x)}$, that is, $F_x \in \mathcal{D}(G_{\xi_0(x)})$. Equivalently, there exist normalizing functions $a_m(x) > 0$ and $b_m(x) \in \mathbb{R}$ such that, for all $z \in \mathbb{R}$,*

$$\lim_{m\to+\infty} F_x^m\big(a_m(x)z + b_m(x)\big) = G_{\xi_0(x)}(z), \tag{3.8}$$

where

$$G_{\xi_0(x)}(z) = \exp\left(-\big(1+\xi_0(x)z\big)^{-1/\xi_0(x)}\right), \qquad 1+\xi_0(x)z > 0.$$

In the case $\xi_0(x) = 0$, the distribution $G_{\xi_0(x)}$ is understood as the continuous limit of $G_{\xi_0}(z)$ as $\xi_0 \to 0$.

This assumption is the conditional version of the convergence in (3.5) and implies that the normalized block maxima converge in distribution to $G_{\xi_0(x)}$ as $m \to +\infty$.

Remark 3.1. *Suppose $F_x \in \mathcal{D}(G_{\xi_0(x)})$ and denote the corresponding normalization sequences $a_m(x)$ and $b_m(x)$. The block maximum $Z_{k,m}$ given $X_{k,m} = x$ has distribution function $F_x^m(z)$ which we shall denote by $F_m(z|x)$ for notational simplicity. In others words, for any $m \geq 1$, we have*

$$F_m(z \mid x) := F_x^m\big(z\big).$$

Note that $F_m(\cdot \mid X_{k,m})$ is the true conditional distribution function of $Z_{k,m}$ given $X_{k,m}$. Under Assumption 3.1, we have the convergence

$$\lim_{m\to+\infty} F_m(a_m(x)z + b_m(x) \mid x) = G_{\xi_0(x)}(z), \qquad z \in \mathbb{R}. \tag{3.9}$$

We model the conditional distribution of $Z_{k,m}$ given $X_{k,m} = x$ by a generalize extreme values distribution with parameters $(\mu(x), \sigma(x), \xi(x))$, when $m \to \infty$. In the conditional setting, the parameters μ, σ, and ξ of the GEV distribution depend on x and are described by the functions $\mu : \mathcal{X} \to \mathbb{R}$, $\sigma : \mathcal{X} \to \mathbb{R}_+^*$, and $\xi : \mathcal{X} \to \mathbb{R}$. To address the extrapolation problem in the tail, we approximate the conditional distribution in (3.1) by this conditional GEV distribution. We then define the estimator $\hat{\theta}_n(x) = \left(\hat{\mu}_n(x), \hat{\sigma}_n(x), \hat{\xi}_n(x)\right)$ of the conditional GEV parameters as the value of θ that maximizes the local weighted log-likelihood function defined by

$$L_n(\theta; x) = \sum_{k=1}^{n} w_k(x)\ell_\theta(z_{k,m}), \quad x \in \mathcal{X}, \tag{3.10}$$

where $\ell_\theta(z)$ is equal to :

$$\ell_\theta(z) = \begin{cases} -\log(\sigma) - (1+\frac{1}{\xi})\log\left(1+\xi\frac{z-\mu}{\sigma}\right) - \left(1+\xi\frac{z-\mu}{\sigma}\right)^{-\frac{1}{\xi}} & \text{if} \quad 1+\xi\frac{z-\mu}{\sigma} > 0, \\ -\infty & \text{otherwise,} \end{cases}$$

when $\xi \neq 0$, and

$$\ell_\theta(z) = -\log(\sigma) - \left(\frac{z-\mu}{\sigma}\right) - \exp\left(-\frac{z-\mu}{\sigma}\right)$$

when $\xi = 0$. The weights $w_k(x)$ are obtained using the generalized random forest method briefly described in Section 3.2.1. The weights $w_k(x)$ depend on the sample size n, since they are derived from the generalized random forest built from the entire dataset. The conditional quantile of probability level τ_n, with τ_n close to 1, is then obtained by substituting the estimated GEV distribution into equation (3.1). It is given, for all $x \in \mathcal{X}$, by

$$\hat{Q}_\tau(x) = \begin{cases} \hat{\mu}(x) + \dfrac{\hat{\sigma}(x)}{\hat{\xi}(x)}\left(\left(-\ln(\tau^m)\right)^{-\hat{\xi}(x)} - 1\right) & \text{if} \quad \hat{\xi}(x) \neq 0, \\ \hat{\mu}(x) - \hat{\sigma}(x)\ln\left(-\ln\left(\tau^m\right)\right) & \text{if} \quad \hat{\xi}(x) = 0. \end{cases} \tag{3.11}$$

The support of the GEV distribution is given by $\{z \in \mathbb{R} : 1 + \xi\frac{z-\mu}{\sigma} > 0\}$, if $\xi \neq 0$, and by $\mathbb{R}$ if $\xi = 0$. This support depends on the parameters of the GEV distribution, which are unknown in practice. This makes it difficult to establish the usual regularity conditions required to guarantee the classical asymptotic properties of the maximum likelihood estimator (MLE). It should be noted that the weighted likelihood defined in equation (3.10) does not admit a global solution. Indeed, (Smith, 1985) showed that the MLE does not exist if $\xi \leq -1$, that it is asymptotically normal when $\xi > -0.5$, and that it nevertheless remains consistent for $\xi > -1$. Along the same lines, (Dombry, 2015) proved the existence and consistency of the MLE of the GEV parameters for all $\xi > -1$, under a first-order extreme value condition. Recently, (Dombry and Ferreira, 2019) provided a formal proof of the asymptotic normality of the GEV MLE. We say that $\hat{\theta}(x)$ is a local weighted maximum likelihood estimator if it maximizes (3.10) over Θ, where $\Theta \subset \mathbb{R} \times (0, +\infty) \times (-1, +\infty)$. This local estimator is defined by

$$\hat{\theta}_n(x) \in \arg\max_{\theta \in \Theta} L_n(\theta; x), \quad \forall x \in \mathcal{X}. \tag{3.12}$$

In order to establish the consistency of this estimator, we introduce a set of assumptions on the weights $w_k(\cdot)$ that ensure the convergence of $\hat{\theta}_n(x)$. Assumptions 3.2 and 3.3 gather the regularity conditions on the forest weights required in our proof. These assumptions are adapted from (Gnecco et al., 2024) and (Athey et al., 2019).

Assumption 3.2. *Let $b = 1, \ldots, B$ denote a tree in the forest, and let $x \in \mathcal{X}$ be a fixed predictor point. Define the diameter of the leaf $R_b(x)$ by*

$$\operatorname{diam}(R_b(x)) := \sup\{\|y - x\|_2 : y \in R_b(x)\}.$$

Let $s < n$ denote the number of observations used to grow the tree. We assume that the diameter

of the leaf $R_b(x)$ converges in probability to zero, that is, for every $\varepsilon > 0$,

$$\mathbb{P}[\text{diam}(R_b(x)) > \varepsilon] \to 0 \quad \textit{as } s \to \infty.$$

Assumption 3.3. *The weights satisfy $\sum_{k=1}^{n} w_k(x) = 1$ and, for all $k = 1, \ldots, n$, we have*

$$\max_{1 \le k \le n} w_k(x) \le \frac{s}{n}.$$

where s denotes the subsample size used in the construction of the trees, with $s \to \infty$ and $s/n \to 0$ as $n \to \infty$. We assume that the forest is honest, following the framework of (Wager and Athey, 2018) and (Athey et al., 2019). Thus, for each k, the weight $w_k(x)$ is independent of Z_k, m conditional on $X_{k,m} = x$.

Finally, we introduce assumptions regarding the distribution of the covariate $X_{k,m}$ concomitant to the maximum, as well as the conditional distribution $F_m(\cdot \mid X = x)$. These conditions are essential for the proof of Lemma 3.2 and are analogous to those considered by (Gnecco et al., 2024) and (Meinshausen and Ridgeway, 2006). Note that $X_{k,m}$ is the covariate associated with the block maximum, which implies that its distribution might differ from the marginal distribution of X. We assume in assumption below that it still admits a well-behaved density.

Assumption 3.4. *Assume that the predictor space $\mathcal{X} \subset \mathbb{R}^p$ is compact and that the distribution of the covariate $X_{k,m}$ associated with the maximum admits a density on $\mathcal{X}$ that is uniformly bounded away from zero and infinity in m. Moreover, assume that there exists a constant $L > 0$ such that $F_m(y \mid X = x)$ is Lipschitz continuous in x with Lipschitz constant L, that is, for all $x, x' \in \mathcal{X}$,*

$$\sup_{y \in \mathbb{R}} \left| F_m(y \mid X = x) - F_m(y \mid X = x') \right| \le L \|x - x'\|_2.$$

We emphasize that the Lipschitz continuity assumption on $F_m(\cdot \mid x)$ holds uniformly in m for m sufficiently large. Assume also that, for every $x \in \mathcal{X}$, the conditional distribution function $F_m(y \mid X = x)$ is strictly monotonically increasing in y.

In the BM approach, the block size m must be chosen carefully : if m is too large, the variance of the estimator increases ; if it is too small, the bias becomes more significant. Therefore, a bias–variance trade-off must be considered when selecting m. (Dombry, 2015) proved consistency under the condition $\frac{m}{\log(n)} \to +\infty$, or equivalently $\frac{n \log(n)}{N} \to 0$. The presence of the logarithmic factor implies that, in practice, the number of observations effectively used in the block maxima approach is of a smaller order than that required by the POT method, which only requires $\frac{n}{N} \to 0$, where n denotes in this setting the number of threshold exceedances ((Zhou, 2009), (Gnecco et al., 2024)). This condition is crucial for establishing the consistency of the estimator proposed in this work, which is stated below. Furthermore, we assume that the block size m depends on the sample size n of the block maxima, that is, $m = m(n)$.

Theorem 3.1 (Existence and Consistency)**.** *Let $x \in \mathcal{X}$. Suppose $F_x \in \mathcal{D}(G_{\xi_0(x)})$ and assume that Assumptions 3.1 to 3.4 hold with $\xi_0(x) > -1$ and $\lim_{n\to+\infty} \frac{m(n)}{\log(n)} = +\infty$. Then, there exists a sequence of estimators $(\hat{\mu}_n(x), \hat{\sigma}_n(x), \hat{\xi}_n(x))$ and an integer $N \geq 1$ such that*

$$\mathbb{P}\left[(\hat{\mu}_n(x), \hat{\sigma}_n(x), \hat{\xi}_n(x)) \text{ is an local MLE for all } n \geq N\right] = 1, \tag{3.13}$$

and

$$\hat{\xi}_n(x) \xrightarrow{\mathbb{P}} \xi_0(x), \qquad \frac{\hat{\mu}_n(x) - b_m(x)}{a_m(x)} \xrightarrow{\mathbb{P}} 0, \qquad \frac{\hat{\sigma}_n(x)}{a_m(x)} \xrightarrow{\mathbb{P}} 1, \quad \text{as } n \to +\infty. \tag{3.14}$$

The condition $\lim_{n\to+\infty} \frac{m(n)}{\log(n)} = +\infty$ comes from the work of (Dombry, 2015) and is required in the proof of Lemma 3.5. The assumption $\xi_0(x) > -1$ is crucial and aligns with the results of (Dombry, 2015) in the unconditional case, as it ensures the first-order condition. That is, the maximum likelihood estimator must satisfy

$$\left(\frac{\partial Ln}{\partial \mu}(\theta; x), \frac{\partial Ln}{\partial \sigma}(\theta; x), \frac{\partial Ln}{\partial \xi}(\theta; x)\right) = (0, 0, 0).$$

Indeed, we have

$$\begin{aligned}
\frac{\partial L_n}{\partial \mu}(\theta; x) &= \sum_{k=1}^{n} w_k(x) \frac{\partial \ell_\theta}{\partial \mu}(Z_{k,m}) \\
&= -\frac{1}{\sigma} \sum_{k=1}^{n} w_k(x) \left[-\left(\frac{1+\xi}{1+\xi \frac{Z_{k,m}-\mu}{\sigma}} \right) + \left(1 + \xi \frac{Z_{k,m}-\mu}{\sigma} \right)^{-1-\frac{1}{\xi}} \right] \\
&= -\frac{1}{\sigma} \sum_{k=1}^{n} w_k(x)\, g'_\xi\left(\frac{Z_{k,m}-\mu}{\sigma} \right),
\end{aligned}$$

with

$$g_\xi(y) = -\left(1 + \frac{1}{\xi}\right) \log(1+\xi y) - (1+\xi y)^{-\frac{1}{\xi}}.$$

According to Proposition 1 of (Dombry, 2015), when $\xi \leq -1$, the function g_ξ is strictly increasing on its domain of definition. Consequently, for any $x \in \mathcal{X}$, we have

$$\frac{\partial Ln}{\partial \mu}(\theta, x) < 0 \quad \text{whenever } \xi \leq -1.$$

It follows that the condition $\xi > -1$ is necessary for the existence of a local extremum.

3.4 Proofs

The proof of Theorem 3.1 builds on the work of (Dombry, 2015). We first introduce some notation and recall a few preliminary results that will be used in the proof.

Notation 3.1. *Let $x \in \mathcal{X}$. We normalize the block maxima as follows*

$$\tilde{Z}_{k,m} = \frac{Z_{k,m} - b_m(x)}{a_m(x)}, \quad \textit{where } Z_{k,m} \textit{ is defined in equation (3.7).}$$

The associated weighted log-likelihood is given by

$$\tilde{L}_n(\theta; x) = \sum_{k=1}^{n} w_k(x)\, \ell_\theta(\tilde{Z}_{k,m}), \tag{3.15}$$

where $w_k(x)$, for all $x \in \mathcal{X}$, denotes the weight assigned to each X_k, with $k = 1, \ldots, n$, obtained using the GRF method, whose explicit form is provided in relation (3.4). We therefore introduce $\Theta = \mathbb{R} \times (0, +\infty) \times (-1, +\infty)$, where a generic element is denoted by $\theta = (\mu, \sigma, \xi)$. The restriction of $\tilde{L}_n(\cdot; x)$ to Θ, that is, $\tilde{L}_n(\cdot; x) : \Theta \to [-\infty, +\infty)$, is a continuous function. Hence, for any compact set $K \subset \Theta$, this restriction is bounded and reaches its maximum on K. Denoting this maximum by $\tilde{\theta}_n^K(x) = \left(\tilde{\mu}_n^K(x), \tilde{\sigma}_n^K(x), \tilde{\xi}_n^K(x)\right)$, it is defined for all $x \in \mathcal{X}$ as follows

$$\tilde{\theta}_n^K(x) = \arg\max_{\theta \in K} \tilde{L}_n(\theta; x). \tag{3.16}$$

For $x \in \mathcal{X}$, let us denote by $\mathbb{P}_n^x$ the empirical distribution defined as

$$\mathbb{P}_n^x = \sum_{k=1}^{n} w_k(x)\, \delta_{\tilde{Z}_{k,m}},$$

where δ_x represents the Dirac measure at the point $x \in \mathbb{R}$. The measure P_n^x is a random probability measure conditional on $(X_1, \ldots, X_n)$, since the weights satisfy $\sum_k w_k(x) = 1$. For any measurable function $f : \mathbb{R} \to [-\infty, +\infty)$, we define

$$\mathbb{P}_n^x[f] = \sum_{k=1}^{n} w_k(x)\, f(\tilde{Z}_{k,m}).$$

With these notations, the weighted log-likelihood defined in (3.15) can be written as

$$\tilde{L}_n(\theta; x) = \mathbb{P}_n^x[\ell_\theta],$$

and the associated empirical distribution function is given by

$$\mathbb{F}_n^x(t) = \mathbb{P}_n^x((-\infty,t]) = \sum_{k=1}^{n} w_k(x)\,\mathbb{1}_{\{\tilde{Z}_{k,m}\leq t\}}, \quad \textit{for all } t\in\mathbb{R}.$$

We now present some essential results that will be used to establish the proof of Theorem 3.1.

Lemma 3.1. *Let $x\in\mathcal{X}$. The triplet $(\hat{\mu}_n(x),\hat{\sigma}_n(x),\hat{\xi}_n(x))$ is a weighted maximum likelihood estimator if and only if $\tilde{L}_n$ has a local maximum at $\left(\frac{\hat{\mu}_n(x)-b_m(x)}{a_m(x)},\frac{\hat{\sigma}_n(x)}{a_m},\hat{\xi}_n(x)\right)$.*

Proof. Let $x\in\mathcal{X}$. According to Lemma 1 in (Dombry, 2015), we have

$$\ell_{(\mu,\sigma,\xi)}\Big(\frac{z-b}{a}\Big) = \ell_{(a\mu+b,a\sigma,\xi)}(z)+\log(a), \quad \text{for all } z\in\mathbb{R}.$$

Hence,

$$\begin{aligned}
\tilde{L}_n\Big(\Big(\frac{\mu-b_m}{a_m},\frac{\sigma}{a_m},\xi\Big);x\Big) &= \sum_{k=1}^{n} w_k(x)\,\ell_{(\frac{\mu-b_m}{a_m},\frac{\sigma}{a_m},\xi)}\Big(\frac{Z_{k,m}-b_m}{a_m}\Big)\\
&= \sum_{k=1}^{n} w_k(x)\,\ell_{(\mu,\sigma,\xi)}(Z_{k,m})+\log(a_m), \quad \text{since } \sum_{k=1}^{n} w_k(x)=1\\
&= L_n(\theta;x)+\log(a_m).
\end{aligned}$$

It follows that

$$L_n((\mu,\sigma,\xi);x) = \tilde{L}_n\Big(\Big(\frac{\mu-b_m}{a_m},\frac{\sigma}{a_m},\xi\Big);x\Big)-\log(a_m).$$

Therefore, the local maxima of L_n and $\tilde{L}_n$ correspond directly, which proves Lemma 3.1. □

Lemma 3.2. *Let $x\in\mathcal{X}$. Suppose that assumptions 3.1 to 3.4 hold and $\lim_{n\to+\infty} m(n)=+\infty$. Then,*

$$\sup_{t\in\mathbb{R}}|\mathbb{F}_n^x(t)-G_{\xi_0(x)}(t)| \xrightarrow{\mathbb{P}} 0.$$

Proof. Let $x\in\mathcal{X}$. Let $F_x\in\mathcal{D}(G_{\xi_0(x)})$, and let $a_m(x)$ and $b_m(x)$ be the corresponding normalizing sequences. We write $F_m(\cdot\mid x)$ for the conditional distribution function of the maximum $Z_{k,m}$ given $X=x$, as introduced in Remark 3.1. For all $t\in\mathbb{R}$, we have

$$\begin{aligned}
\big|\mathbb{F}_n^x(t)-G_{\xi_0(x)}(t)\big| &= \left|\sum_{k=1}^{n} w_k(x)\mathbb{1}_{\{\tilde{Z}_{k,m}\leq t\}}-F_m(a_m t+b_m|x)+F_m(a_m(x)t+b_m(x|x)-G_{\xi_0(x)}(t)\right|\\
&\leq |A_{n,m}(t)|+|B_{n,m}(t)|,
\end{aligned}$$

with

$$A_{n,m}(t)=\sum_{k=1}^{n} w_k(x)\mathbb{1}_{\{\tilde{Z}_{k,m}\leq t\}}-F_m(a_m(x)t+b_m(x)|x) \quad \text{and} \quad B_{n,m}(t)=F_m(a_m t+b_m|x)-G_{\xi_0(x)}(t).$$

Under Assumption 3.1, and in particular from equation (3.9), we obtain that for every $x \in \mathcal{X}$ and $m = m(n) \to \infty$,

$$\sup_{t\in\mathbb{R}} |B_{n,m}(t)| \longrightarrow 0.$$

It therefore remains to prove that $\sup_{t\in\mathbb{R}} |A_{n,m}(t)| \xrightarrow{p} 0$. The proof of this convergence follows the same strategy as that of Theorem 1 in (Meinshausen and Ridgeway, 2006), which establishes the uniform convergence in probability of a weighted conditional empirical distribution function to the true conditional distribution in a non-extreme setting, where the weights come from the classic random forest method. To this end, define the random variables $U_{k,m}$, $k = 1,\ldots,n$, as the conditional probability integral transforms of the observations $Z_{k,m}$ (the non-normalized block maxima) given $X_{k,m}$:

$$U_{k,m} = F_m(Z_{k,m} \mid X_{k,m}),$$

where $F_m(\cdot \mid X_{k,m})$ denotes the true conditional distribution function of $Z_{k,m}$ given $X_{k,m}$. By construction, conditionally on $X_{k,m}$, the variables $U_{k,m}$ are i.i.d. uniformly distributed on $[0,1]$, and they are independent across k. Now, for a fixed $t \in \mathbb{R}$, set $y = a_m(x)t + b_m(x)$. Then, we have

$$\{\tilde{Z}_{k,m} \le t\} = \{Z_{k,m} \le a_m(x)t + b_m(x)\} = \{U_{k,m} \le F_m(y \mid X_{k,m})\}.$$

Therefore, the weighted empirical distribution function can be written as

$$\begin{aligned}
\mathbb{F}_n^x(t) &= \sum_{k=1}^{n} w_k(x) \mathbb{1}_{\{\tilde{Z}_{k,m} \le t\}} \\
&= \sum_{k=1}^{n} w_k(x) \mathbb{1}_{\{U_{k,m} \le F_m(y|X_{k,m})\}}.
\end{aligned}$$

Hence,

$$\begin{aligned}
|A_{n,m}(t)| &= \left| \sum_{k=1}^{n} w_k(x) \left(\mathbb{1}_{\{U_{k,m} \le F_m(y|X_{k,m})\}} - \mathbb{1}_{\{U_{k,m} \le F_m(y|x)\}} \right) \right. \\
&\qquad \left. + \sum_{k=1}^{n} w_k(x) \mathbb{1}_{\{U_{k,m} \le F_m(y|x)\}} - F_m(y \mid x) \right| \\
&\le \left| \sum_{k=1}^{n} w_k(x) \left(\mathbb{1}_{\{U_{k,m} \le F_m(y|X_{k,m})\}} - \mathbb{1}_{\{U_{k,m} \le F_m(y|x)\}} \right) \right| \\
&\qquad + \left| F_m(y \mid x) - \sum_{k=1}^{n} w_k(x) \mathbb{1}_{\{U_{k,m} \le F_m(y|x)\}} \right|.
\end{aligned}$$

Taking the supremum over $t \in \mathbb{R}$ in the second term of the above inequality, we obtain

$$\sup_{y\in\mathbb{R}} \left| F_m(y \mid x) - \sum_{k=1}^{n} w_k(x) \mathbf{1}_{\{U_{k,m} \le F_m(y|x)\}} \right| = \sup_{z\in[0,1]} \left| z - \sum_{k=1}^{n} w_k(x) \mathbf{1}_{\{U_{k,m} \le z\}} \right|,$$

since $F_m(\cdot \mid x)$ is strictly increasing (Assumption 3.4) and maps $\mathbb{R}$ onto $(0,1)$. This term corresponds to the variance-type term studied in (Meinshausen and Ridgeway, 2006). The authors showed, after applying Bonferroni's inequality, that this term $\sup_{z\in[0,1]} \left| z - \sum_{k=1}^{n} w_k(x) \mathbf{1}_{\{U_{k,m} \le z\}} \right|$ converges to zero in probability provided that,

$$\sum_{k=1}^{n} w_k^2(x) \longrightarrow 0 \qquad \text{as } n \to \infty.$$

This condition is satisfied by the generalized random forest weights $w_k(x)$ used in our setting. Indeed, by Assumption 3.3, we have

$$\max_{1\le k\le n} w_k(x) \le \frac{s}{n}, \qquad \text{with } \frac{s}{n} \to 0 \text{ as } n \to \infty.$$

Since $\sum_{k=1}^{n} w_k(x) = 1$, it follows that

$$\sum_{k=1}^{n} w_k^2(x) \le \max_{1\le k\le n} w_k(x) \longrightarrow 0, \qquad n \to \infty. \tag{3.17}$$

Consequently, for every $x \in \mathcal{X}$,

$$\sup_{z\in[0,1]} \left| z - \sum_{k=1}^{n} w_k(x) \mathbf{1}_{\{U_{k,m} \le z\}} \right| \xrightarrow{p} 0.$$

We now turn to the first term in the upper bound of $|A_{n,m}(t)|$. Recall that the conditional distribution function $F_m(\cdot \mid x)$ satisfies the same regularity assumptions imposed in (Meinshausen and Ridgeway, 2006) on the conditional distribution $F(\cdot \mid X = x)$. As pointed out by the author, it therefore suffices to show that

$$\left| \sum_{k=1}^{n} w_k(x) \left(\mathbb{1}_{\{U_{k,m} \le F_m(y|X_{k,m})\}} - \mathbb{1}_{\{U_{k,m} \le F_m(y|x)\}} \right) \right| \xrightarrow{p} 0,$$

in order to conclude that

$$\sup_{t\in\mathbb{R}} |A_{n,m}(t)| \xrightarrow{p} 0.$$

This is the convergence that remains to be established. To this end, for each $k \in \{1, \ldots, n\}$ and

fixed y such that $y = a_m(x)t + b_m(x)$, define

$$D_{k,m} = \mathbb{1}_{\{U_{k,m} \leq F_m(y|X_{k,m})\}} - \mathbf{1}_{\{U_{k,m} \leq F_m(y|x)\}}.$$

Since $U_{k,m}$, $k = 1, \ldots, n$, are uniformly distributed on $[0,1]$, it follows that

$$\mathbb{E}\Big(D_{k,m}\Big) = F_m(y \mid X_{k,m}) - F_m(y \mid x).$$

Using (3.17) together with the independence of the $U_{k,m}$, $k = 1, \ldots, n$, as $n \to \infty$, we obtain, based on a strategy similar to that of (Meinshausen and Ridgeway, 2006),

$$\left| \sum_{k=1}^{n} w_k(x) \Big(\mathbf{1}_{\{U_{k,m} \leq F_m(y|X_{k,m})\}} - \mathbf{1}_{\{U_{k,m} \leq F_m(y|x)\}} \Big) \right| \xrightarrow{\mathbb{P}} \left| \sum_{k=1}^{n} w_k(x) \big(F_m(y \mid X_{k,m}) - F_m(y \mid x) \big) \right|.$$

We now control the deterministic term. By the triangle inequality and by Assumption 3.4 (mainly the Lipschitz property), we have

$$\begin{aligned} \left| \sum_{k=1}^{n} w_k(x) \big(F_m(y \mid X_{k,m}) - F_m(y \mid x) \big) \right| &\leq \sum_{k=1}^{n} w_k(x) \left| F_m(y \mid X_{k,m}) - F_m(y \mid x) \right| \\ &\leq L \sum_{k=1}^{n} w_k(x) \left\| X_{k,m} - x \right\|_2. \end{aligned}$$

By Lemma 3.3, we have

$$\sum_{k=1}^{n} w_k(x) \left\| X_{k,m} - x \right\|_2 \xrightarrow{\mathbb{P}} 0.$$

Hence, for every fixed $y \in \mathbb{R}$,

$$\left| \sum_{k=1}^{n} w_k(x) \big(F_m(y \mid X_{k,m}) - F_m(y \mid x) \big) \right| \xrightarrow{\mathbb{P}} 0.$$

Combining the previous results, we conclude that

$$\left| \sum_{k=1}^{n} w_k(x) \Big(\mathbf{1}_{\{U_{k,m} \leq F_m(y|X_{k,m})\}} - \mathbf{1}_{\{U_{k,m} \leq F_m(y|x)\}} \Big) \right| \xrightarrow{\mathbb{P}} 0.$$

Therefore,

$$\sup_{t \in \mathbb{R}} |A_{n,m}(t)| \xrightarrow{\mathbb{P}} 0.$$

Finally, combining this result with the convergence of the bias term $B_{n,m}(t)$, we obtain that for

every $x \in \mathcal{X}$ and $m = m(n) \to \infty$,

$$\sup_{t\in\mathbb{R}} \left| \mathbb{F}_n^x(t) - G_{\xi_0(x)}(t) \right| \xrightarrow{\mathbb{P}} 0.$$

□

This lemma 3.2 allows us to conclude, by Lemma 2.2 in (Van Der Vaart, 1998), that for any bounded continuous function $f : \mathbb{R} \to \mathbb{R}$, we have

$$\mathbb{P}_n^x[f] \xrightarrow{\mathbb{P}} G_{\xi_0(x)}[f] := \int_{\mathbb{R}} f(t)\, dG_{\xi_0(x)}(t).$$

Lemma 3.3. *Under the assumptions of Lemma 3.2, it holds that, for every fixed $x \in \mathcal{X}$,*

$$\sum_{k=1}^{n} w_k(x) \left\| X_k - x \right\|_2 \xrightarrow{\mathbb{P}} 0.$$

The proof of this lemma is given in Appendix 3.A.

Lemma 3.4. *Let $x \in \mathcal{X}$. Assume that assumptions 3.1 to 3.4 hold and that $\lim_{n\to+\infty} m(n) = +\infty$. Then, for all upper semi-continuous function $f : [-\infty, +\infty) \to \mathbb{R}$ that is bounded above,*

$$\limsup_{n\to+\infty} \mathbb{P}_n^x[f] \leq G_{\xi_0(x)}[f] \quad \textit{in probability.}$$

Proof. Let M denote the supremum of f. Define $\tilde{f} = M - f$, which is lower semicontinuous and nonnegative. We have

$$\begin{aligned} \mathbb{P}_n^x[f] &= \sum_{k=1}^{n} w_k(x) \left(M - \tilde{f}\right)(\tilde{Z}_{k,m}) \\ &= M \sum_{k=1}^{n} w_k(x) - \sum_{k=1}^{n} w_k(x) \tilde{f}(\tilde{Z}_{k,m}) \\ &= M - \mathbb{P}_n^x[\tilde{f}] \quad \text{since } \sum_{k=1}^{n} w_k(x) = 1, \end{aligned}$$

and

$$\begin{aligned} G_{\xi_0(x)}[f] &= \int_{\mathbb{R}} \left(M - \tilde{f}\right)(t) dG_{\xi_0(x)}(t) \\ &= M - G_{\xi_0(x)}[\tilde{f}]. \end{aligned}$$

It is therefore sufficient to prove that

$$\liminf_{n\to+\infty} \mathbb{P}_n^x[\tilde{f}] \geq G_{\xi_0(x)}[\tilde{f}] \quad \text{in probability.} \qquad (\star)$$

Indeed,

$$\begin{aligned}\limsup_{n\to+\infty}\mathbb{P}_n^x[f]\le G_{\xi_0(x)}[f] &\Leftrightarrow \limsup_{n\to+\infty}\big(M-\mathbb{P}_n^x[\tilde f]\big)\le M-G_{\xi_0(x)}[\tilde f]\\ &\Leftrightarrow M-\liminf_{n\to+\infty}\mathbb{P}_n^x[\tilde f]\le M-G_{\xi_0(x)}[\tilde f]\\ &\Leftrightarrow \liminf_{n\to+\infty}\mathbb{P}_n^x[\tilde f]\ge G_{\xi_0(x)}[\tilde f].\end{aligned}$$

To establish $(\star)$, we use the representations

$$\mathbb{P}_n^x[\tilde f]=\int_0^1\tilde f\big(F_n^{x\leftarrow}(u)\big)\,du \quad\text{and}\quad G_{\xi_0(x)}[\tilde f]=\int_0^1\tilde f\big(G^{\leftarrow}_{\xi_0(x)}(u)\big)\,du,$$

with $F_n^{x\leftarrow}(u)=\inf\{t\in\mathbb{R}\mid F_n^x(t)\ge u\}$. By Lemma 3.2, we know that $\mathbb{F}_n^x(t)\xrightarrow{\mathbb{P}}G_{\xi_0(x)}(t)$ for all $t\in\mathbb{R}$. Furthermore, since $\mathbb{F}_n^x$ is nondecreasing and $G_{\xi_0(x)}$ is continuous, it follows that

$$\mathbb{F}_n^{x\leftarrow}(u)\xrightarrow{\mathbb{P}}G^{\leftarrow}_{\xi_0(x)}(u),\quad\forall u\in(0,1). \tag{3.18}$$

Let $(u_j)_{j\ge1}$ be a dense sequence in the interval $(0,1)$. Let (n_k) be an arbitrary subsequence of $(\mathbb{F}_n^{x\leftarrow})_{n\ge1}$. According to the convergence in probability given by the relation (3.18), we can extract from this arbitrary subsequence a diagonal sub-subsequence (n_{k_ℓ}) such that, almost surely

$$\forall j\ge1,\quad F_{n_{k_\ell}}^{x\leftarrow}(u_j)\to G^{\leftarrow}_{\xi_0(x)}(u_j)\quad\text{as }\ell\to+\infty.$$

The functions $u\mapsto F_{n_{k_\ell}}^{x\leftarrow}(u)$ are nondecreasing. By density of the set $\{u_j\}$ and continuity of $G^{\leftarrow}_{\xi_0(x)}$, we then obtain that, for almost every $u\in(0,1)$, $F_{n_{k_\ell}}^{x\leftarrow}(u)\xrightarrow{a.s.}G^{\leftarrow}_{\xi_0(x)}(u)$.

Since $\tilde f$ is lower semicontinuous, we obtain for all $u\in(0,1)$

$$\liminf_{\ell\to+\infty}\tilde f\left(\mathbb{F}_{n_{k_\ell}}^{x\leftarrow}(u)\right)\ge\tilde f\left(G^{\leftarrow}_{\xi_0(x)}(u)\right)\quad\text{a.s.},$$

and integrating yields

$$\begin{aligned}\int_0^1\tilde f\left(G^{\leftarrow}_{\xi_0(x)}(u)\right)du &\le \int_0^1\liminf_{\ell\to+\infty}\tilde f\left(\mathbb{F}_{n_{k_\ell}}^{x\leftarrow}(u)\right)du\quad\text{a.s.}\\ &\le \liminf_{\ell\to+\infty}\int_0^1\tilde f\left(\mathbb{F}_{n_{k_\ell}}^{x\leftarrow}(u)\right)du\quad\text{by Fatou's lemma,}\end{aligned}$$

which gives

$$G_{\xi_0(x)}[\tilde f]\le\liminf_{\ell\to+\infty}\mathbb{P}^x_{n_{k_\ell}}[\tilde f]\quad\text{a.s.}$$

We have shown that every subsequence (n_k) admits a sub-subsequence (n_{k_ℓ}) along which $\liminf_{\ell\to\infty}\mathbb{P}^x_{n_{k_\ell}}[\tilde f]\ge G_{\xi_0(x)}[\tilde f]$ almost surely. By Lemma 4.2 in (Kallenberg, 2002)), it then

follows that

$$\liminf_{n\to\infty} \mathbb{P}_n^x[\tilde{f}] \geq G_{\xi_0(x)}[\tilde{f}] \quad \text{in probability.}$$

This completes the proof of the Lemma 3.4. □

The Lemma 3.5 and proposition 3.1, which we will state below, are very useful for proving the convergence theorem.

Lemma 3.5. *Let $x \in \mathcal{X}$. We assume that Assumptions 1–4 are satisfied, that $\xi_0(x) > -1$, and that $\lim_{n\to\infty} \frac{m(n)}{\log n} = +\infty$. Let $Y_{k,m} = \ell_{\theta_0}\big((Z_{k,m} - b_m(x))/a_m(x)\big)$ and $f_m(x) = \mathbb{E}[Y_{k,m} \mid X_{k,m} = x]$. We further assume that $\{f_m\}$ is equicontinuous and that there exists a constant C such that, for all n and all $k = 1, \ldots, n$,*

$$\mathbb{E}\Big[\big|Y_{k,m}\big|^2 \;\Big|\; X_{k,m}\Big] \leq C. \tag{H}$$

Then

$$\mathbb{P}_n^x[\ell_{\theta_0}] = \sum_{k=1}^{n} w_k(x)\, Y_{k,m} \xrightarrow{\mathbb{P}} G_{\xi_0(x)}[\ell_{\theta_0}] \quad \text{as } n \to +\infty,.$$

with $\theta_0 = (0, 1, \xi_0(x))$.

The proof of this lemma is thecnical and is given in appendix 3.B.

Proposition 3.1. *Let $x \in \mathcal{X}$ and let $K \subset \Theta$ be a compact neighborhood of $\theta_0(x)$, with $\theta_0(x) = (0, 1, \xi_0(x))$. Under the conditions of Theorem 3.1, we have:*

$$\tilde{\theta}_n^K(x) \xrightarrow{\mathbb{P}} \theta_0(x) \quad \text{as } n \to +\infty,$$

where, $\tilde{\theta}_n^K$ is defined in (3.16).

Proof. Let $x \in \mathcal{X}$. From Lemmas 5 and 6 of (Dombry, 2015), we have the following results:

(i) For all $\theta \in \Theta$, we have $G_{\xi_0(x)}[\ell_\theta] \leq G_{\xi_0(x)}[\ell_{\theta_0}]$, with equality if and only if $\theta = \theta_0$.

(ii) For all $B \subset \Theta$, define $\ell_B(z) = \sup_{\theta\in B} \ell_\theta(z)$ for all $z \in \mathbb{R}$. For all $\theta \in \Theta$, let $B(\theta, \varepsilon)$ denote the open ball in Θ centered at θ with radius $\varepsilon > 0$. Then,

$$\lim_{\varepsilon\to 0} G_{\xi_0(x)}[\ell_{B(\theta,\varepsilon)}] = G_{\xi_0(x)}[\ell_\theta].$$

From (i) and (ii), for all $\theta \in K$ with $\theta \neq \theta_0$, there exists $\varepsilon_\theta > 0$ such that

$$G_{\xi_0(x)}[\ell_{B(\theta,\varepsilon_\theta)}] < G_{\xi_0(x)}[\ell_{\theta_0}]. \tag{3.19}$$

Fix $\delta > 0$ and define $\Delta = \{\theta \in K : \|\theta - \theta_0\| \geq \delta\}$. Δ is compact and can be covered by the open balls $\{B(\theta, \varepsilon_\theta), \theta \in \Delta\}$. Let $B_i = B(\theta_i, \varepsilon_{\theta_i})$, $i = 1, \ldots, p$, be a finite subcover (by the Borel–Lebesgue theorem). Since $\tilde{L}_n(\theta; x) = \mathbb{P}_n^x[\ell_\theta]$, we have

$$\begin{aligned}\sup_{\theta\in\Delta}\tilde{L}_n(\theta;x) &\leq \sup_{\theta\in\cup_{i=1}^{p}B_i}\tilde{L}_n(\theta;x)\\ &\leq \max_{1\leq i\leq p}\{\sup_{\theta\in B_i}\tilde{L}_n(\theta;x)\}\\ &\leq \max_{1\leq i\leq p}\{\sup_{\theta\in B_i}\sum_{k=1}^{n}w_k(x)\ell_\theta(\tilde{Z}_{k,m})\} \qquad (3.20)\end{aligned}$$

For each $i\in\{1,\dots,p\}$ and $k\in\{1,\dots,n\}$, we have

$$\forall\theta\in B_i,\quad \ell_\theta(\tilde{Z}_{k,m})\leq\sup_{\theta\in B_i}\ell_\theta(\tilde{Z}_{k,m}),$$

so that

$$\begin{aligned}\sum_{k=1}^{n}w_k(x)\ell_\theta(\tilde{Z}_{k,m}) &\leq \sum_{k=1}^{n}w_k(x)\sup_{\theta\in B_i}\ell_\theta(\tilde{Z}_{k,m})\\ &\leq \sum_{k=1}^{n}w_k(x)\ell_{B_i}(\tilde{Z}_{k,m}).\end{aligned}$$

Hence, $\forall i\in[|1,p|]$

$$\begin{aligned}\sup_{\theta\in B_i}\sum_{k=1}^{n}w_k(x)\ell_\theta(\tilde{Z}_{k,m}) &\leq \sum_{k=1}^{n}w_k(x)\ell_{B_i}(\tilde{Z}_{k,m})\\ &\leq \mathbb{P}_n^x[\ell_{B_i}]\end{aligned}$$

(3.20) therefore becomes

$$\sup_{\theta\in\Delta}\tilde{L}_n(\theta;x)\leq\max_{1\leq i\leq p}\mathbb{P}_n^x[\ell_{B_i}].$$

Since the function ℓ_{B_i} is upper semi-continuous and bounded from above, the assumptions of Lemma 3.4 are satisfied, and it follows that

$$\begin{aligned}\limsup_{n\to+\infty}\left(\sup_{\theta\in\Delta}\tilde{L}_n(\theta;x)\right) &\leq \limsup_{n\to+\infty}\left(\max_{1\leq i\leq p}\mathbb{P}_n^x[\ell_{B_i}]\right)\\ &\leq \max_{1\leq i\leq p}\left(\limsup_{n\to+\infty}\mathbb{P}_n^x[\ell_{B_i}]\right)\\ &\leq \max_{1\leq i\leq p}\left(G_{\xi_0(x)}[\ell_{B_i}]\right)\quad\text{in probability, by Lemma 3.4}\\ &< G_{\xi_0(x)}[\ell_{\theta_0}]\quad\text{in probability, by equation (3.19).}\end{aligned}$$

Moreover, for all $\theta\in K$, we have $\tilde{L}_n(\theta;x)\leq\sup_{\theta\in K}\tilde{L}_n(\theta;x)$, and in particular $\tilde{L}_n(\theta_0;x)\leq\sup_{\theta\in K}\tilde{L}_n(\theta;x)$,

so that

$$\begin{aligned}\liminf_{n\to+\infty}\left(\sup_{\theta\in K}\tilde{L}_n(\theta;x)\right) &\geq \liminf_{n\to+\infty}\left(\tilde{L}_n(\theta_0,x)\right)\\ &\geq G_{\xi_0(x)}[\ell_{\theta_0}] \quad \text{in probability, by Lemma 3.5.}\end{aligned}$$

Since $\tilde{\theta}_n^K(x)$ maximizes $\tilde{L}_n$ over K, we have

$$\liminf_{n\to+\infty}\left(\tilde{L}_n(\tilde{\theta}_n^K(x))\right) \geq G_{\xi_0(x)}[\ell_{\theta_0}] > \limsup_{n\to+\infty}\left(\sup_{\theta\in\Delta}\tilde{L}_n(\theta;x)\right).$$

which implies $\tilde{\theta}_n^K(x) \in K\backslash\Delta$ for large n. That is, $\|\tilde{\theta}_n^K(x) - \theta_0(x)\| < \delta$ for n sufficiently large. Since δ is arbitrary, this shows that

$$\tilde{\theta}_n^K(x) \xrightarrow{\mathbb{P}} \theta_0(x), \quad n \to \infty.$$

□

Proof of Theorem 3.1. Let $x \in \mathcal{X}$ and consider $K \subset \Theta$ a compact neighborhood of $\theta_0(x)$, and let $\tilde{\theta}_n^K(x) = \left(\tilde{\mu}_n^K(x), \tilde{\sigma}_n^K(x), \tilde{\xi}_n^K(x)\right)$ be defined as in Proposition 3.1. It suffices to take $\hat{\theta}_n(x) = \left(a_m(x)\tilde{\mu}_n^K(x) + b_m(x), a_m(x)\tilde{\sigma}_n^K(x), \tilde{\xi}_n^K(x)\right)$ to obtain the results (3.13) and (3.14) of Theorem 3.1. Indeed, with $\hat{\theta}_n(x) = (\hat{\mu}_n(x), \hat{\sigma}_n(x), \hat{\xi}_n(x))$ considered, we have

$$\left(\frac{\hat{\mu}_n(x) - b_m(x)}{a_m(x)}, \frac{\hat{\sigma}_n(x)}{a_m(x)}, \hat{\xi}_n(x)\right) = \left(\tilde{\mu}_n^K(x), \tilde{\sigma}_n^K(x), \tilde{\xi}_n^K(x)\right) = \tilde{\theta}_n^K(x).$$

By Proposition 3.1, $\tilde{\theta}_n^K(x) \xrightarrow{\mathbb{P}} \theta_0(x)$, which is equivalent to

$$\hat{\xi}_n(x) \xrightarrow{\mathbb{P}} \xi_0(x), \quad \frac{\hat{\mu}_n(x) - b_m(x)}{a_m(x)} \xrightarrow{\mathbb{P}} 0, \quad \frac{\hat{\sigma}_n(x)}{a_m(x)} \xrightarrow{\mathbb{P}} 1 \quad \text{as } n \to \infty,$$

and thus (3.14) holds. Moreover, since $\tilde{\theta}_n^K(x) = \arg\max_{\theta\in K} \tilde{L}_n(\theta;x)$, $\tilde{L}_n(\theta;x)$ attains a local maximum at $\tilde{\theta}_n^K(x)$ whenever $\tilde{\theta}_n^K(x) \in \text{Int}(K)$ (see proof of Theorem 2 in (Dombry, 2015)). As $\theta_0(x) \in \text{Int}(K)$, we have $\tilde{\theta}_n^K(x) \in \text{Int}(K)$ for n sufficiently large. Therefore, $\tilde{L}_n(\theta;x)$ admits a local maximum at $\tilde{\theta}_n^K(x)$ for large n, and by Lemma 3.1, $\hat{\theta}_n(x)$ is a maximum likelihood estimator for n sufficiently large. This proves (3.13) and completes the proof of Theorem 3.1.

□

3.5 Conclusion

In this paper, we proposed a method for estimating conditional quantiles at high probability levels. This approach addresses current challenges in quantile regression, particularly the dif-

ficulty of estimation when the covariate dimension is high and when the relationship between the response and explanatory variables is complex and highly non-linear. The performance of the proposed method has been illustrated through simulation studies and real-data applications in (Vidagbandji et al., 2025) and (Vidagbandji et al., 2026), which focused on two penalized versions of the estimator introduced here. The present work aims to establish the theoretical proof of existence and convergence of the weighted maximum likelihood estimator, where the weights are derived from the generalized random forests method introduced by (Athey et al., 2019), in the context of quantile regression.

Appendix

3.A Proof of lemma 3.3

Under the assumptions of Lemma 3.2, consider each tree $b = 1, \ldots, B$. For a fixed $x \in \mathcal{X}$, let $R_b(x)$ denote the leaf of tree b containing x, and let $|R_b(x)|$ be the number of observations in that leaf. The diameter of $R_b(x)$ is defined as

$$\operatorname{diam}(R_b(x)) = \sup\{\|y - x\|_2 : y \in R_b(x)\}.$$

By construction of the random forest, for $x \in \mathcal{X}$, the weights assigned to the observation X_k can be written as

$$w_k(x) = \frac{1}{B} \sum_{b=1}^{B} \frac{\mathbf{1}_{\{X_k \in R_b(x)\}}}{|R_b(x)|}.$$

Therefore,

$$\sum_{k=1}^{n} w_k(x) \|X_k - x\|_2 = \frac{1}{B} \sum_{b=1}^{B} \frac{1}{|R_b(x)|} \sum_{X_k \in R_b(x)} \|X_k - x\|_2.$$

For each tree b, every point in the leaf $R_b(x)$ lies at a distance from x bounded above by the diameter of the leaf. Hence, for all $b \in \{1, \ldots, B\}$,

$$\frac{1}{|R_b(x)|} \sum_{X_k \in R_b(x)} \|X_k - x\|_2 \leq \operatorname{diam}(R_b(x)).$$

It follows that

$$\sum_{k=1}^{n} w_k(x) \|X_k - x\|_2 \leq \frac{1}{B} \sum_{b=1}^{B} \operatorname{diam}(R_b(x)).$$

Assumption 3.2 ensures that, for each tree b, $\operatorname{diam}(R_b(x)) \xrightarrow{p} 0 \quad \text{as } s \to \infty$, and therefore as $n \to \infty$. Since the number of trees B is fixed, the average diameter also converges to zero in

probability. Indeed, for any $\varepsilon > 0$,

$$\mathbb{P}\left(\frac{1}{B}\sum_{b=1}^{B}\operatorname{diam}(R_b(x)) > \varepsilon\right) \leq \sum_{b=1}^{B}\mathbb{P}(\operatorname{diam}(R_b(x)) > \varepsilon) \longrightarrow 0.$$

Consequently,

$$\sum_{k=1}^{n} w_k(x)\|X_k - x\|_2 \xrightarrow{\mathbb{P}} 0.$$

3.B Proof of lemma 3.5

The proof of lemma 3.5, which is somewhat technical, is facilitated by Lemma 3.6 below.

Lemma 3.6. *Let $x \in \mathcal{X}$. Suppose that the assumptions of Lemma 3.5 are satisfied and let $T_{k,m} = Y_{k,m} - \mathbb{E}[Y_{k,m} \mid X_{k,m}]$. Then*

$$\sum_{k=1}^{n} w_k(x)\, T_{k,m} \xrightarrow{\mathbb{P}} 0.$$

Proof. Let $T_n = \sum_{k=1}^{n} w_k(x) T_{k,m}$. By the property of conditional expectation, we have

$$\begin{aligned}
\mathbb{E}[T_{k,m}^2 \mid X_{k,m}] &= \mathbb{E}\left[Y_{k,m}^2 + [\mathbb{E}(Y_{k,m} \mid X_{k,m})]^2 - 2Y_{k,m}\mathbb{E}(Y_{k,m} \mid X_{k,m})\middle| X_{k,m}\right] \\
&= \mathbb{E}\left[Y_{k,m}^2\middle| X_{k,m}\right] + \mathbb{E}\left[[\mathbb{E}(Y_{k,m} \mid X_{k,m})]^2\middle| X_{k,m}\right] - 2\left[\mathbb{E}(Y_{k,m}\middle| X_{k,m})\right]^2 \\
&\leq \mathbb{E}\left[Y_{k,m}^2\middle| X_{k,m}\right] - \mathbb{E}\left[Y_{k,m}\middle| X_{k,m}\right]^2 \\
&\leq \mathbb{E}\left[Y_{k,m}^2\middle| X_{k,m}\right] \\
&\leq C \quad \text{by assumption H.}
\end{aligned}$$

Therefore, since the $(T_{k,m}, X_{k,m})$ are independent for $k = 1, \cdots, n$, we have

$$\begin{aligned}
\mathbb{E}[T_n^2] &\leq \mathbb{E}\left[\sum_{k=1}^{n} w_k^2(x)\mathbb{E}(T_{k,m}^2 \mid X_{k,m})\right] + 2\mathbb{E}\left[\sum_{i<j} w_i(x)w_j(x)\mathbb{E}(T_{i,m}T_{j,m} \mid \sigma(X_{i,m}, X_{j,m}))\right] \\
&\leq C\mathbb{E}\left[\sum_{k=1}^{n} w_k^2(x)\right] + 2\mathbb{E}\left[\sum_{i<j} w_i(x)w_j(x)\mathbb{E}(T_{i,m} \mid X_{i,m})\mathbb{E}(T_{j,m} \mid X_{j,m})\right] \\
&\leq C\mathbb{E}\left[\sum_{k=1}^{n} w_k^2(x)\right] \quad \text{since} \quad \mathbb{E}(T_{j,m} \mid X_{j,m}) = 0.
\end{aligned}$$

Since $\sum_{k=1}^{n} w_k(x)^2 \leq \max_k w_k(x) \leq s/n$, it follows that

$$\mathbb{E}[T_n^2] \leq C\frac{s}{n} \to 0.$$

Therefore, $T_n \xrightarrow{L^2} 0$ and consequently $T_n \xrightarrow{\mathbb{P}} 0$.

□

Proof. **(Proof of lemma 3.5)** We have:

$$\sum_{k=1}^{n} w_k(x) Y_{k,m} - G_{\xi_0(x)}[\ell_{\theta_0}] = \underbrace{\sum_{k=1}^{n} w_k(x) \left(Y_{k,m} - \mathbb{E}[Y_{k,m} \mid X_{k,m}]\right)}_{T_n} + \underbrace{\sum_{k=1}^{n} w_k(x) \left(\mathbb{E}[Y_{k,m} \mid X_{k,m}] - G_{\xi_0(x)}[\ell_{\theta_0}]\right)}_{V_n},$$

By Lemma 3.6, we have $T_n \xrightarrow{\mathbb{P}} 0$. It remains to show the convergence in probability of V_n to 0. To this end, we use the local uniform convergence of $f_m(x) = \mathbb{E}[Y_{k,m} \mid X_{k,m} = x]$ toward the continuous function $\phi(x) := G_{\xi_0(x)}[\ell_{\theta_0}]$.

First, we establish the pointwise convergence of $f_m(\cdot)$ to $\phi(\cdot)$. For each fixed $x \in \mathcal{X}$, Assumption 3.1 ensures that the conditional distribution of $(Z_{k,m} - b_m(x))/a_m(x)$ given $X_{k,m} = x$ converges to $G_{\xi_0(x)}$. By the continuous mapping theorem, we obtain the convergence in distribution of $Y_{k,m}$ given $X_{k,m} = x'$ toward $\ell_{\theta_0}(Z)$, where $Z \sim G_{\xi_0(x')}$. Moreover, Hypothesis H provides a uniform L^2 bound, hence the sequence is uniformly integrable. Consequently,

$$f_m(x) := \mathbb{E}[Y_{k,m} \mid X_{k,m} = x] \longrightarrow \phi(x), \quad \forall x \in \mathcal{X}. \tag{3.21}$$

We now prove uniform convergence in a neighborhood of x. The limit function ϕ is continuous on the compact set $\mathcal{X}$. Let $\varepsilon > 0$. By continuity of ϕ at x, there exists $\delta_1 > 0$ such that for all x' with $\|x' - x\| < \delta_1$,

$$|\phi(x') - \phi(x)| < \frac{\varepsilon}{2}. \tag{3.22}$$

On the other hand, the pointwise convergence (3.21) together with the equicontinuity of $\{f_m\}$ implies uniform convergence on every compact set. In particular, on the closed ball, we have

$$\sup_{\|x'-x\|_2 \le \delta_1} |f_m(x') - \phi(x')| \longrightarrow 0 \quad \text{as } m \to \infty.$$

Thus, there exists N_1 such that for all $n \ge N_1$ (and hence $m = m(n)$ sufficiently large),

$$\sup_{\|x'-x\| \le \delta_1} |f_m(x') - \phi(x')| < \frac{\varepsilon}{2}. \tag{3.23}$$

Combining (3.22) and (3.23), for all $n \ge N_1$ and all x' such that $\|x' - x\|_2 < \delta_1$, we obtain

$$|f_m(x') - \phi(x)| \le |f_m(x') - \phi(x')| + |\phi(x') - \phi(x)| < \varepsilon.$$

Setting $\delta = \delta_1$, we obtain the following property: for every $\varepsilon > 0$, there exist $\delta > 0$ and N such

that for all $n \geq N$ and all x' satisfying $\|x' - x\| < \delta$,

$$|f_m(x') - \phi(x)| < \varepsilon,$$

that is,

$$\left|\mathbb{E}[Y_{k,m} \mid X_{k,m} = x'] - G_{\xi_0(x)}[\ell_{\theta_0}]\right| < \varepsilon. \tag{3.24}$$

Define $A_n = \{k : \|X_{k,m} - x\|_2 \geq \delta\}$. We have the following decomposition

$$|V_n| \leq \sum_{k \notin A_n} w_k(x)\left|\mathbb{E}[Y_{k,m} \mid X_{k,m}] - G_{\xi_0(x)}[\ell_{\theta_0}]\right| + \sum_{k \in A_n} w_k(x)\left|\mathbb{E}[Y_{k,m} \mid X_{k,m}] - G_{\xi_0(x)}[\ell_{\theta_0}]\right|.$$

The first term is bounded by ε. Indeed,

$$\begin{aligned}\sum_{k \notin A_n} w_k(x)\left|\mathbb{E}[Y_{k,m} \mid X_{k,m}] - G_{\xi_0(x)}[\ell_{\theta_0}]\right| &\leq \varepsilon \sum_{k \notin A_n} w_k(x) \quad \text{by equation (3.24)} \\ &\leq \varepsilon,\end{aligned}$$

since $\sum_{k \notin A_n} w_k(x) \leq 1$. For the second term, we use the following uniform bound provided by the Cauchy–Schwarz inequality and Hypothesis H:

$$\begin{aligned}\left|\mathbb{E}[Y_{k,m} \mid X_{k,m}]\right| &\leq \sqrt{\mathbb{E}[Y_{k,m}^2 \mid X_{k,m}]} \\ &\leq \sqrt{C}.\end{aligned}$$

Similarly, the function $x \mapsto |G_{\xi_0(x)}[\ell_{\theta_0}]|$ is bounded on the compact set $\mathcal{X}$. Hence, there exists a constant M such that

$$\left|\mathbb{E}[Y_{k,m} \mid X_{k,m}] - G_{\xi_0(x)}[\ell_{\theta_0}]\right| \leq M.$$

Therefore,

$$\sum_{k \in A_n} w_k(x)\left|\mathbb{E}[Y_{k,m} \mid X_{k,m}] - G_{\xi_0(x)}[\ell_{\theta_0}]\right| \leq M \sum_{k \in A_n} w_k(x).$$

Consequently,

$$|V_n| \leq \varepsilon + M \sum_{k \in A_n} w_k(x).$$

Moreover, for every $k \in A_n$, we have

$$1 \leq \frac{1}{\delta}\|X_{k,m} - x\|_2,$$

hence

$$\sum_{k \in A_n} w_k(x) \leq \frac{1}{\delta} \sum_{k=1}^{n} w_k(x)\|X_{k,m} - x\|_2.$$

Thus,

$$\begin{aligned}\mathbb{P}\left(|V_n| > 2\varepsilon\right) &\leq \mathbb{P}\left(\varepsilon < M \sum_{k\in A_n} w_k(x)\right) \\ &\leq \mathbb{P}\left(\varepsilon < \frac{M}{\delta}\sum_{k=1}^{n} w_k(x)\|X_{k,m}-x\|_2\right) \\ &\to 0,\end{aligned}$$

since

$$\sum_{k=1}^{n} w_k(x)\|X_{k,m}-x\|_2 \xrightarrow{\mathbb{P}} 0$$

by Lemma 3.3. We conclude that

$$V_n \xrightarrow{\mathbb{P}} 0.$$

In summary,

$$\sum_{k=1}^{n} w_k(x)Y_{k,m} - G_{\xi_0(x)}[\ell_{\theta_0}] = T_n + V_n,$$

with $V_n \xrightarrow{\mathbb{P}} 0$ and $T_n \xrightarrow{\mathbb{P}} 0$. Therefore,

$$\mathbb{P}_n^x[\ell_{\theta_0}] \xrightarrow{\mathbb{P}} G_{\xi_0(x)}[\ell_{\theta_0}].$$

□

CHAPTER QUATRE

GENERALIZED RANDOM FOREST FOR EXTREME QUANTILE REGRESSION

Les résultats présentés dans ce chapitre ont fait l'objet d'un article de recherche publié. Vidagbandji, L. M., Berred, A., Bertelle, C., & Amanton, L. (2025). Generalized random forest for extreme quantile regression. *Communications in Statistics - Simulation and Computation*, 1–24. https://doi.org/10.1080/03610918.2025.2543854 .

Contents

4.1 Introduction

Extreme events, such as floods, earthquakes, rogue waves, or financial crises, are rare but potentially devastating. Studying them helps assess associated risks. Extreme quantile regression models the relationship between covariates and the high-level quantiles of a response variable, capturing the factors influencing rare, tail-end outcomes. Specifically, in the univariate context, if $Y \in \mathcal{Y} \subset \mathbb{R}$ represents a random variable describing a risk factor dependent on a set of covariates represented by the random vector $X \in \mathcal{X} \subset \mathbb{R}^p$, the objective is to estimate the extreme conditional quantile defined by:

$$Q_x(\tau) = \inf\{y : F_{Y|X=x}(y) \geq \tau\},\ x \in \mathcal{X}, \tag{4.1}$$

with τ close to 1 and $F_{Y|X=x}$ being the conditional distribution of $Y|X = x$. Observations in the tails of a distribution are inherently rare, which poses a significant challenge when estimating the quantity (4.1). This difficulty stems from the limited data available in these extreme regions. Yet, rare events often play a crucial role in fields such as risk management, strategic planning, and decision-making. In extreme quantile regression, developing robust and accurate methods to handle these infrequent observations is essential to improve the reliability of models and predictions where extreme outcomes are critical. For instance, in risk analysis or hydrology, it is important when estimating unobserved water levels ((Gumbel, 1963), (Hu and Franzke, 2020)).

Let n denote the sample size available for analysis and $\tau = \tau_n$ (dependent on the sample size) the level of the quantile we seek to estimate. Classical quantile estimation methods work well when $n(1-\tau_n) \to \infty$ as $\tau_n \to 1$ (for $n \to +\infty$). In this case, the quantile we seek to estimate is within the sample, and there is also a large amount of data in the region of the quantile to be estimated. However, the situation differs when $n(1-\tau_n) \in [0, +\infty[$ as $\tau_n \to 1$ (when $n \to +\infty$). In this latter case, estimation involves extrapolating beyond the data range or into the distribution's tail, and τ_n is referred to as an extreme quantile level. In other words, the sought quantile lies outside the range of the available sample, making estimation more complex and requiring specific approaches to handle these borderline situation. Asymptotic results from extreme value theory (De Haan and Ferreira, 2006) allow extrapolation beyond observed data. Leveraging this, (Chernozhukov, 2005) introduced an extreme quantile regression method for linear relationships between Y and $X \in \mathbb{R}^p$. Subsequent models have been proposed by (Yu and Moyeed, 2001), (Koenker and Hallock, 2001), (Daouia et al., 2013), (Takeuchi et al., 2006), (Laksaci and Maref, 2009), (Gardes and Stupfler, 2015), (Gardes and Stupfler, 2019), and (El Methni and Stupfler, 2017).

The second difficulty faced by classical extreme quantile regression methods arises when the dimension of the predictor space $\mathcal{X}$ is large (i.e., p is large) and the relationship between

the characteristic variables and the response variable is complex. In high dimensions, a simple quantile regression model may introduce additional bias (Gnecco et al., 2024). Recent literature has seen the development of machine learning approaches to address this challenge. First, we can cite the work of (Meinshausen and Ridgeway, 2006), who proposed a method based on the random forest method of (Breiman, 2001). (Athey et al., 2019) proposed a quantile regression model based on the generalized random forest method. Their methods encounter difficulties mainly when the quantile of interest is extreme (as shown by the simulation studies in section 4.4). Quantile regression methods that combine extreme value theory and machine learning have emerged very recently and are competitive with existing methods when the quantile of interest is extreme and when the predictor space is high-dimensional. To our knowledge, the first model proposed in this sense is that of (Farkas et al., 2024), based on regression trees and the Peaks Over Threshold (POT) approach of extreme value theory. Next, we can mention the method of (Velthoen et al., 2023), which models the distribution of equation 4.1 using the conditional generalized Pareto distribution and the gradient boosting method. (Pasche and Engelke, 2024) propose a method that still combines extreme value theory to approximate the conditional distribution of equation 4.1 and neural networks to estimate the parameters of this distribution. (Gnecco et al., 2024) propose a method combining the POT approach and generalized random forests.

Previous work has mainly relied on the *Peaks Over Threshold* (POT) approach to address challenges in estimating extremes. While widely recognized for its effectiveness, POT has notable limitations, particularly when the available data consist solely of block maxima aggregated over fixed periods, such as annual maxima from long historical records or large-scale simulations. In such cases, the block maxima (BM) approach provides a more suitable alternative. The BM method is extensively used in environmental sciences, where the Generalized Extreme Value (GEV) distribution effectively models phenomena such as annual or monthly maximum temperatures and extreme river discharges. As highlighted by (Ferreira and De Haan, 2015), the BM approach offers several practical advantages: it is particularly appropriate when only periodic maxima are available, allows for temporal dependence within blocks (e.g., seasonality or short-term correlations), and is often simpler to implement since block periods naturally align with existing temporal structures, such as years or seasons.

In this study, applying the block maxima approach, we approximate the conditional distribution $Y|X = x$ of block maxima by a generalized extreme value distribution, whose parameters are functions of the features $x \in \mathcal{X}$. These parameters are estimated by a weighted maximum likelihood method, with weights obtained using the generalized random forest method developed by (Athey et al., 2019). Section 4.2 provides a reviews of quantile regression, the BM approach to extreme value theory and the generalized random forest method. Section 4.3 presents our proposed model in detail. Section 4.4 examines the results obtained when applying our extreme quantile regression method to simulated data under several scenarios. Finally,

section 4.5 presents an application to daily weather data from the Fort Collins station in Colorado (USA).

4.2 Framework and Related Work

4.2.1 Generalized extreme value distribution

In this section, we will present the approach of extreme value theory that we will use to address the first quantile regression challenge mentioned in the introduction: the block maxima approach. This method is based on the results of (Fisher and Tippett, 1928) and (Gnedenko, 1943), which provide the possible asymptotic distributions of the maximum of a sequence of random variables $X_1, \cdots, X_m$ drawn independently and identically distributed from a variable X with probability distribution F. These authors showed that there exist normalization constants $a_m > 0$ and $b_m \in \mathbb{R}$ such that:

$$\lim_{m \to +\infty} F^m(a_m x + b_m) = G_\xi(x),\ x \in \mathbb{R} \text{ and } \xi \in \mathbb{R}, \tag{4.2}$$

where G_ξ is a non-degenerate probability distribution defined by:

$$G_\xi(x) = \exp\left(-(1+\xi x)^{-\frac{1}{\xi}}\right) \text{ with } 1+\xi x > 0,$$

which is called the extreme value distribution. Any function F that satisfies equation (4.2) is said to belong to the max-domain of attraction of the extreme value distribution G_ξ and is denoted as $F \in \mathcal{D}(G_\xi)$. For further details on the latter, see Chapter 1 or the pioneering work of (De Haan and Ferreira, 2006). Considering $Y_1, \cdots, Y_N$ as a sequence of independent and identically distributed random variables according to the random variable Y with distribution function F, the block maxima method involves dividing the data into n blocks of equal size $m > 1$ (or nearly equal), denoted as

$$B_{k,m} = \{Y_{(k-1)m+1}, \cdots, Y_{km}\}, \text{ with } k = 1, \cdots, n.$$

For any $m > 1$, the distribution of $Z_k = \max_{B_{k,m}}(Y_i)$ is F^m and satisfies (4.2) with a certain normalization constant $a_m > 0$ and $b_m \in \mathbb{R}$. Therefore, its distribution is given by the generalized extreme value distribution with parameters (a_m, b_m, ξ_0). The BM method assumes that these block maxima, denoted Z_k, exactly follow the GEV distribution, and that the sequence of random variables $Z_1, \cdots, Z_n$ thus formed is also independent and identically distributed. The specificity of this method lies in the choice of the optimal block size. Increasing the block size m leads to an increase in the variance of the estimation, while decreasing the block size introduces a bias. Therefore, it is essential to find a trade-off between bias and variance when

defining the blocks to ensure accurate estimation. The BM method is commonly presented, discussed, and employed in the literature for modeling extremes. Among the reference works, one can refer to (Coles, 2001) and (De Haan and Ferreira, 2006). The general form of the Generalized Extreme Value distribution, first introduced by (Jenkinson, 1955), is expressed through the function $G_{\mu,\sigma,\xi}(.)$ given in the following equation

$$G_{\mu,\sigma,\xi}(z) = \begin{cases} \exp\left(-\left(1+\xi\frac{z-\mu}{\sigma}\right)_+^{-\frac{1}{\xi}}\right), & \xi \neq 0, \\ \exp\left(-\exp\left(-\frac{z-\mu}{\sigma}\right)\right), & \xi = 0, \end{cases} \qquad \forall z \in \mathbb{R}, \tag{4.3}$$

where $\mu \in \mathbb{R}$, $\sigma > 0$ and $\xi \in \mathbb{R}$. Depending on the sign of ξ, three domains of attraction are distinguished: the Fréchet domain of attraction if $\xi > 0$, the Gumbel domain of attraction if $\xi = 0$, and the Weibull domain of attraction if $\xi < 0$. The $\tau - th$ extreme quantile of the Generalized Extreme Value distribution is obtained using equation (4.1) and is given by

$$Q_\tau = \begin{cases} \mu + \frac{\sigma}{\xi}\left((-\ln(\tau))^{-\xi} - 1\right) & \text{if } \xi \neq 0, \\ \mu - \sigma\ln\left(-\ln\left(\tau\right)\right) & \text{if } \xi = 0, \end{cases}$$

where τ close to 1, $\mu \in \mathbb{R}$, $\sigma > 0$, and $\xi \in \mathbb{R}$. Estimating the quantile of the GEV distribution therefore requires the estimation of the parameters μ, σ and ξ. Among the various methods available, the maximum likelihood method is the most commonly used and is the focus of this study. The theoretical validity of this method for GEV parameter estimation has been demonstrated by (Dombry, 2015), (Dombry and Ferreira, 2019) and (Bücher and Segers, 2017). Denoting the parameter vector of GEV as $\theta = (\mu, \sigma, \xi)$, the negative log-likelihood of the GEV distribution for the sample $z_1, \cdots, z_n$ is given by

$$L(\theta; z_1, \cdots, z_n) = \frac{1}{n}\sum_{i=1}^{n} \ell_\theta(z_i), \tag{4.4}$$

where ℓ_θ is defined by

$$\ell_\theta(z_i) = \begin{cases} \log(\sigma) + (1+\frac{1}{\xi})\log\left(1+\xi\frac{z_i-\mu}{\sigma}\right) + \left(1+\xi\frac{z_i-\mu}{\sigma}\right)^{-\frac{1}{\xi}} & \text{if } 1+\xi\frac{z_i-\mu}{\sigma} > 0, \\ +\infty & \text{otherwise,} \end{cases}$$

when $\xi \neq 0$, and

$$\ell_\theta(z_i) = \log(\sigma) + \left(\frac{z_i-\mu}{\sigma}\right) + \exp\left(-\frac{z_i-\mu}{\sigma}\right),$$

when $\xi = 0$. The maximum likelihood estimator $\hat{\theta}$ satisfies

$$\hat{\theta} \in \arg\min_{\theta \in \Theta} L(\theta; z_1, \cdots, z_n),$$

where $\Theta \subset \mathbb{R} \times]0, +\infty[\times \mathbb{R}$.

4.2.2 Relevance of GRF Similarity Weights in the GEV Approach

In this section, we introduce the machine learning method we will be using to address the second problem, which aims to handle the high-dimensional space $\mathcal{X}$ of predictors and capture the complex relationship between the response variable and the features. The generalized random forests (Athey et al., 2019) is a generalization of the classical random forest method proposed by (Breiman, 2001). It retains the attractive features of classical random forests but allows customization of the loss function used for tree construction. Random forest is a learning method used for both classification and regression tasks. It belongs to the family of ensemble learning methods and provides a non-parametric estimation of the conditional mean. It involves aggregating B trees trained in parallel on bootstrap samples from the original training set, similar to the bagging method introduced by (Breiman, 1996). A key difference with bagging lies in how feature variables are used to split nodes in each tree. Instead of using all features at every node split, a random subset of features is selected. This random selection introduces additional variability between the trees, promoting diversity and improving the model's generalization (Breiman, 2001). Each tree provides an estimate of the conditional mean, which is obtained as the function minimizing the mean squared error. Further details can be found in Chapter 2 and in the foundational article (Athey et al., 2019). If $(X_i, Y_i) \in \mathcal{X} \times \mathcal{Y}$ where $\mathcal{X} \subset \mathbb{R}^p$ and $\mathcal{Y} \subset \mathbb{R}$, random forests are used to estimate $\mu(x) = \mathbb{E}(Y_i | X_i = x)$, for $x \in \mathcal{X}$. Let $\eta_b(x)$ denote the predicted value by the b^{th} tree for data $x \in \mathcal{X}$. In the case of regression, this prediction is given by

$$\eta_b(x) = \sum_{i=1}^{n} \frac{\mathbb{1}_{\{X_i \in R_b(x)\}}}{|\{i : X_i \in R_b(x)\}|} Y_i, \; b = 1, \cdots, B,$$

where $R_b(x) \subset \mathcal{X}$ denotes the region containing x in tree b, and $|E|$ denotes the cardinal of E. The random forest gives as its prediction the average of the predictions of the B trees, thus it predicts the value $\eta(x)$ given below for data x

$$\begin{aligned} \eta(x) &= \frac{1}{B} \sum_{b=1}^{B} \eta_b(x) \\ &= \sum_{i=1}^{n} w_n(x, X_i) Y_i, \end{aligned}$$

with

$$w_n(x,X_i) = \frac{1}{B}\sum_{b=1}^{B}\frac{\mathbb{1}_{\{X_i\in R_b(x)\}}}{|\{i : X_i \in R_b(x)\}|}. \tag{4.5}$$

The $w_n(x,X_i)$, for $i = 1,\cdots,n$, represent the similarity weights assigned by the random forest for each observation X_i given $x \in \mathcal{X}$. We have also

$$\begin{aligned}\sum_{i=1}^{n} w_n(x,X_i) &= \frac{1}{B}\sum_{i=1}^{n}\sum_{b=1}^{B}\frac{\mathbb{1}_{\{X_i\in R_b(x)\}}}{|\{i : X_i \in R_b(x)\}|}\\ &= 1.\end{aligned}$$

While the similarity weights $w_n(x,X_i)$ assigned by the generalized random forests (GRF) are computed using the same formula as in the classical forest, the key difference lies in the loss functions employed during tree construction. In classical random forests, a high weight is typically assigned to an observation X_i when $\mathbb{E}(Y|X = X_i) \approx \mathbb{E}(Y|X = x)$ (Meinshausen and Ridgeway, 2006). However, as shown by Figure 2 in (Athey et al., 2019) and Figure 1 in (Gnecco et al., 2024), these weights can still be large even when $Q(Y|X = X_i) \not\approx Q(Y|X = x)$, indicating that classical forests often fail to capture heterogeneity in the conditional quantile function. This limitation is addressed in GRF by incorporating the quantile loss into the tree construction process, enabling the method to target conditional quantiles directly. This approach has been successfully applied for both moderate and extreme quantile modeling by (Athey et al., 2019) and (Gnecco et al., 2024), respectively. Forests-based methods have the advantage of requiring few tuning parameters to provide effective predictions. This characteristic makes generalized random forests a natural choice for constructing similarity weights, especially as they capture the heterogeneity of conditional quantiles, a central aspect in our study. Unlike classical approaches such as kernel or nearest-neighbor methods, which are often limited when faced with non-linear structures or high-dimensional data, GRF exploit tree structure to accurately model local variability. They thus produce similarity weights that are both more robust and better suited to estimating extreme quantiles. What's more, compared with techniques such as gradient boosting or neural networks, GRF combine simplicity of use with sound theoretical foundations (Athey et al., 2019). They also offer better high-dimensional adaptability than generalized additive models (Koenker, 2011). Furthermore, the BM approach is better suited to random forests, as it generates i.i.d. observations from fixed block maxima, in line with the assumptions underlying random forests. Conversely, the POT method produces potentially correlated excesses, requiring additional adjustments, and relies on an often empirical threshold, which can weaken learning. This combination of robustness and flexibility makes GRF-based weights particularly well-suited to our framework. For clarity, we denote $w_i(x) \equiv w_n(x,X_i)$ for all $x \in \mathcal{X}$ and $i = 1,\ldots,n$.

4.3 GEV Extremal Random Forest

In this section, we present our method of extreme quantile regression that addresses the issues outlined in the introduction. To address the first issue, we model the tail of the conditional distribution $F(.|X=x)$ from equation (4.1) using the generalized extreme value distribution, as described in section 4.2.1. To tackle the second issue, we employ the generalized random forest method to determine the weights $w_n(x,X_i)$, which will be used to estimate the parameters of the GEV distribution. These parameters are estimated using weighted maximum likelihood estimators, following the approach proposed by (Gnecco et al., 2024) within the peaks-over-threshold framework. To proceed, let's consider an independent and identically distributed sample $(X_1,Y_1),\ldots,(X_N,Y_N)$ from the random vector $(X,Y)\in\mathcal{X}\times\mathcal{Y}$. Assume $N=n\times m$ and divide the sample into n blocks of size m, such that the k^{th} block for $k=1,\ldots,n$ is given by

$$B_{k,m}=\{(X_{(k-1)m+1},Y_{(k-1)m+1}),\cdots,(X_{km},Y_{km})\}.$$

Noting $Z_{k,m}=\max\{Y_i:(X_i,Y_i)\in B_{k,m}\}$ and $X_{k,m}$ the X_i corresponding to the Y_i maximizing $\{Y_i:(X_i,Y_i)\in B_{k,m}\}$. Thus, for all $m>1$, $\{(X_{k,m},Z_{k,m})\}_{k=1,\ldots,n}$ represents the sample of block maxima relative to the variable Y. For any $x\in\mathcal{X}$, we assume that the distribution F_x of $Z_{k,m}|X_{k,m}=x$ follows a Generalized Extreme Value distribution with parameter depending to the covariate x. More precisely, $\mu(\cdot):\mathcal{X}\to\mathbb{R}$ is the location parameter function, $\sigma(\cdot):\mathcal{X}\to\mathbb{R}_+^*$ is the scale parameter function and $\xi(\cdot):\mathcal{X}\to\mathbb{R}$ is the shape parameter function. The conditional GEV distribution is obtained by substituting $\theta=(\mu,\sigma,\xi)$ in (4.3) with $\theta(x)=(\mu(x),\sigma(x),\xi(x))$ for all $x\in\mathcal{X}$. The τ-th extreme quantile of the conditional GEV is obtained using equation (4.1) and is given for all $x\in\mathcal{X}$ by

$$Q_x(\tau)=\begin{cases}\mu(x)+\dfrac{\sigma(x)}{\xi(x)}\left((-\ln(\tau^m))^{-\xi(x)}-1\right) & \text{if } \xi(x)\neq 0,\\ \mu(x)-\sigma(x)\ln\left(-\ln\left(\tau^m\right)\right) & \text{if } \xi(x)=0.\end{cases} \tag{4.6}$$

Estimating this conditional quantile involves estimating the parameters $\mu(x)$, $\sigma(x)$, and $\xi(x)$ for all $x\in\mathcal{X}$. Our proposed method here consists of estimating these parameters and then substituting them into equation (4.6) to obtain the estimation of $Q_x(\tau)$. For this purpose, we propose a weighted form of the maximum likelihood estimator given in (4.4). As proposed by (Gnecco et al., 2024) for the POT approach, we estimate $\theta(x)$ by $\hat{\theta}(x)=(\hat{\mu}(x),\hat{\sigma}(x),\hat{\xi}(x))$, which is a weighted maximum likelihood estimator, i.e., it minimizes $L_n(\theta;x)$ defined by

$$L_n(\theta;x)=\sum_{i=1}^{n}w_i(x)\ell_\theta(z_i)\text{ for all }x\in\mathcal{X}, \tag{4.7}$$

with $\ell_\theta(z_i)$ defined in section 4.2.1. To capture the heterogeneity of the quantile function and improve estimation in a high-dimensional feature space, we obtain the weights $w_i(x)$ for $i = 1,\ldots,n$ using the generalized random forests method, as described in section 4.2.2.

It should be noted that the support of the GEV distribution depends on the extreme value index ξ, which is unknown (see (Dombry, 2015)). The usual regularity conditions ensuring good asymptotic properties of the maximum likelihood estimator (MLE) are not satisfied. This problem is first studied by (Smith, 1985), who shows that there is no maximum likelihood estimator if $\xi \leq -1$, that asymptotic normality holds if $\xi > -0.5$, and consistency holds if $\xi > -1$. (Dombry, 2015) demonstrated the existence and consistency of this estimator for the GEV distribution parameters when $\xi > -1$. The asymptotic normality of the MLE for the block maxima approach was proven by (Dombry and Ferreira, 2019). In general, $L_n(\theta;x)$ does not admit a global minimum, and we refer to the work of (Dombry, 2015) to state that $\hat{\theta}(x)$ is a weighted maximum likelihood estimator if $\hat{\theta}(x)$ is a local minimum. Thus, we define $\hat{\theta}(x)$ to be the vector that minimizes $L_n(\theta;x)$ over a large compact set $\Theta \subset \mathbb{R}\times]0,+\infty[\times]-1,+\infty[$. The weighted maximum likelihood estimator is given by

$$\hat{\theta}(x) \in \arg\min_{\theta\in\Theta} L_n(\theta;x). \tag{4.8}$$

We take the intermediate order $\tau_0 = 0.8$ as considered in the works of (Gnecco et al., 2024) and (Velthoen et al., 2023). Thus, we estimate the conditional quantile of order $\tau > \tau_0$ by:

$$\hat{Q}_\tau(x) = \begin{cases} \hat{\mu}(x) + \dfrac{\hat{\sigma}(x)}{\hat{\xi}(x)}\left(\left(-\ln(\tau^m)\right)^{-\hat{\xi}(x)} - 1\right) & \text{if} \quad \hat{\xi}(x) \neq 0, \\ \hat{\mu}(x) - \hat{\sigma}(x)\ln\left(-\ln\left(\tau^m\right)\right) & \text{if} \quad \hat{\xi}(x) = 0. \end{cases} \tag{4.9}$$

where $\hat{\mu}(x)$, $\hat{\sigma}(x)$ and $\hat{\xi}(x)$ are the estimated parameters obtained using the weighted maximum likelihood method. The algorithm 2 details our method for predicting the extreme conditional quantiles, which we have named GEV-erf. This method combines the use of the generalized extreme value distribution with generalized random forests to estimate extreme conditional quantiles precisely and efficiently.

Among the parameters of the GEV distribution, the shape parameter ξ is of particular importance, as it provides information on the shape of the tail of the distribution. A distribution is considered to have a heavy tail when $\xi > 0$, a light tail when $\xi = 0$, and a finite endpoint distribution (i.e. $x_F = \{y \in \mathbb{R}, F(y) < 1\}$ finite) when $\xi < 0$. Because of its importance, this parameter has received considerable attention in the literature. Many methods have been proposed for estimating ξ, the most notable examples being Hill's estimator, which focuses on the most important data points for estimating tail behavior, and Pickands' estimator. In line with the proposals of (Gnecco et al., 2024) for maximum likelihood in the context of the POT approach, and (Bücher et al., 2021) for the maximum likelihood of the GEV, we penalize the

negative log-likelihood in the following form

$$L_n^{pen}(\theta,x) = \sum_{i=1}^{n} w_i(x)\ell_\theta(z_i) + \lambda(\xi - \xi_0)^2.$$

The parameter ξ_0 used here corresponds to the shape parameter derived from the unconditional GEV distribution, which means it is obtained by setting $w_i(x) = 1$ for all indices $i \in \{1,\cdots,n\}$ in equation (4.7). The penalty parameter λ is determined by the cross-validation method, thus regularizing the negative log-likelihood in the estimation process. The resulting estimator, $\hat{\theta}_{pen}(x)$, is the weighted and penalized maximum likelihood estimator, which adjusts the GEV parameters by considering both the data and the model complexity, thereby ensuring better generalization. It is defined as :

$$\hat{\theta}_{pen}(x) \in \arg\min_{\theta\in\Theta} L_n^{pen}(\theta,x).$$

The algorithm to obtain the conditional quantile using $\hat{\theta}_{pen}(x)$ is similar to the algorithm 2 with the difference that the log-likelihood $L_n(\theta,x)$ is replaced by the penalized likelihood $L_n^{pen}(\theta,x)$. In the simulation study described in section 4.4, we provide a detailed explanation of the validation method used to determine the parameter λ. This method includes a series of tests and analyses to ensure the accuracy and reliability of the estimates. Specifically, we discuss the different steps of the validation process, the criteria used to evaluate the performance of the method, and the adjustments made to optimize the estimation of the parameter λ. The algorithm detailed below outlines our proposed method.

Algorithm

The algorithm for our extreme quantile regression method, denoted GEV-erf, consists of two sub-algorithms: GEV-erf-fit and GEV-erf-predict. The first sub-algorithm, GEV-erf-fit, is used for training the generalized random forest to obtain the weights $w_i(x)$ and to obtain the sample formed by the block maxima. The second sub-algorithm, GEV-erf-predict, is used to make predictions of the GEV distribution parameters and hence extreme conditional quantiles. It is presented as follows :

4.4 Simulation Study

In this section, we present a simulation study to demonstrate the effectiveness of our proposed extreme quantile regression method. We generate an independent and identically distributed sample of size $N = 90,000$ from the random variable X, following a uniform distribution over $[-1,1]^p$. We assume that the conditional distribution of $Y|X = x$ is of the form $\gamma(x)T_{\nu(x)}$, where

Algorithm 1 GEV-erf

Let $\mathcal{D}_N = \{(X_i, Y_i)\}_{i=1\cdots N}$ denote the original sample, $\mathcal{D}'_n = \{(X_i, Z_i)\}_{i=1\cdots n}$ denote the sample of block maxima, where m is the block size, and α is the vector containing hyperparameters associated with the generalized random forest.

1: **procedure** **GEV-erf-fit**

2: **Input**:($\mathcal{D}_N$, m, α)

3: $\mathcal{D}'_n \leftarrow makeBloc(\mathcal{D}_N, m)$

4: $w_i(.) \leftarrow GRF(\mathcal{D}'_n, \alpha)$

5: **Output:** GEV-erf $\leftarrow (\mathcal{D}'_n, w_i(.), m)$

1: **procedure** **GEV-erf-predict**

2: **Input**:(GEV-erf, x, τ)

3: $\hat{\theta}(x) \leftarrow \arg\min_{\theta\in\Theta} L_n(\theta; x)$

4: $\hat{Q}_x(\tau) \leftarrow GEV(\hat{\theta}(x))$

5: **Output:** $\hat{Q}_x(\tau)$

- The function **makeBloc** divides the initial sample into blocks of a given size m and returns the sample of block maxima.
- The **GRF** function is used to adjust the weights using the generalized random forests method proposed by (Athey et al., 2019).

T_k denotes the Student's distribution with k degrees of freedom. We conduct various experiments by varying the values of $\gamma(x)$ and $\nu(x)$. The choice of block size is a key challenge in the BM approach, especially since, as shown in Appendix 4.C, the GEV-erf model is sensitive to this parameter. We therefore select a fixed block size of $m = 40$ for all scenarios, as our analysis indicates that model performance remains stable for block sizes ranging from 25 to 50. For more details on sensitivity analysis, see Appendix 4.C.

In our first simulation study (scenario 1), we set the dimension of the feature space to $p = 40$, and evaluate the performance of our method. We then compare its ability to address the first concern stated in the introduction with other methods based on statistical learning approaches. The N data points are used for the training process, while the evaluation of the method is conducted on an independent sample from the training set, $\{(x_i, y_i)\}_{i=1}^{n'}$, where $x_i \in \mathcal{X}$ for $i = 1, \ldots, n'$ is generated using the Halton sequence method by (Halton, 1964) with $n' = 8000$. In the second simulation study (scenario 2), we demonstrate our method's capability to handle the second concern, namely complex dependence on covariates. For this, we conduct

simulations with functions $\gamma(.)$ and $\nu(.)$ chosen to introduce complex dependence with the covariates. In the last simulation study of this section (scenario 3), we show the effectiveness of our method in addressing both concerns simultaneously.

We proceed with comparing the performance of our method with two other random forest-based approaches: Quantile Regression Forests (QRF) by (Meinshausen and Ridgeway, 2006) and Generalized Random Forests (GRF) by (Athey et al., 2019). To assess these performances, we first employ the Mean Integrated Squared Error (MISE), which is computed as the average of m' Integrated Squared Error (ISE) values for the estimated conditional quantile, derived from repeating the training and testing process m' times. This metric aligns with the approach used by (Gnecco et al., 2024) in his work. Additionally, we evaluate performance using two other metrics: the Mean Absolute Error (MAE) and the Median Absolute Error (MedAE). The ISE on the test sample $\{(x_i, y_i)\}_{i=1}^{n'}$ is defined as follows:

$$ISE = \frac{1}{n'} \sum_{i=1}^{n'} \left(\hat{Q}_{x_i}(\tau) - Q_{x_i}(\tau)\right)^2, \tag{4.10}$$

where $x \mapsto Q_x(\tau)$ is the true quantile function, $x \in \mathcal{X}$ and $\tau > 0.8$.

4.4.1 Simulations scenarios

We demonstrate the capability of our proposed method across three scenarios:

▷ **Scenario 1:** We take

$$\gamma(x) = 1 + \mathbb{1}_{\{x_1 > 0\}} \text{ and } \nu(x) = 4 - (x_1^2 - 2x_2^2 + x_3^2).$$

The goal is to assess our method's ability to estimate extreme conditional quantiles in the presence of noise and when p is high.

▷ **Scenario 2:** We take

$$\gamma(x) = 2 + \frac{1}{1 + \exp(x_1^2 + x_2)} \text{ and } \nu(x) = 4 - (x_1^2 - 2x_2^2 + x_3^2).$$

The objective is to evaluate our method's capability to predict extreme conditional quantiles under complex forms of the quantile function. We set the feature size to $p = 40$, where only the first three components contribute to the quantile function formation, and the others are noise.

▷ **Scenario 3:** We consider

$$\gamma(x) = 1 + 2\pi\varphi(2x_1, 2x_2) \text{ and } \nu(x) = 3 + \frac{7}{1 + \exp(4x_1 + 1.2)}.$$

In this study, φ represents the density of the centered bivariate normal distribution with unit variance and a correlation coefficient of 0.75. The parameters chosen here are the same as those used in experiment 3 of the simulation conducted by (Gnecco et al., 2024). The aim of this scenario is to evaluate our method's ability to accurately estimate the conditional quantile in situations where the quantile function is complex, the dimension of feature variables is high, and noise is present.

4.4.2 Parameters tuning choice

We obtain the penalty parameter λ and the parameters for generalized random forest (specifically min.node.size and num.trees) using cross-validation. Instead of using the classic form (as explained in Section 7.10 of (Hastie, 2017)), we follow the approach used by (Gnecco et al., 2024), using GEV deviance as the metric. More clearly, if we consider that the training sample we have is $\mathcal{D}_n = \{(x_i, z_i)\}_{i=1,\cdots,n}$, and we wish to perform a K-cross validation, the method involves partitioning the sample into K folds of approximately equal size, denoted as $\mathcal{D}^j$ for $j = 1, \cdots, K$. For each tuning parameter $\alpha_1, \cdots, \alpha_S$, the model is trained on $\mathcal{D}_n \backslash \mathcal{D}^j$ using the GEV-erf-fit algorithm, and then parameters $\hat{\theta}(x)$ are estimated on $\mathcal{D}^j$ for each $j = 1, \cdots, K$. The performance metric, here the negative log-likelihood of GEV, is calculated for each fold, and the average error across the K folds provides an overall estimate of model performance. The α_s achieving the best performance is selected, corresponding to minimizing the cross-validation error (CV-error) across s,

$$CV(\alpha_s) = \frac{1}{K} \sum_{j=1}^{K} \sum_{(x_i, z_i) \in \mathcal{D}^j} \ell_{(\hat{\theta}(x_i), \alpha_s)}(z_i), \tag{4.11}$$

where $\ell_{(\hat{\theta}(x), \alpha_s)}$ represents the negative log-likelihood defined in equation (4.4) and the α_s for all $s \in \{1, \cdots, S\}$ designate the tuning parameters from which we seek to select the optimal choice. The selection of the parameters λ and *min.node.size* for the different scenarios is detailed in section 4.A of the appendix. Our model is called GEV-erf, but we use the abbreviation *gev* to represent it in the various plots and tables showcasing the model performances.

4.4.3 Performance of GEV-erf with Scenario 1

In Scenario 1, we rigorously evaluated the performance of our method by comparing it with Quantile Regression Forest (QRF) proposed by (Meinshausen and Ridgeway, 2006) and Generalized Random Forest (GRF) by (Athey et al., 2019). For this comparative analysis, we fixed the number of predictors to $p = 40$ and plotted $\log(\text{MISE})$ against the quantile level τ. The MISE was computed from $m' = 50$ repetitions of the ISE (defined in (4.10)), ensuring robust and statistically significant evaluation. Figure 4.1 clearly illustrates that our model is compet-

itive compared to QRF and GRF models, especially for moderate quantile orders. This figure also highlights that our method provides better estimation of conditional quantiles, particularly as the quantile level is closed to 1.

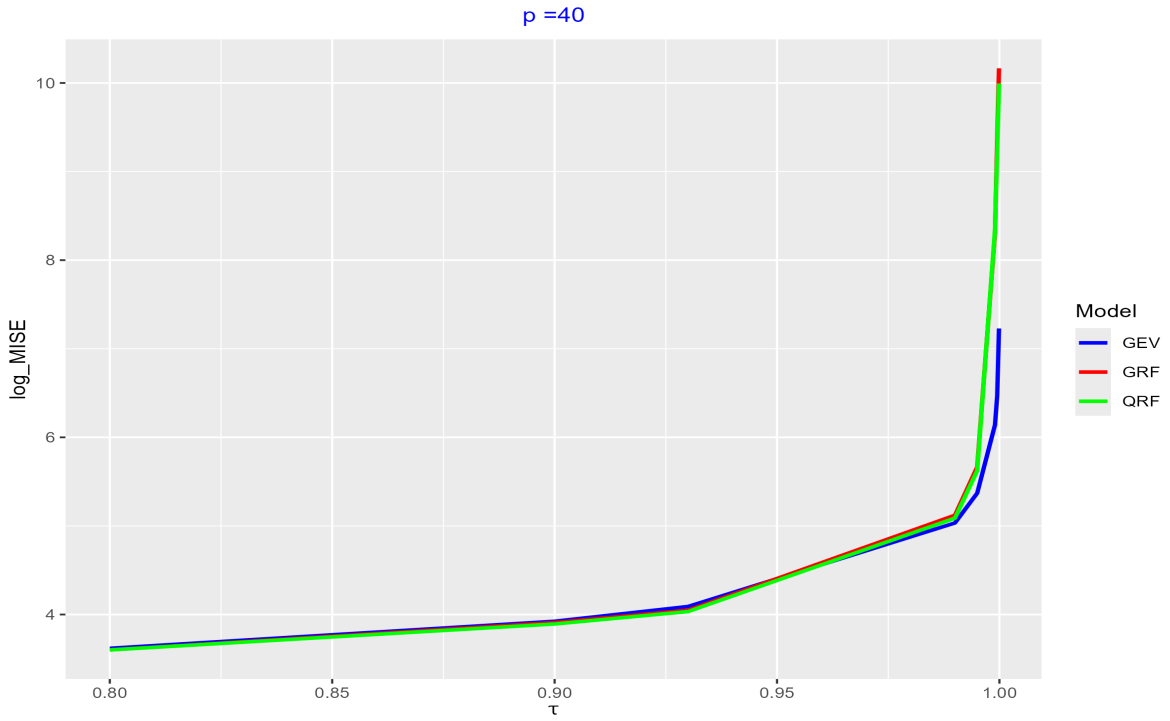

Figure 4.1: Logarithm of MISE for different methods as a function of τ.

Furthermore, our method demonstrates increased stability and reduced variance in quantile predictions, thanks to our innovative approach of local weighting. These results suggest that our model not only competes with existing methods but also outperforms them in scenarios where precise estimates of extreme quantiles are crucial. This is particularly relevant in applications requiring accurate risk management and decision-making based on reliable predictions of rare events.

4.4.4 Performance of GEV-erf with Scenario 2

In the figure 4.2, we evaluate the performance of our model for scenario 2. To do this, we present box plots of log(ISE) for different quantile orders $\tau \in \{0.99, 0.995, 0.999, 0.9995\}$, with p fixed at 40. The functions $\gamma(.)$ and $\nu(.)$ are chosen to make the conditional quantile function complex. The boxplots in this figure demonstrate that our model outperforms the QRF model and the GRF model for various quantile orders. This indicates the ability of our model to make accurate predictions even when the quantile function is complex. This scenario showcases our method's ability to effectively predict the conditional quantile of the dependent variable, outperforming the GRF and QRF methods even in the presence of noise and when

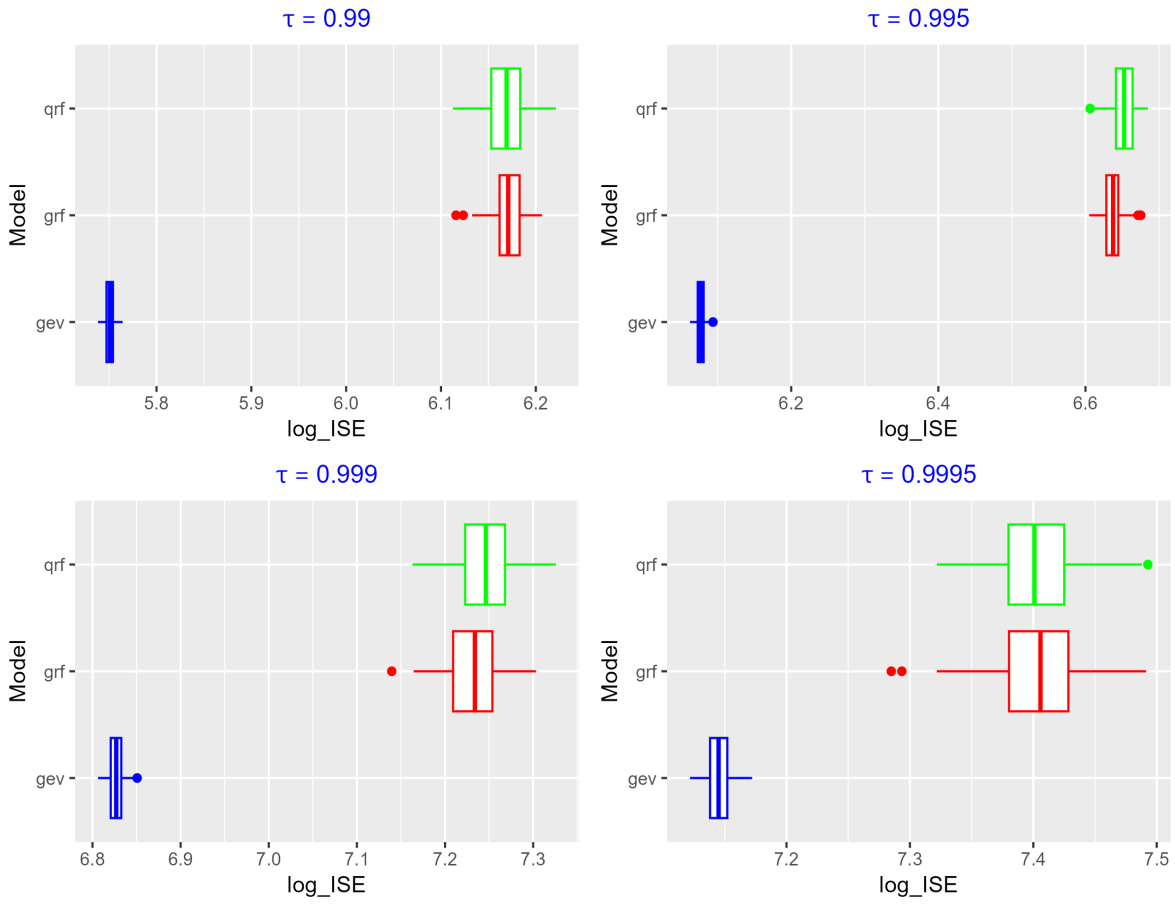

Figure 4.2: Boxplot of $log(ISE)$ over 100 replication, for $p = 40$ and $\tau = (0.99, 0.995, 0.999, 0.9995)$ in Scenario 2.

$Q_\tau(x)$ is complex. This confirms that our method addresses the previously outlined challenges, providing accurate predictions under difficult conditions where the relationship between the conditional quantile function and the covariates is complex and noisy.

4.4.5 Performance of GEV-erf with Scenario 3

In this scenario, the covariate size is fixed at $p = 50$, and the functions $\gamma(.)$ and $\nu(.)$ are chosen to make more complex the relationship between the covariates and the conditional quantile of the dependent variable. This scenario highlights the performance of our GEV-erf method in predicting extreme quantiles when the covariate size is large, the quantile function is complex, and in the presence of noise. Figure 4.3 displays the boxplot of $\log(ISE)$ for our model as well as for the GRF and QRF models, across different quantile orders. This figure demonstrates that our GEV-erf method outperforms the GRF and QRF methods, indicating its effective capture of the complex structure of extreme quantiles, its adaptation to the high dimensionality of covariates, and its ability to address the challenges outlined in the introduction. In figure 4.4, we present the boxplot of $\log(ISE)$ for various methods, considering different covariate sizes

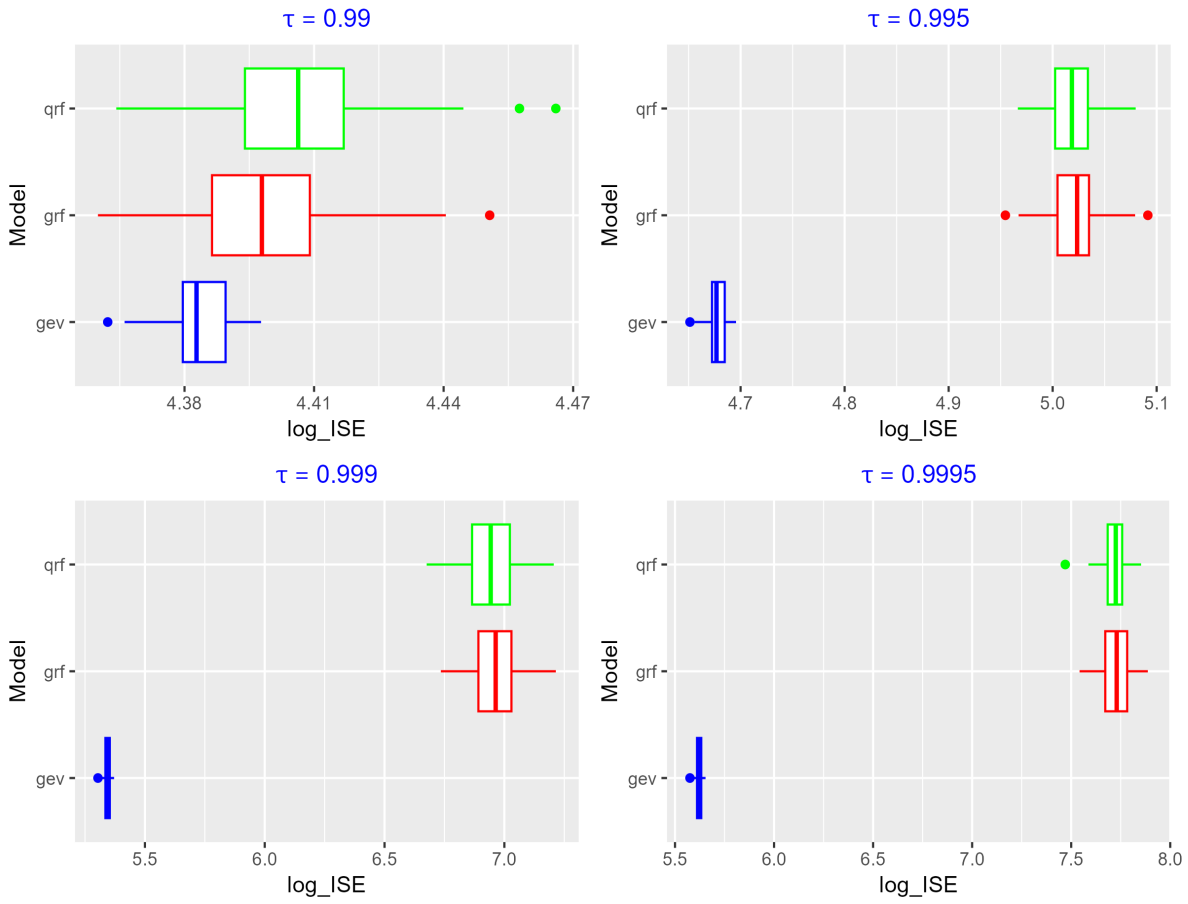

Figure 4.3: Boxplot of $log(ISE)$ for $p = 50$ and $\tau = (0.99, 0.995, 0.999, 0.9995)$ (Scenario 3).

and with quantile order set at 0.999. This graph shows that our method is efficient for different values of p. In terms of performance, our method is superior to GRF and QRF for extreme quantile regression analysis, particularly when the quantile function is complex and the size of the covariates is large. This is particularly evident in scenarios where capturing the subtleties of the quantile function is crucial. The performance of our method remains robust and reliable, highlighting its advantages over other methods.

We evaluate the performance of the models using two additional error metrics: the mean absolute error (MAE) and the median absolute error (MedAE). The results obtained are summarized in Tables 4.1, 4.2, and 4.3, corresponding to the three scenarios of the study. The analysis of these results, conducted for three levels of high quantiles ($\tau = 0.99, 0.995, 0.9995$), demonstrates that the GEV-erf model systematically outperforms the GRF and QRF models. Indeed, GEV-erf exhibits significantly lower errors as well as increased stability, even for extreme quantiles. Conversely, the GRF and QRF models show a strong sensitivity to the highest quantiles ($\tau = 0.9995$), resulting in a marked increase in errors. Moreover, the detailed analysis by scenario confirms this trend: in scenario 1, GEV-erf maintains stable performance, while the errors for the GRF and QRF models increase sharply for extreme quantiles; in scenario

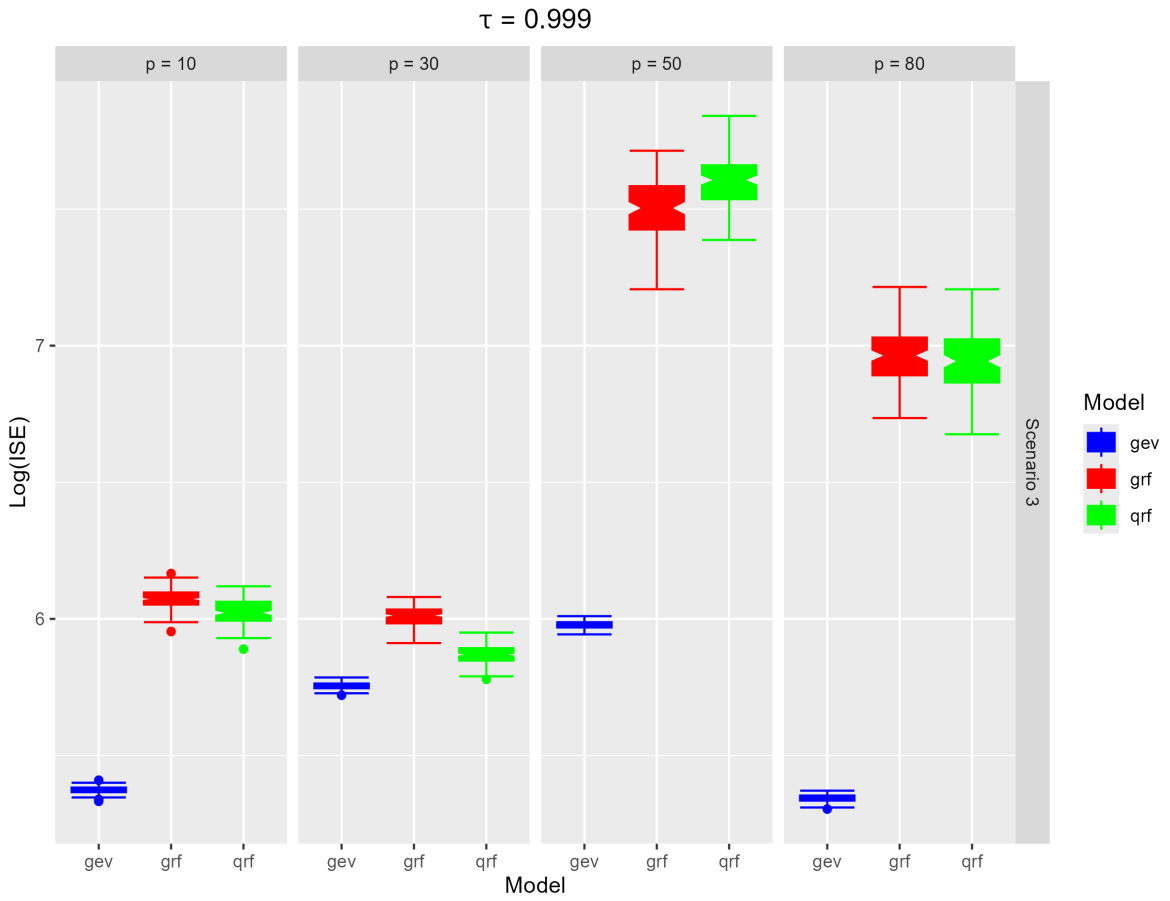

Figure 4.4: Boxplot of $log(ISE)$ for $p \in \{10, 30, 50, 80\}$ and $\tau = 0.999$ (Scenario 3).

2, a general degradation in performance is observed, but GEV-erf retains relatively higher accuracy; finally, in scenario 3, overall performance improves for all models, with GEV-erf nonetheless confirming its superiority. In summary, the analysis highlights the robustness and precision of the GEV-erf model for estimating high quantiles, while the GRF and QRF models reveal their limitations in dealing with extreme quantiles. These findings further support the conclusions previously obtained.

Scenario 1						
Model	**MAE**			**MedAE**		
	$\tau = 0.99$	$\tau = 0.995$	$\tau = 0.9995$	$\tau = 0.99$	$\tau = 0.995$	$\tau = 0.9995$
gev	12.4	14.6	24.9	12.3	14.7	25.5
grf	12.8	16.5	71.8	12.8	16.3	31.7
qrf	12.5	15.9	68.7	12.5	15.1	31.1

Table 4.1: Performance of models for different metrics.

In the appendix 4.B, we provide additional graphical representations for the different scenarios considered in this work. These graphs validate our conclusions by illustrating the

Scenario 2						
Model	**MAE**			**MedAE**		
	$\tau = 0.99$	$\tau = 0.995$	$\tau = 0.9995$	$\tau = 0.99$	$\tau = 0.995$	$\tau = 0.9995$
gev	17.9	21.0	34.8	18.1	21.3	36.5
grf	18.5	22.2	83.7	18.3	21.8	38.7
qrf	18.8	22.9	88.1	18.5	22.0	37.4

Table 4.2: Performance of models for different metrics.

Scenario 3						
Model	**MAE**			**MedAE**		
	$\tau = 0.99$	$\tau = 0.995$	$\tau = 0.9995$	$\tau = 0.99$	$\tau = 0.995$	$\tau = 0.9995$
gev	9.50	11.3	19.7	9.08	10.8	19.3
grf	10.1	12.7	36.2	9.85	12.2	33.7
qrf	10.3	12.8	34.1	10.9	11.8	31.7

Table 4.3: Performance of models for different metrics.

method's performance for different orders of high quantiles. In so doing, we provide an overview of the efficiency and robustness of our approach in handling extreme quantile regression tasks.

4.5 Applications to real datasets

To evaluate the performance of the extreme quantile regression model GEV-erf in a real-world setting, we use daily meteorological data from the Fort Collins station in Colorado (USA), recorded between January 1, 1900, and December 31, 1999. This dataset, available in (Sielenou and Alain, 2020), has been used in previous studies such as (Dkengne et al., 2020) and (Katz et al., 2002). In our analysis, the response variable Y corresponds to the daily maximum temperature (in degrees Fahrenheit), while the covariates include daily accumulated precipitation (in inches), daily accumulated snowfall, and two transformed variables: the variable *Season*, which takes values 1 (winter), 2 (spring), 3 (summer), and 4 (fall), and the maximum temperature of the previous day. Observations with missing values in any of the variables were excluded, resulting in a total of 35,793 complete observations. To increase the dimensionality of the covariate space and test the robustness of the model in a high-dimensional setting, we add six independent random variables generated from a uniform distribution on $[-1,1]$, bringing the total number of covariates to $p = 10$. The analysis focuses on monthly maximum temperatures, obtained by forming monthly blocks from the daily data. For model training, 70% of the monthly maxima are used as the training set, with the remaining 30% reserved for testing.

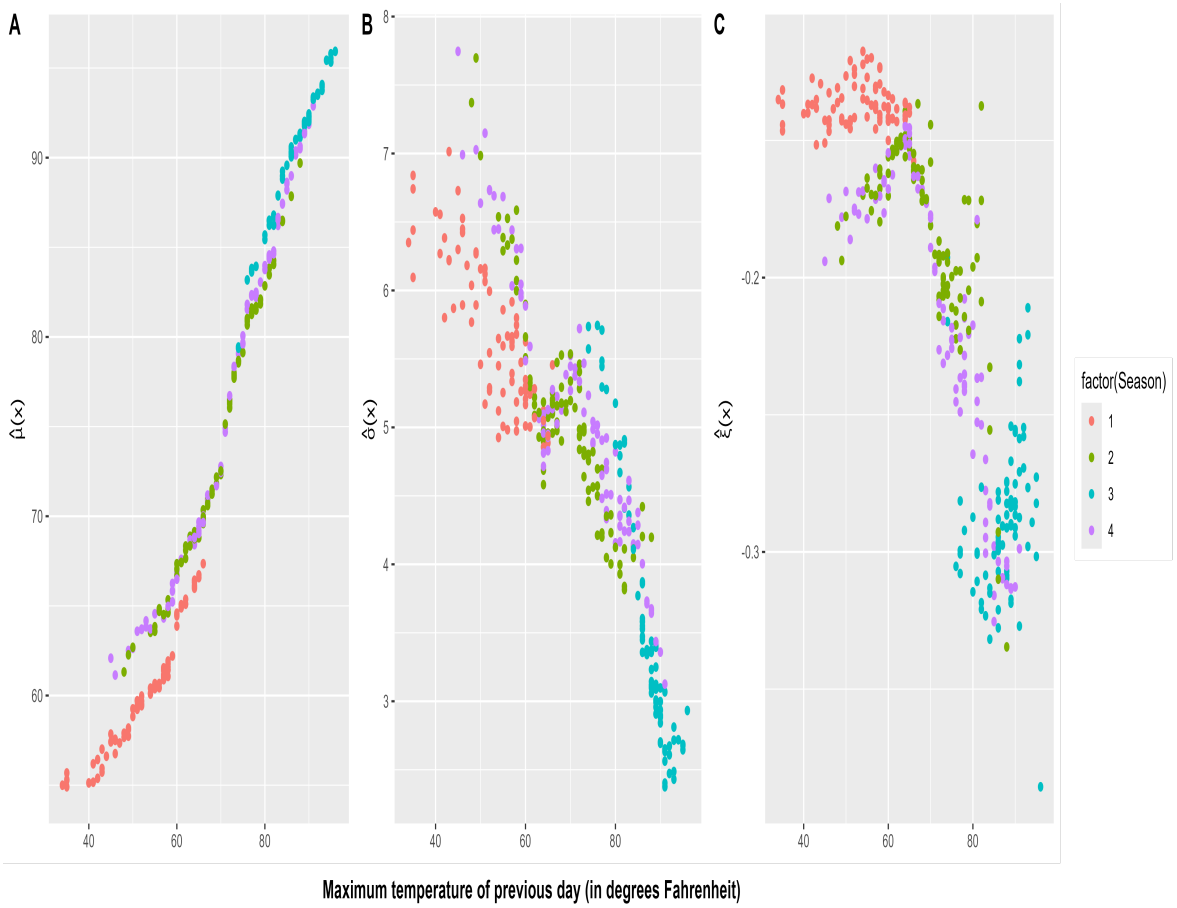

Figure 4.5: Variation of the estimated parameters $\hat{\theta}(x)$ as a function of the previous day's maximum daily temperature, by season.

The results reveal a strong dependence between the monthly maximum temperature and the previous day's maximum temperature. Figure 4.5 illustrates the variation of the estimated GEV distribution parameters $\hat{\theta}(x)$ with respect to this covariate, with a distinction made across seasons. A nearly linear increasing relationship is observed between the previous day's maximum temperature and the location parameter $\hat{\mu}(x)$, suggesting that warmer preceding days raise the baseline level of expected extreme temperatures. The scale parameter $\hat{\sigma}(x)$ decreases as the previous day's temperature increases, indicating reduced variability in the extremes. The shape parameter $\hat{\xi}(x)$ exhibits a more complex dynamic: for moderate temperatures, $\hat{\xi}(x)$ remains close to zero, indicating a Gumbel-type tail (moderately heavy). However, for higher temperatures—particularly during summer and autumn—$\hat{\xi}(x)$ becomes significantly negative. This implies a shorter tail, suggesting the presence of an upper bound on extreme temperatures. These findings are consistent with the conclusions of (Dkengne et al., 2020), who showed that the unconditional distribution of daily maximum temperature belongs to the Weibull domain of attraction ($\xi < 0$), implying a finite right endpoint. Our conditional modeling using the `GEV-erf` model confirms this property, with estimated values of the shape parameter ranging

between -0.386 and -0.118, as shown in panel C of Figure 4.5. We conduct a quantitative performance evaluation of the `GEV-erf` method by comparing it with the `GRF` approach of (Athey et al., 2019) and the `QRF` method of (Meinshausen and Ridgeway, 2006), using the metric proposed by (Wang and Li, 2013), defined as follows:

$$R_{n'}\left(\hat{Q}_{\cdot}(\tau)\right) = \frac{\sum_{i=1}^{n'} \mathbb{1}\{Y_i < \hat{Q}_{X_i}(\tau)\} - n'\tau}{\sqrt{n'\tau(1-\tau)}}, \tag{4.12}$$

where the function $\hat{Q}_x(\tau)$ denotes the estimated conditional quantile, evaluated on the test sample $(x_i, y_i)_{i=1,\ldots,n'}$. The predictive performance of the different models is estimated using 5-fold cross-validation. As an evaluation criterion, we use the absolute value of the metric defined in (4.12). This procedure is repeated 20 times, and the average of the resulting errors is computed

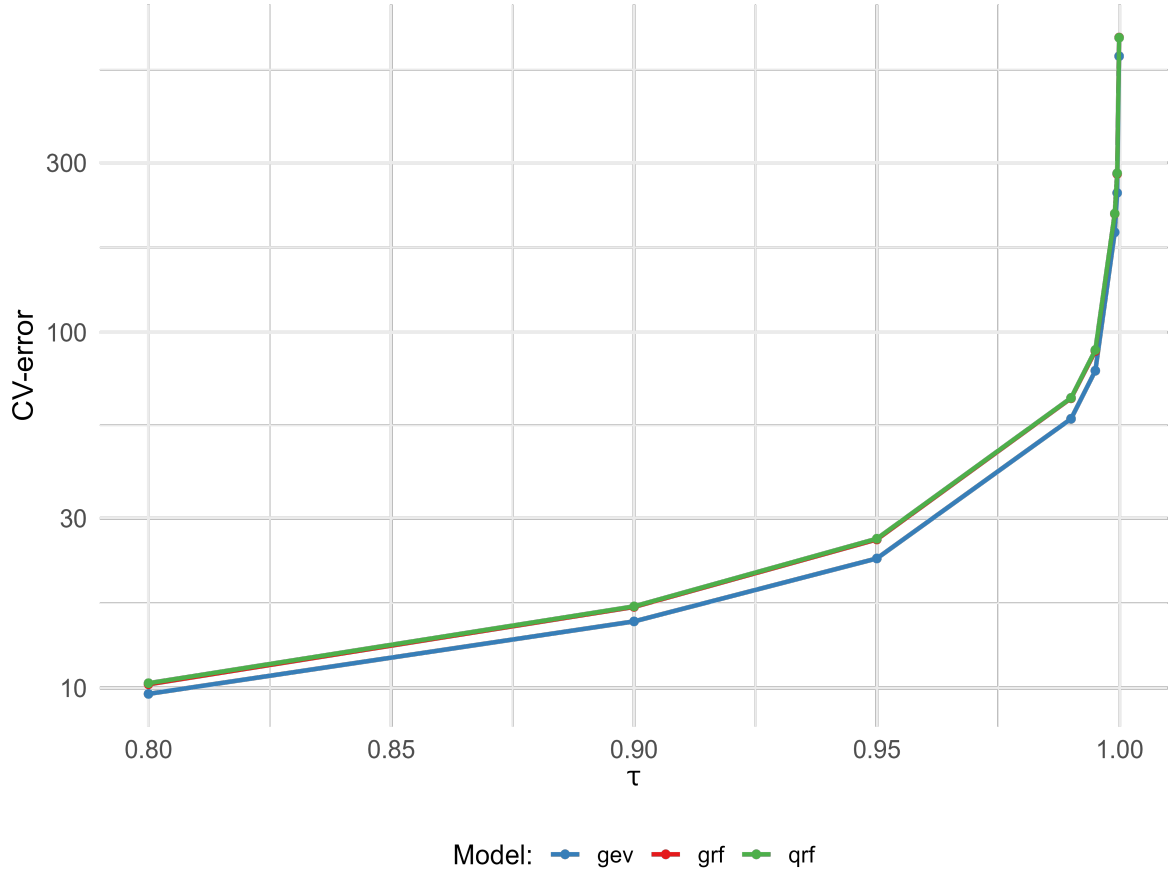

Figure 4.6: Average prediction error for the different models as a function of the extreme quantile level.

to stabilize the evaluation. Figure 4.6 illustrates the evolution of the average prediction error as a function of the probability level (on a logarithmic scale) for each of the evaluated models. The results clearly show that the `GEV-erf` method outperforms the `GRF` and `QRF` approaches, the latter two exhibiting very similar performance regardless of the extreme probability level

considered.

4.6 Conclusion

Extreme quantile regression is a powerful statistical tool that allows for the analysis and prediction of behaviours in the tails of distributions, where rare and extreme events occur. The existing literature on modeling methods for this approach is limited, particularly in the context of using the Block Maxima approach to ensure tail extrapolation. Most available methods use the Peak-Over-Threshold approach, such as (Pasche and Engelke, 2024), (Velthoen et al., 2023), (Farkas et al., 2024) and (Gnecco et al., 2024). In this work, we propose a flexible quantile regression method based on the BM approach and the generalized random forest method to address the issues encountered with classical quantile regression methods, primarily the limitation to low-dimensional covariate spaces and the potential complexity of the quantile function. Our proposed method effectively addresses these issues. Using the BM approach, we model the tail of the conditional distribution $Y|X = x$ for the generalized extreme value distribution with parameters dependent on covariates. These parameters are estimated using a penalized weighted maximum likelihood method, with weights obtained through the generalized random forest method. Simulation studies and application to real dataset show that our method better captures the complex structure of the quantile function and provides good estimates even when the characteristic variables are high-dimensional and in the presence of noise. Our method demonstrates strong performance compared to other quantile regression methods that utilize learning algorithms, such as (Meinshausen and Ridgeway, 2006) and (Athey et al., 2019) method. While this work is more application-oriented, future research should aim to theoretically prove the consistence of our method, as done by (Gnecco et al., 2024) for the POT approach. Another perspective is to explore the multidimensional case of Y (i.e., $Y \in \mathcal{Y} \subset \mathbb{R}^q$ with $q > 1$) and examine the spatial aspect of our method.

Appendix

4.A Selection of parameters λ and min.node.size

The table 4.4 shows the results of the cross-validation used to select the penalty λ and the tuning parameter *min.node.size* for the generalized random forest. To optimize the execution time of the cross-validation, we set the number of trees in the forest to $num.trees = 2000$ (the default value in the GRF method). The various λ values tested are $\{10^{-4}, 10^{-3}, 0.005, 0.01, 0.05, 0.1\}$ and those of *min.node.size* are $\{5, 10, 20, 50\}$ using 5-fold cross-validation (i.e. we take $K = 5$ in (4.11)). We select the pair $(\lambda; min.node.size)$ that minimizes the cross-validation error defined in (4.11). The table 4.4 shows the cross-validation errors for all combinations of these

hyper-parameters in the set under consideration, and for the different scenarios performed. According to this table, the smallest cross-validation error is obtained for the combination ($\lambda = 0.001$; *min.node.size* $= 10$) for scenario 1, ($\lambda = 0.001$; *min.node.size* $= 50$) for scenario 2 and ($\lambda = 0.001$; *min.node.size* $= 5$) for scenario 3.

Grid parameter for CV		**CV-Error**		
λ	**min.node.size**	**Scenario 1**	**Scenario 2**	**Scenario 3**
1e-04	5	11.53729	13.10445	10.42870
1e-03	5	11.53557	13.10410	**10.41969**
5e-03	5	11.53842	13.10315	10.42660
1e-02	5	11.53971	13.10379	10.42475
5e-02	5	11.53710	13.10341	10.42564
1e-01	5	11.54149	13.10148	10.42543
1e-04	10	11.53785	13.09767	10.42836
1e-03	10	**11.53538**	13.09863	10.42696
5e-03	10	11.53980	13.09724	10.43051
1e-02	10	11.53842	13.09817	10.42812
5e-02	10	11.53824	13.09793	10.42688
1e-01	10	11.53911	13.09822	10.42852
1e-04	20	11.53810	13.09522	10.43054
1e-03	20	11.53865	13.09498	10.42747
5e-03	20	11.53897	13.09506	10.42786
1e-02	20	11.53886	13.09487	10.43035
5e-02	20	11.53847	13.09440	10.42615
1e-01	20	11.53947	13.09515	10.42804
1e-04	50	11.54066	13.09314	10.42667
1e-03	50	11.53954	**13.09209**	10.42739
5e-03	50	11.53960	13.09295	10.42588
1e-02	50	11.54043	13.09330	10.42886
5e-02	50	11.54106	13.09272	10.42899
1e-01	50	11.54099	13.09276	10.42816

Table 4.4: Adjustment parameters with different combinations of λ and *min.node.size*.

4.B Additional simulation study

In this section, we perform additional experiments to demonstrate the performance of our method in different scenarios and for various quantile orders. Whatever the quantile order considered, our model maintains high performance across the various scenarios and covariate sizes studied, as shown in Figures 4.7, 4.8, and 4.9.

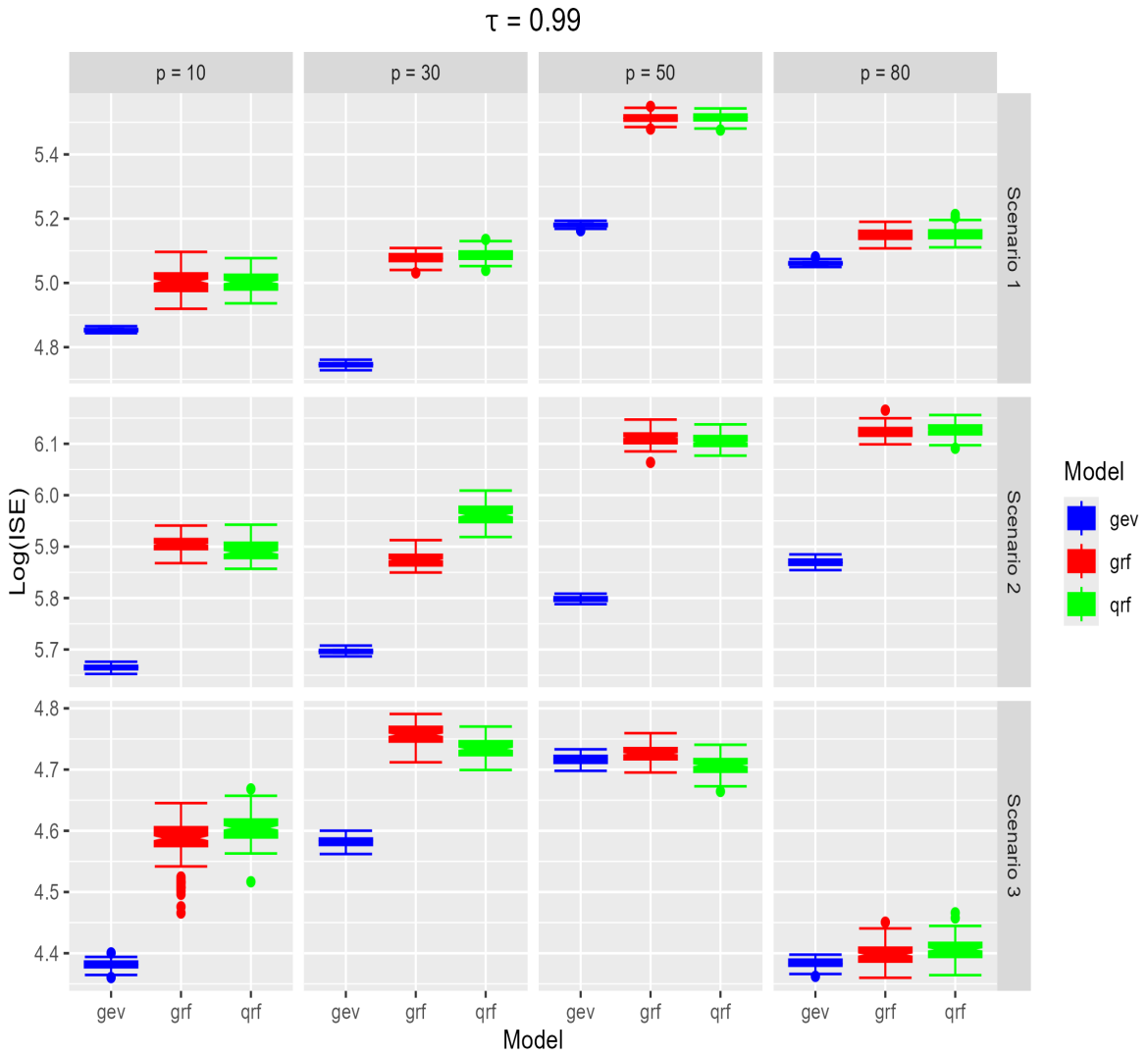

Figure 4.7: Boxplot of $log(ISE)$ over 100 replication, for $p \in \{10, 30, 50, 80\}$, $\tau = 0.99$ and different scenario.

4.C Sensitivity analysis of block size m

We study here the sensitivity of the block maxima (BM) method to the block size m, a crucial parameter for estimating extreme quantiles. An excessively large block size increases the variance of the estimators, while an insufficient block size induces bias. The literature does not provide a universal method for selecting the block size m, although several recent contributions have attempted to address this issue. For example, (Özari et al., 2019), (Özarı et al., 2018) present a computational approach illustrated by a financial case study. (Wang et al., 2016) propose a multi-criteria method combining graphical analyses and goodness-of-fit tests (Kolmogorov–Smirnov, χ^2). (Dkengne et al., 2020) develop an automatic method applied in engineering and meteorology, while (Cervantes et al., 2024) assess the fit of the GEV distribution across nine representative block sizes using QQ-plots and statistical tests (Kol-

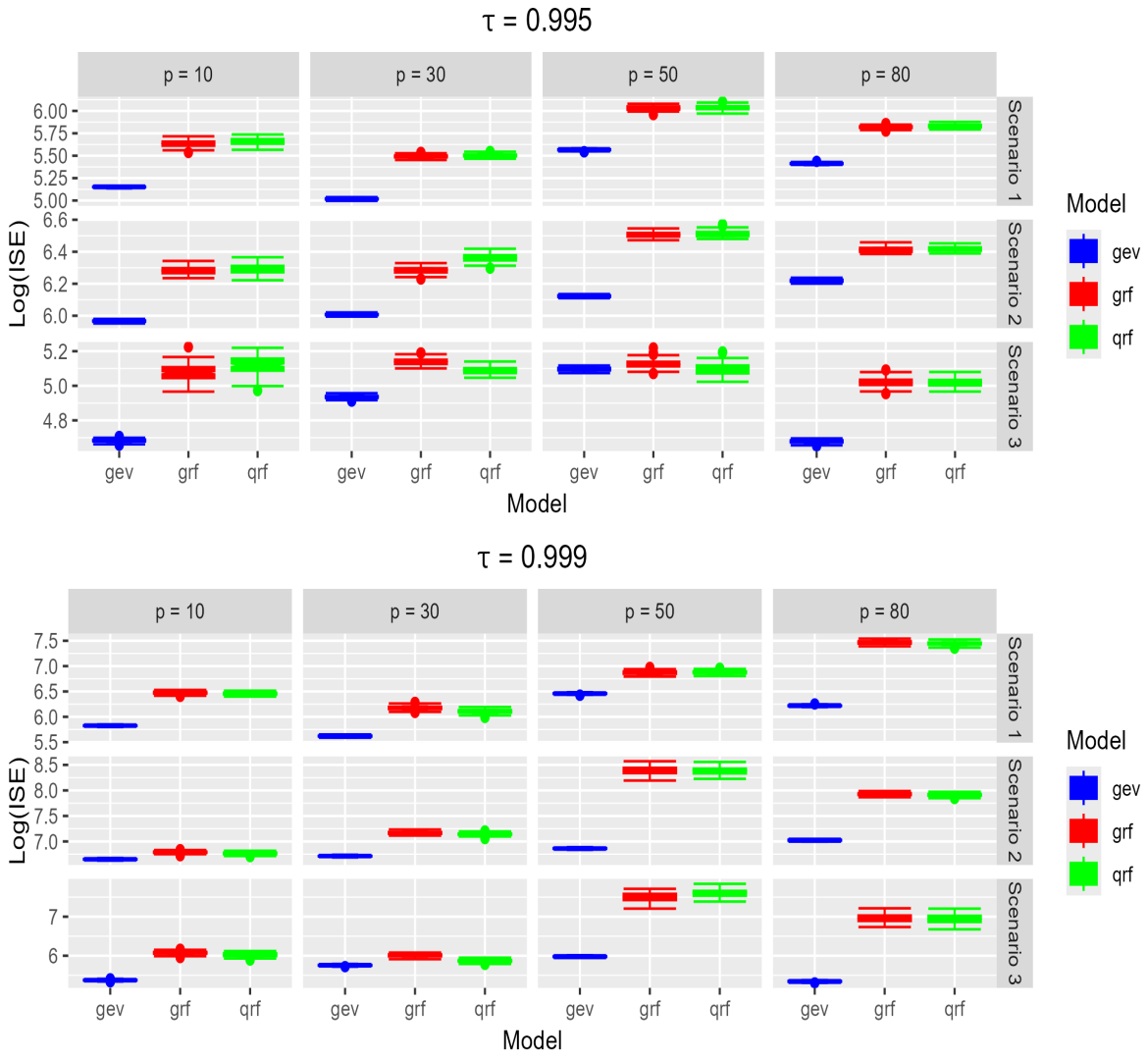

Figure 4.8: Boxplot of $log(ISE)$ over 100 replication, for $p \in \{10, 30, 50, 80\}$, $\tau \in \{0.995, 0.999\}$ and different scenario.

mogorov–Smirnov, Anderson–Darling, Cramér–von Mises).

Building on these contributions, we analyze the impact of the block size m on both the fit of the estimated GEV distribution and the estimation of the conditional quantile. For each scenario, we consider a range of block sizes from 10 to 100 in increments of 5. For each value of m, the model is trained on a dataset of size N, and the conditional quantile is then estimated on an independent test sample $\{(x_i, y_i)\}_{i=1}^{n'}$, as described in Section 4.4. The covariate dimension is fixed at $p = 40$, the regularization parameter at $\lambda = 0.001$, and the parameters *min.node.size* and *num.trees* are retained at their default values from the `grf` package (Athey et al., 2019). The GEV-ERF model is fitted to the block maxima samples corresponding to each block size. For each value of m, two steps are performed using the test sample. First, the MISE (as defined in Section 4.4) is computed for various quantile levels ($\tau \in \{0.9, 0.99, 0.999\}$), in order to identify the minimum block size $m_{\min}$ beyond which the MISE stabilizes. Second, for

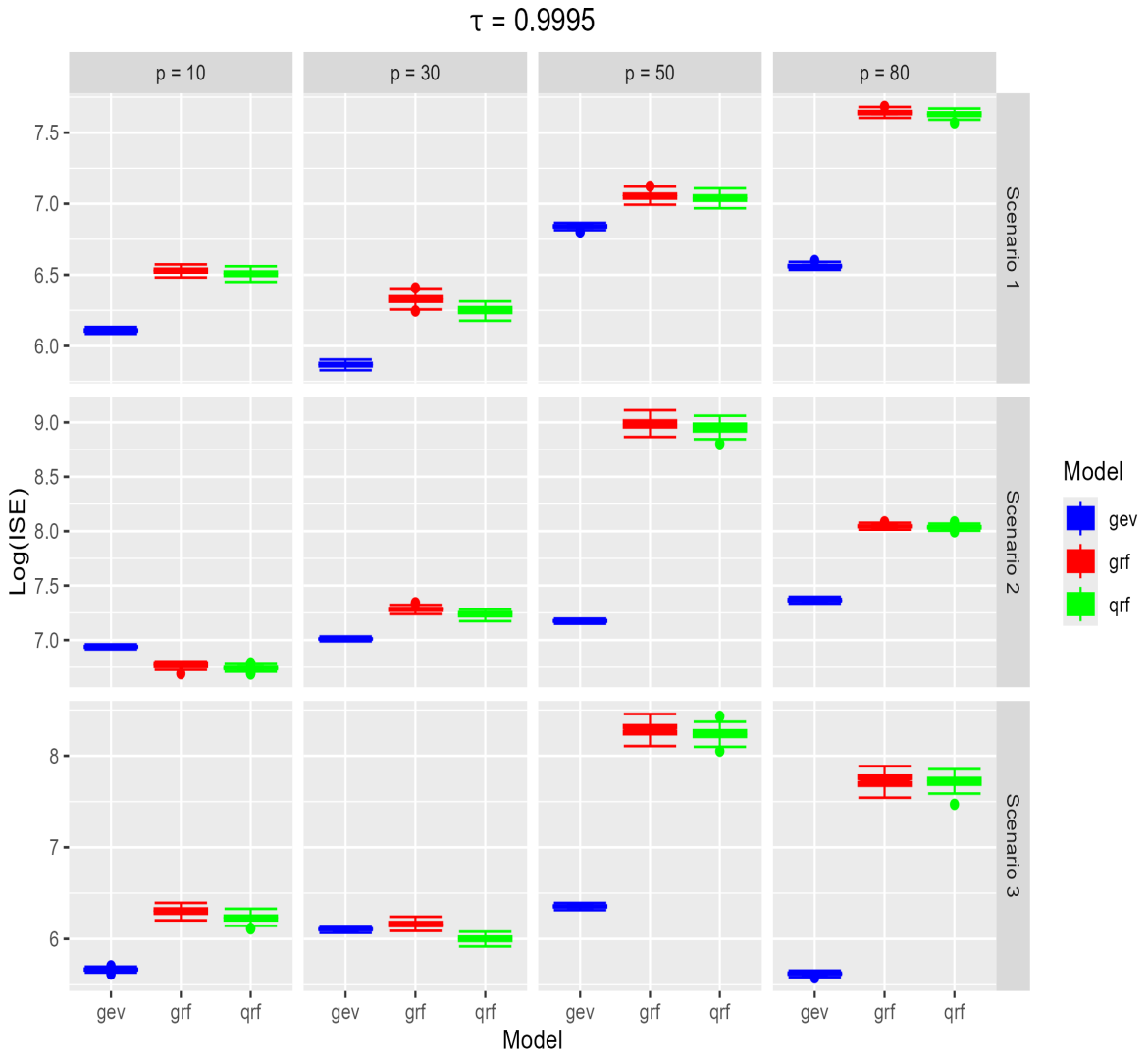

Figure 4.9: Boxplot of $log(ISE)$ over 100 replication, for $p \in \{10, 30, 50, 80\}$, $\tau = 0.9995$ and different scenario.

all $m \geq m_{\min}$, we assess the goodness-of-fit of the conditional GEV distribution estimated by the GEV-ERF method using three statistical tests: Kolmogorov–Smirnov (KS), Cramér–von Mises (CVM), and Anderson–Darling (AD). These tests are applied to the probability integral transform of the estimated distribution. More precisely, for each $i \in \{1, \ldots, n'\}$, we compute $u_i = G_{\theta(X_i)}(z_i)$, where z_i is the block maximum from the test sample, and test whether the u_i values follow a uniform distribution.

The goodness-of-fit tests are performed using the functions `ad.test` (AD test) and `cvm.test` (CVM test) from the `goftest` package, and `ks.test` (KS test) from the `stats` package. Lower values of the test statistics indicate a better fit between the model and the data. Figure 4.10 shows the evolution of the logarithm of the MISE as a function of block size for various quantile levels across the scenarios. The results indicate that while increasing block size tends to raise the MISE of the estimated quantiles, this increase becomes less variable be-

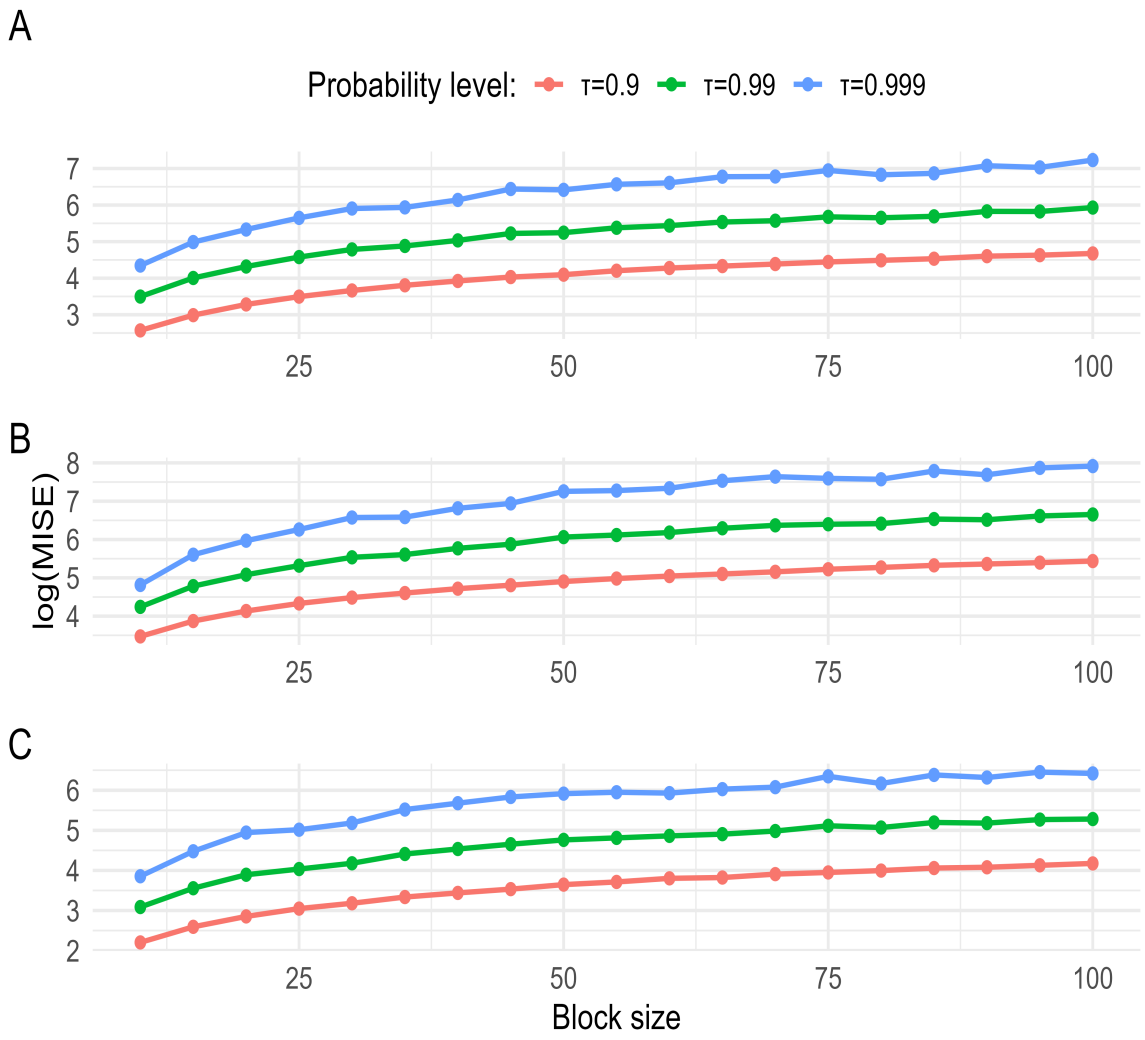

Figure 4.10: Log(MISE) of Estimated Conditional Quantiles vs. Block Size for scanrio 1 (A), scenario 2 (B) and scenario 3 (C)

yond $m_{\min} = 30$ in Scenarios 1 and 2, and $m_{\min} = 25$ in Scenario 3. This trend can be explained by the fact that smaller blocks provide more observations for training, thereby improving the performance of the estimation method.

Table 4.5 presents the KS, AD, and CVM test statistics for block sizes $m \geq m_{\min}$. The smallest values for the KS and CVM statistics occur around $m = 45$ in Scenarios 1 (0.061 and 0.172, respectively) and 2. In Scenario 3, the minimum values across all three tests are achieved at $m = 25$, indicating a good model fit at that block size. These findings confirm that block size significantly influences the quality of fit of the GEV-ERF model. The results highlight the method's sensitivity to this parameter and suggest that block sizes in the range of 25 to 50 generally provide a good trade-off between bias and variance, ensuring a satisfactory fit between the estimated GEV distribution and the block maxima. The choice of block size remains a central challenge in extreme value analysis. In practice, it is common to select

Block size	**Scenario 1**			**Scenario 2**			**Scenario 3**		
	KS Stat	AD stat	CVM stat	KS stat	AD stat	CVM stat	KS stat	AD stat	CVM
25							0,053	1,572	0,234
30	0,072	1,468	0,206	0,046	0,453	0,074	0,080	1,950	0,350
35	0,071	1,290	0,203	0,071	0,950	0,180	0,101	3,845	0,728
40	0,077	1,975	0,296	0,055	0,561	0,088	0,111	3,717	0,704
45	0,061	1,308	0,172	0,036	0,237	0,028	0,119	4,659	0,914
50	0,073	1,571	0,232	0,059	0,468	0,083	0,109	3,523	0,648
55	0,085	1,753	0,245	0,069	0,652	0,103	0,120	3,243	0,573
60	0,090	1,492	0,242	0,055	0,324	0,035	0,103	2,672	0,491
65	0,095	1,686	0,300	0,064	0,451	0,062	0,139	4,437	0,848
70	0,090	1,372	0,223	0,056	0,272	0,031	0,103	1,851	0,307
75	0,102	1,478	0,256	0,047	0,285	0,030	0,097	2,171	0,368
80	0,140	2,296	0,431	0,091	0,980	0,159	0,112	2,876	0,463
85	0,121	2,074	0,385	0,142	2,956	0,442	0,138	3,946	0,735
90	0,102	1,210	0,206	0,058	0,309	0,025	0,128	2,525	0,458
95	0,112	1,647	0,259	0,058	0,281	0,023	0,149	4,444	0,849
100	0,132	2,118	0,388	0,113	1,380	0,176	0,124	1,773	0,306

Table 4.5: Statistical values for the various tests as a function of block size (verifying $m \geq m_{min}$) and scenario.

block sizes based on natural time units—such as a year, season, or month—depending on the temporal resolution of the data.

CHAPTER CINQ

PENALIZED ESTIMATION OF GEV PARAMETERS FOR EXTREME QUANTILE REGRESSION

Les résultats présentés dans ce chapitre ont fait l'objet d'un article de recherche qui a été accepté pour publication.
Vidagbandji, L. M., Berred, A., Bertelle, C., & Amanton, L. (2026). Penalized estimation of GEV parameters for extreme quantile regression. *Accepted for publication in Journal of Statistical Theory and Practice.*

Contents

5.1 Introduction

The study of extreme phenomena has become an unavoidable necessity in the current context, where their impacts are increasingly concerning. Whether related to climate change (Arnell, 1988), financial crises, or technological failures, extreme events—though rare by definition—often lead to severe and disproportionate consequences. Positioned in the extreme tails of statistical distributions, these events exhibit unique characteristics that cannot be captured by average or typical behaviors. Their modeling and understanding, therefore, require specialized tools capable of grasping the complexity and dynamics inherent to these phenomena. Classical statistical methods, which primarily focus on analyzing the central values of distributions, provide powerful frameworks for examining average trends and central relationships between variables. However, they prove inadequate when it comes to exploring or predicting behaviors in the extreme regions of distributions. The tails of distributions, where extreme values reside, are often underrepresented in standard models due to the low data density and the difficulty of capturing the complex relationships that prevail in these regions. In response to these limitations, new methodologies have emerged to address the specific needs of extreme event analysis. Among them, extreme quantile regression stands out for its relevance and constitutes the main focus of this work.

The quantile regression model, as introduced by (Koenker and Bassett, 1978), aims to estimate a conditional quantile from a sample $(X_1,Y_1),\ldots,(X_n,Y_n)$, representing independent copies of the random vector (X,Y), where $X \in \mathcal{X} \subset \mathbb{R}^p$ and $Y \in \mathcal{Y} \subset \mathbb{R}$. This model allows for the estimation of the conditional quantile $Q_x(\tau)$, defined as

$$Q_x(\tau) = F^{-1}_{Y|X=x}(\tau), \tag{5.1}$$

where $x \in \mathcal{X}$, $\tau \in (0,1)$, and $F^{-1}_{Y|X=x}$ represents the generalized inverse of the conditional cumulative distribution function of Y, given $X = x$. This model has revolutionized the statistical approach by providing a powerful alternative to classical regression (Beyerlein, 2014b). Unlike the latter, which is limited to estimating the conditional mean $E(Y|X = x)$, quantile regression allows for the estimation of any conditional quantile. This flexibility enables a deeper understanding of heterogeneous relationships between variables, highlighting dynamics that traditional methods fail to capture. Numerous quantile regression models have been proposed in the literature, including those by (Angrist et al., 2006), (Koenker, 2017), (Benziadi et al., 2016), (Cade and Noon, 2003), and (Wang and Tsai, 2009). These approaches are effective for moderate quantiles, where sufficient observations are available in the sample. However, when focusing on extreme quantiles, located in the tails of the distribution, new challenges arise.

Classical quantile regression methods perform well when $\tau_n \to 1$ and $n(1-\tau_n) \to \infty$ as $n \to \infty$. In contrast, when $\tau_n \to 1$ and $n(1-\tau_n) \to d \in (0,+\infty)$, estimation requires extrapolation

into the tails. This is problematic because observations are scarce in these regions and the statistical properties of extremes differ fundamentally from those in the central part of the distribution. Extreme quantile regression, leveraging advancements from extreme value theory (EVT) and incorporating robust estimation techniques, emerges as an appropriate response to these challenges. Several studies have explored the integration of extreme value theory in this framework, including those by (Chernozhukov, 2005), (Zheng et al., 2017), (Zhu et al., 2022), (Schaumburg, 2012), (Saulo et al., 2022), (Chernozhukov et al., 2020), and (Kithinji et al., 2021). A more detailed presentation of extreme quantile regression approaches can be found in the work of (Chernozhukov et al., 2017).

Even when sufficient data is available, classical regression methods can encounter difficulties if the conditional quantile function is nonlinear or heterogeneous. In such cases, overly simplistic models risk introducing bias (Gnecco et al., 2024). To address this issue, several models combining statistical learning methods have been proposed. Among the notable approaches are those based on neural networks, such as (Cannon, 2018), as well as methods relying on decision trees and random forests, including those proposed by (Chaudhuri and Loh, 2002), (Meinshausen and Ridgeway, 2006), (Buchinsky, 1998), and (Athey et al., 2019). Additionally, studies utilizing other machine learning techniques, such as those by (**?**), (Yao et al., 2022), and (Tyralis et al., 2019), also provide significant contributions.

In high-dimensional settings, identifying neighbors close to x becomes problematic, and kernel or nearest-neighbor methods suffer from the curse of dimensionality. Recently, several approaches have combined EVT with statistical learning to simultaneously address extrapolation in the tails, complexity of quantile fonction, and high dimensionality of covariate space. Examples include generalized additive models proposed by (Youngman, 2019), gradient boosting introduce by (Velthoen et al., 2023), neural networks by (Pasche and Engelke, 2023), and generalized random forests proposed by (Gnecco et al., 2024) and (Vidagbandji et al., 2025). To enable extrapolation in the distribution tails, these authors use the generalized Pareto distribution (GPD) within the Peak-over-Threshold (POT) approach of extreme value theory. Specifically, these models approximate the conditional distribution $Y|X = x$ in expression (5.1) with the GPD, where the parameters depend on the covariate x, and thus obtain the extreme conditional quantile using various techniques for estimating the GPD parameters.

In this paper, we adopt the block maxima approach from EVT (see (De Haan and Ferreira, 2006), (Coles, 2001) for a detailed presentation), which models block maxima using the generalized extreme value (GEV) distribution. We propose an extreme quantile regression method where the GEV parameters (μ, σ, ξ) depend on covariates $x \in \mathcal{X}$. These parameters are estimated through a weighted version of the penalized maximum likelihood estimator, originally introduced by (Coles and Dixon, 1999), where the weights are derived from generalized

random forests (Athey et al., 2019). The generalized extreme value distribution is given by

$$G_{(\mu,\sigma,\xi)}(z) = \begin{cases} \exp\left(-(1+\xi\frac{z-\mu}{\sigma})_+^{-\frac{1}{\xi}}\right), & \text{if } \xi \neq 0, \\ \exp\left(-\exp\left(-\frac{z-\mu}{\sigma}\right)\right), & \text{if } \xi = 0, \end{cases} \tag{5.2}$$

defined on $\{z \in \mathbb{R} : 1+\xi\frac{z-\mu}{\sigma} > 0\}$, where $\mu \in \mathbb{R}$ is the location parameter, $\sigma \in \mathbb{R}_+^*$ is the scale parameter, and $\xi \in \mathbb{R}$ is the extreme value index. Inspired by (Gnecco et al., 2024), our approach simultaneously addresses three challenges: extrapolation in the tails through the GEV distribution, complex quantile structures via adaptive similarity weights, and high-dimensional predictors space through the partitioning properties of grf. Finally, the penalty function employed in this work ensures the asymptotic optimality of the maximum likelihood estimator in large samples and improves estimation accuracy in small-sample settings.

This article is structured as follows. Section 5.2 introduces the fundamental concepts related to extreme quantile regression, the block maxima approach that we will use in our method, and the technique of generalized random forests. Section 5.3 is dedicated to a detailed description of our methodology, with a particular focus on estimation and validation techniques suited to the tails of distributions. Finally, in Sections 5.4 and 5.5, we evaluate the effectiveness of the proposed method through simulation studies and an empirical application using salary census data from the United States for the year 1980 (Angrist et al., 2006).

5.2 Extreme quantile regression

The literature on extreme quantile estimation relies on the asymptotic results of extreme value theory, which allow for extrapolation beyond the range of observed data (De Haan and Ferreira, 2006; **?**). In the case where no covariate is available, consider a sample $(Y_1, \ldots, Y_n)$ consisting of independent and identically distributed realizations of a random variable Y with a cumulative distribution function F. The goal is to estimate the quantile $Q(\tau) = F^{-1}(\tau)$ for an extreme probability level $\tau \in (0,1)$. When this level depends on the sample size n, let $\tau = \tau_n$. A probability level is considered extreme if $\tau_n \to 1$ and $n(1-\tau_n) \to d \geq 0$ as $n \to +\infty$. In this case, the number of observations exceeding the quantile $Q(\tau_n)$ becomes limited as n increases, making empirical estimation particularly difficult. Estimating these extreme quantiles requires extrapolation beyond the range of available data (De Haan and Ferreira, 2006), which necessitates robust asymptotic theory to make these extrapolations accurately and reliably. Let τ_0 denote the intermediate probability level from which the quantile order becomes extreme.

One of the main results concerning extrapolation is the asymptotic theorem by (Fisher and Tippett, 1928) and (Gnedenko, 1943), which establishes the limiting distribution of the maximum of a sequence of independent and identically distributed random variables. These

authors demonstrated that if there exist sequences $a_n > 0$ and $b_n \in \mathbb{R}$ such that

$$F^n(b_n + a_n x) \to G(x), \quad as \quad n \to \infty, \tag{5.3}$$

where F^n denotes the n-th power of the cumulative distribution function of the random variable Y, then G is a non-degenerate distribution given by

$$G_\xi(x) = \exp\left(-(1+\xi x)^{-\frac{1}{\xi}}\right),$$

for all x such that $1+\xi x > 0$ and $\xi \in \mathbb{R}$. The case $\xi = 0$ corresponds to the limit of $G_\xi(x)$ as $\xi \to 0$, and is given by $G_0(x) = \exp(-e^{-x})$. A cumulative distribution function F is said to belong to the domain of attraction of an extreme value distribution, denoted $F \in \mathcal{D}(G_\xi)$, when there exist sequences $a_n > 0$ and $b_n \in \mathbb{R}$ such that the equality (5.3) is satisfied. Depending on the sign of ξ, three domains of attraction are distinguished: the Fréchet domain of attraction (when $\xi > 0$), the Gumbel domain of attraction (when $\xi = 0$), and the Weibull domain of attraction (when $\xi < 0$). The shape parameter ξ plays a key role in characterizing the behavior of the upper tail of the distribution. The Weibull class is characterized by a finite upper bound, while the Gumbel and Fréchet classes feature infinite tails with distinct rates of decay. More specifically, the Fréchet distribution, which has heavy tails, decays polynomially, while the Gumbel class decays exponentially. This difference in tail behavior is crucial, as it reflects structural differences in the modeling of extreme values. The unification of these three classes into a generalized extreme value distribution, whose explicit form is given by equation (5.2), offers the major advantage of allowing the data to determine the most appropriate family, thanks to the estimation of the ξ parameter using inference methods. Thus, the choice of tail behavior is made without the need for prior decisions. Additionally, the uncertainty surrounding the estimation of ξ allows for an assessment of the relative relevance of the three types of distributions for their application to the available data (De Paola et al., 2018). In practice, special attention is given to the estimation of this parameter.

In this work, the GEV distribution is used to facilitate extrapolation in the tail of the distribution and thus the estimation of extreme quantiles. By inverting this distribution, the quantile corresponding to an extreme probability level $\tau_n > \tau_0$ is obtained, expressed by

$$Q(\tau) = \begin{cases} \mu + \dfrac{\sigma}{\xi}\left((-\ln(\tau))^{-\xi} - 1\right) & \text{if } \xi \neq 0, \\ \mu - \sigma \ln(-\ln(\tau)) & \text{if } \xi = 0. \end{cases} \tag{5.4}$$

The estimation of the extreme quantile involves determining the parameters μ, σ, and ξ. Many works have focused on estimating these parameters while integrating statistical learning methods capable of addressing the issues raised in the introduction, namely the high dimensionality of the predictor space and the complexity of the relationships between the response variable

and the explanatory variables. Within the framework of the Peak-over-Threshold (POT) approach in extreme value theory, (Velthoen et al., 2023) proposed a method based on gradient boosting to estimate the parameters of the conditional generalized Pareto distribution (GPD), thereby providing an estimate of the extreme conditional quantile. Meanwhile, (Pasche and Engelke, 2023) developed a method using neural networks to estimate these conditional quantiles. In a similar approach, (Gnecco et al., 2024) introduced a method for estimating the parameters of the conditional GPD via a weighted version of the maximum likelihood estimator, with the weights determined by the generalized random forest method. Although the maximum likelihood method is widely used and presents interesting asymptotic properties ((Bücher and Segers, 2017), (Dombry, 2015), (Dombry and Ferreira, 2019)), it is often criticized for its limited performance in the case of small samples. This weakness is a major obstacle in practice, where the analysis of extreme events often relies on a limited amount of data. Indeed, the rarity of extreme events means that, even over long periods of observation, the data available for fitting an extreme value model can remain scarce. To address this issue, (Coles and Dixon, 1999) proposed a penalized maximum likelihood estimator for the parameters of the GEV distribution. This approach retains the asymptotic properties of the classical maximum likelihood estimator (MLE) while improving its robustness and performance in the presence of small samples. The likelihood function associated with the independent block maxima $z_1, \ldots, z_n$ is given by

$$L(\mu, \sigma, \xi) = \prod_{i=1}^{n} \frac{dG(z_i; \mu, \sigma, \xi)}{dz_i}, \tag{5.5}$$

and the maximum likelihood estimator of the parameters θ of the GEV distribution is given by

$$\hat{\theta} = \arg\max_{\theta \in \Theta} L(\mu, \sigma, \xi),$$

with $L(\mu, \sigma, \xi)$ given in equation (5.5) and $\Theta \subset \mathbb{R} \times \mathbb{R}_+^* \times \mathbb{R}$. To address the limitations of the maximum likelihood method with small samples, our approach incorporates the penalized likelihood function proposed by (Coles and Dixon, 1999), as detailed in Section 5.3.2. Specifically, we develop an estimation procedure for the parameters of the conditional GEV distribution based on a weighted extension of the estimator introduced by (Coles and Dixon, 1999), where the weights are derived from the generalized random forest method described below.

Generalized Random Forest

The generalized random forest (grf), introduced by (Athey et al., 2019), extends the classical random forest method (Breiman, 2001). Like traditional random forests, grf aggregates predictions from B decision trees, each grown on a bootstrap sample and built using random feature selection at each split, which promotes diversity and improves generalization. The key novelty

of grf is the flexibility to adapt the loss function guiding tree construction, making it suitable for a wide range of estimation tasks, such as conditional means, conditional quantiles (Athey et al., 2019), or extremal quantiles (see (Gnecco et al., 2024) and (Vidagbandji et al., 2025)). Let $\{(X_i, Y_i)\}_{i=1,\cdots,n}$ be the training data. For a test point $x \in \mathcal{X}$, each tree provides an estimate of the conditional mean: $\eta_b(x) = \sum_{i=1}^{n} \frac{\mathbb{1}_{\{X_i \in R_b(x)\}}}{|\{j:X_j \in R_b(x)\}|} Y_i$, where $R_b(x)$ denotes the region (or leaf) of the b-th tree containing x. Equivalently, $R_b(x)$ corresponds to the set of training observations X_i that fall in the same terminal node as x. The forest prediction $\eta(x)$ is obtained by taking the average of all the predictions from the B trees

$$\eta(x) = \frac{1}{B} \sum_{b=1}^{B} \eta_b(x) = \sum_{i=1}^{n} w_n(x, X_i) Y_i,$$

with similarity weights defined as

$$w_n(x, X_i) = \frac{1}{B} \sum_{b=1}^{B} \frac{\mathbb{1}_{\{X_i \in R_b(x)\}}}{|\{j : X_j \in R_b(x)\}|}, \tag{5.6}$$

where $|E|$ denotes the cardinality of the set E. The weights $w_n(x, X_i)$ quantify the contribution of observation X_i to the prediction at x. They can be understood as the normalized frequency with which X_i and x fall in the same leaf across the ensemble of trees. Thus, $w_n(x, X_i)$ defines an adaptive similarity measure between x and X_i, reflecting how strongly X_i influences the prediction at x. This provides an intuitive interpretation: $R_b(x)$ identifies the local neighborhood of x within a tree, and $w_n(x, X_i)$ aggregates these neighborhoods across the forest.

In standard random forests, the similarity weights implicitly favor observations X_i with $\mathbb{E}(Y|X = X_i) \approx \mathbb{E}(Y|X = x)$. However, this mechanism may fail to capture heterogeneity in quantile functions: an observation X_i may have a low weight even if $\mathbb{Q}_{Y|X=X_i}(\tau) \approx \mathbb{Q}_{Y|X=x}(\tau)$ (see (Athey et al., 2019)). By contrast, grf incorporates a loss function tailored to quantiles, ensuring that the similarity weights emphasize observations relevant for conditional quantile estimation. Moreover, grf-based weights adaptively partition the covariate space in a way that highlights the most informative variables. This is particularly advantageous in high-dimensional settings: unlike kernel or nearest-neighbor methods, which are severely affected by the curse of dimensionality, grf constructs data-driven neighborhoods around x that capture complex and possibly nonlinear dependencies between the response and the covariates. Therefore, weights based on the grf are a robust and flexible tool for approximating conditional quantiles, even in heterogeneous and high-dimensional environments. They play a central role in the methodology we propose. In what follows, these weights will be denoted by $w_i(x)$ for all $x \in \mathcal{X}$ and $i = 1, \ldots, n$.

5.3 Model and inference procedure

In this section, we present in detail the procedure for estimating the parameters of the conditional GEV distribution, as well as our flexible method for estimating the conditional quantiles associated with extreme probability levels, in order to address the challenges discussed in section 5.2.

5.3.1 Setup for extreme quantile regression

Let $(X_1,Y_1),(X_2,Y_2),\ldots$ be a sequence of independent and identically distributed random vectors, coming from the random vector (X,Y). Suppose that the conditional distribution function of Y given $X=x$, denoted $F_x(\cdot)$, belongs to the domain of attraction of the extreme value distribution, i.e., $F_x \in \mathcal{D}\left(G_{\xi(x)}\right)$, with the corresponding normalizing sequences $a_m(x)$ and $b_m(x)$ defined in (5.3). We divide the sequence of vectors into n blocks of size m, such that the k-th block, for $k \geq 1$, is given by

$$B_{k,m} = \{(X_{(k-1)m+1}, Y_{(k-1)m+1}), \ldots, (X_{km}, Y_{km})\}.$$

Let $Z_{k,m} = \max\{Y_i : (X_i,Y_i) \in B_{k,m}\}$ and $X_{k,m}$ be the X_i associated with the Y_i maximizing $\{Y_i : (X_i,Y_i) \in B_{k,m}\}$. Thus, for any $m > 1$, the sequence $\{(X_{k,m}, Z_{k,m})\}_{k\geq 1}$ represents the block maxima, where the maximum is taken with respect to the Y-component. For any $x \in \mathcal{X}$ and $m \geq 1$, the variables $\{Z_{k,m}|X_{k,m} = x\}_{k\geq 1}$ are independent and identically distributed with a distribution function F_x^m. Thus, for any $x \in \mathcal{X}$ and $m > 1$, we have

$$\frac{Z_{k,m} - b_m(x)}{a_m(x)}\Big|X_{k,m} = x \xrightarrow{d} G_{\xi(x)} \text{ as } m \to +\infty,$$

with $G_{\xi(x)}(z) = \exp\left(-(1+\xi(x)z)^{-\frac{1}{\xi(x)}}\right)$. This allows us to approximate, for all $x \in \mathcal{X}$, the distribution of $Z_{k,m}|X_{k,m} = x$ by the conditional GEV distribution, where the parameters depend on the covariate x. More specifically, the location, scale, and shape parameters are functions of x, defined respectively by $\mu(\cdot) : \mathcal{X} \to \mathbb{R}$, $\sigma(\cdot) : \mathcal{X} \to \mathbb{R}_+^*$, and $\xi(\cdot) : \mathcal{X} \to \mathbb{R}$. The conditional GEV distribution is given by

$$G_{\theta(x)}(z) = \exp\left[-\left(1+\xi(x)\frac{z-\mu(x)}{\sigma(x)}\right)_+^{-\frac{1}{\xi(x)}}\right], \tag{5.7}$$

with $\theta(x) = (\mu(x), \sigma(x), \xi(x))$ for all $x \in \mathcal{X}$. The case $\xi(x) = 0$ corresponds to the limit of the expression given in (5.7) as $\xi(x) \to 0$, and is given by

$$G_{(\mu(x),\sigma(x),0)}(z) = \exp\left[-\exp\left(-\frac{z-\mu(x)}{\sigma(x)}\right)\right].$$

The quantile of order τ (with τ close to 1) of the conditional GEV distribution is obtained by inverting the distribution given in (5.7). It is given, for all $x \in \mathcal{X}$, by

$$Q_x(\tau) = \begin{cases} \mu(x) + \dfrac{\sigma(x)}{\xi(x)} \left((-\ln(\tau^m))^{-\xi(x)} - 1 \right) & \text{if } \xi(x) \neq 0, \\ \mu(x) - \sigma(x) \ln\left(-\ln(\tau^m)\right) & \text{if } \xi(x) = 0. \end{cases} \tag{5.8}$$

The conditional quantile depends on the three parameters of the conditional GEV distribution, namely $\mu(x)$, $\sigma(x)$, and $\xi(x)$ for each $x \in \mathcal{X}$. Therefore, its estimation requires the prior estimation of these parameters, which we propose to obtain initially, before substituting them into equation (5.8) to calculate an estimate of $Q_x(\tau)$. There are various methods to estimate these parameters, among which the maximum likelihood estimator is recognized for its flexibility and effectiveness in modeling extremes. However, this estimator performs worse than an alternative method based on probability-weighted moments (PWM) when applied to small samples. According to (Coles and Dixon, 1999), the superiority of the PWM method for small samples can be attributed to the assumption of a restricted parameter space, corresponding to finite population moments. In order to incorporate similar information into a likelihood-based approach, the authors propose a penalized maximum likelihood estimator, wich will be presented in section 5.3.2. This estimator retains the flexibility of modeling and the asymptotic optimality of maximum likelihood, while improving its performance in the presence of small samples. The weighting procedure for this estimator, proposed in this work, which addresses the issues stated in the introduction, is detailed in the following section.

5.3.2 Penalized weighted likelihood estimator

As described in section 5.2, the estimation of the conditional quantile presents two major challenges: the first is encountered when estimating quantiles for high probability levels (τ close to 1), and the second arises when the quantile function is complex and the dimension of the predictor space is high. Our methodology addresses both of these challenges simultaneously. To facilitate extrapolation in the tail of the response variable Y conditional on $X = x$, we rely on the asymptotic results of extreme value theory. More specifically, we model block maxima using the conditional GEV distribution, as defined in section 5.3.1. Furthermore, to capture the complex structure of the quantile function and ease the estimation in a high-dimensional covariate space, we use the weights $w_i(x)$, which are estimated using the method of generalized random forests, detailed in section 5.2. The limited performance of the maximum likelihood estimator for estimating extreme quantiles in small samples can be attributed to the marked positive skewness in the distribution of the estimated parameter ξ, which amplifies errors due to the nonlinear relationship between the quantile $Q_x(\tau)$ and ξ, leading to significant biases. In contrast, the estimator based on probability-weighted moments (PWM) a priori assumes that

$\xi < 1$, which limits the variation of ξ and generates a moderate negative bias, less severely penalizing errors (Coles and Dixon, 1999). This bias-variance trade-off allows PWM to provide more reliable estimates of extreme quantiles.

Since the maximum likelihood estimator may yield unstable results with small samples and produce values of the shape parameter outside the admissible range, (Coles and Dixon, 1999) proposed the use of a penalty function. This function incorporates into the likelihood the information that $\xi < 1$, while making values close to 1 less probable than smaller values. In this way, it imposes a coherent restriction on the parameter space, analogous to that used for the PWM estimator, thereby improving the estimation of GEV parameters in small-sample settings. The penalization acts as weak prior information, ensuring consistent estimates without compromising the main advantages of likelihood-based inference. The proposed penalty function is defined as follows

$$P_{\alpha,\lambda}(\xi) = \begin{cases} 1 & \text{if } \xi \leq 0, \\ \exp\left\{-\lambda\left(\frac{1}{1-\xi} - 1\right)^{\alpha}\right\} & \text{if } 0 < \xi < 1, \\ 0 & \text{if } \xi \geq 1, \end{cases} \tag{5.9}$$

where $\alpha \geq 0$ and $\lambda \geq 0$ are tuning parameters. Larger values of α impose a stronger relative penalty on values of ξ close to 1, while λ controls the overall weighting attached to the penalty. To account for heterogeneous data structures, we propose an adaptive procedure to select the optimal values of the tuning parameters, based on the cross-validation method described in Appendix 5.A. According to (Coles and Dixon, 1999), the combination $\lambda = \alpha = 1$ provides good performance for a wide range of ξ values and sample sizes. Although our adaptive selection increases computational cost, it yields more stable and flexible performance in our applications, while also capturing heterogeneous structures in the data. As illustrated in Figure 5.4, the optimal parameter values differ from the default values. However, it is important to note that, across the considered parameter grid, the variation in the cross-validation error is small, and therefore does not completely contradict (Coles and Dixon, 1999) recommendation. These experiments thus confirm that the results remain robust to moderate changes in the tuning parameters. The resulting penalized maximum likelihood estimator (PMLE) provides more stable estimates of ξ and, consequently, of extreme quantiles, enhancing the reliability of inference in extreme value analysis. For $\xi < 0$, the PMLE is practically indistinguishable from the classical MLE. In contrast, for $\xi \geq 0$, its behavior resembles that of the probability-weighted moments estimator, exhibiting lower variance at the cost of a slight negative bias (see (Coles and Dixon, 1999)). Overall, in terms of the bias–variance trade-off, the PMLE performs at least as well as the probability-weighted moments estimator.

The maximum penalized likelihood estimator retains all the asymptotic properties of the classical estimator, and that its performance in the presence of small samples is comparable to

that of the PWM estimator. We incorporate this approach in the estimation of the parameters of the conditional GEV distribution in the context of this work. We define the estimator $\hat{\theta}(x)$ for the parameters of the conditional GEV, as the parameter $\theta(x)$ that minimizes the weighted and penalized (negative) log-likelihood, given by

$$L_n^{pen}(\theta;x) = \sum_{i=1}^{n} w_i(x)\ell_\theta(z_i) + \log(P_{\alpha,\lambda}(\xi)), \quad \forall x \in \mathcal{X}, \tag{5.10}$$

where $\ell_\theta(z_i) = -\log\left(\frac{dG(z_i;\mu,\sigma,\xi)}{dz_i}\right)$ and $P_{\alpha,\lambda}(\xi)$ is the penalty function defined in (5.9). The weights $w_i(x)$ for $i = 1,\ldots,n$ are obtained through the grf method, which is described in section 5.2. The theoretical guarantees of the maximum likelihood estimator for the GEV distribution are well documented in the literature. The existence and convergence of the estimator are proved by (Dombry, 2015) and (Bücher and Segers, 2017) for $\xi > -1$, and the asymptotic normality is demonstrated by (Dombry and Ferreira, 2019) for $\xi > -\frac{1}{2}$. Regarding the penalized estimator, (Coles and Dixon, 1999) have shown that the penalized maximum likelihood retains the asymptotic properties of the classical MLE, since the penalty only acts as weak prior information when ξ close to 1. In our framework, we extend this estimator to the weighted setting, where the weights are obtained from generalized random forests. Asymptotic guarantees for grf-based weights have been established by (Athey et al., 2019), who proved consistency and asymptotic normality for localized parameter estimates. Combining these results with the properties of the penalized MLE, we expect the weighted penalized estimator to inherit the same asymptotic guarantees under regularity conditions. A complete theoretical analysis is beyond the scope of this paper, but we acknowledge the importance of this point and plan to address it in future work.

The likelihood function of the GEV distribution does not have a global maximum, but, by abuse of language, we will refer to $(\hat{\mu}(x), \hat{\sigma}(x), \hat{\xi}(x))$ as a weighted penalized maximum likelihood estimator if $L_n^{\text{pen}}(\theta;x)$ has a local minimum at $(\hat{\mu}(x), \hat{\sigma}(x), \hat{\xi}(x))$. The parameter space of the GEV distribution, denoted $\theta(\mathcal{X}) = \{\zeta \in \mathbb{R} \times (0,+\infty) \times (-1,+\infty) : \zeta = \theta(x) \text{ for some } x \in \mathcal{X}\}$, is unknown in practice. We define our estimator $\hat{\theta}(x)$ as the value of the parameter θ that optimizes (5.10) over a compact set $\Theta \subset \mathbb{R} \times (0,+\infty) \times (-1,+\infty)$ such that $\theta(\mathcal{X}) \subset \Theta$. It is thus defined by

$$\hat{\theta}(x) \in \arg\min_{\theta \in \Theta} L_n^{pen}(\theta;x). \tag{5.11}$$

These estimated parameters are then substituted into (5.8) and the final estimation of the conditional quantile of order $\tau > \tau_0$ is given by

$$\hat{Q}_x(\tau) = \begin{cases} \hat{\mu}(x) + \dfrac{\hat{\sigma}(x)}{\hat{\xi}(x)}\left((-\ln(\tau^m))^{-\hat{\xi}(x)} - 1\right) & \text{if } \hat{\xi}(x) \neq 0, \\ \hat{\mu}(x) - \hat{\sigma}(x)\ln(-\ln(\tau^m)) & \text{if } \hat{\xi}(x) = 0. \end{cases} \tag{5.12}$$

Based on the work of (Gnecco et al., 2024), (Pasche and Engelke, 2023), and (Velthoen et al., 2023), we adopt in this study the value $\tau_0 = 0.8$, which corresponds to the probability level beyond which $Q_x(\tau)$ is considered an extreme quantile. Thus, our method applies to probability levels $\tau \geq \tau_0$. The algorithm outlining our approach is presented below in algorithm 2.

Algorithm

Let $\mathcal{D}_N = \{(X_i, Y_i)\}_{i=1\cdots N}$ denote the initial sample, and $\mathcal{D}'_n = \{(X_i, Z_i)\}_{i=1\cdots n}$ the sample of block maxima defined in Section 5.2, where m represents the block size, and υ is the vector containing the hyperparameters associated with the generalized random forest. The function

Algorithm 2 erf_Pen

procedure **erf_Pen-fit**
Input:$(\mathcal{D}_N, m, \upsilon)$
$\mathcal{D}'_n \leftarrow makeBloc(\mathcal{D}_N, m)$,
$w_i(.) \leftarrow grf(\mathcal{D}'_n, \upsilon)$,
Output: erf_Pen $\leftarrow (\mathcal{D}'_n, w_i(.), m)$.

procedure **erf_Pen-predict**
Input: (erf_Pen, x, τ, α, λ)
$\hat{\theta}(x) \leftarrow \arg\min_{\theta\in\Theta} L_n^{pen}(\theta; x)$ as defined in (5.11),
$\hat{Q}_x(\tau)$ is computed by substituting $\hat{\theta}(x)$ into equation (5.12).
Output: $\left(\hat{\theta}(x), \hat{Q}_x(\tau)\right)$.

makeBloc divides the initial sample into blocks of a given size m and returns the sample of block maxima. The grf function is used to adjust the weights using the generalized random forests method proposed by (Athey et al., 2019) and defined in section 5.2.

5.4 Simulation Study

In this section, we conduct simulation studies to demonstrate the performance of the proposed method and its ability to address the issues outlined in the introduction. An independent sample of size $N = 90,000$ is generated for the random vector (X, Y), where the covariate $X \in \mathbb{R}^p$ follows a uniform distribution over the hypercube $[-1, 1]^p$, and the conditional response variable $Y|X = x$ follows a heavy-tailed distribution, according to the simulation study. The goal is to estimate the extreme conditional quantile $Q_x(\tau)$ for high probability levels, with $\tau \in \{0.9, 0.99, 0.995, 0.9995, 0.9999\}$. The sensitivity analysis with respect to the block size is provided in Appendix 5.B. This analysis indicates that the model is sensitive to the choice

of block size m and suggests an optimal range of $30-70$ for Scenario 1 and $35-110$ for Scenario 2, ensuring good agreement between the estimated GEV distribution and the block maxima. To ensure a sufficient number of observations for the learning process of our method, we set the block size to $m = 40$, which produces $n = 2,225$ observations from the $N = 90,000$ generated data.

We compare our erf_Pen method with two other widely used quantile regression methods in the literature, namely the quantile regression forest (qrf) method proposed by (Meinshausen and Ridgeway, 2006) and the generalized random forest (grf) method proposed by (Athey et al., 2019). We also consider a third method dedicated to extreme values, the Generalized Additive Extreme Value Model (evgam) introduced by (Youngman, 2022), in which the parameters of the GEV distribution are expressed as generalized additive functions. In our study, the location, scale, and shape parameters are modeled as smooth additive functions of the covariates, without including interaction effects. In order to demonstrate both the necessity and superiority of the penalized model, we included the unpenalized model (denoted by Unpen_ERF) in our comparisons. This approach clearly illustrates the advantage of penalization. To evaluate the performance of the methods, we generate a test dataset $\{x_i\}_{i=1,\dots,n'}$ with $n' = 8,000$ (after forming the blocks with $m = 40$, which gives 200 observations for the test) using the Halton sequence (Halton, 1964). We use three metrics to compare the performance of the different methods. For each $\tau \geq \tau_0$, we calculate the integrated squared error (ISE) for the estimated conditional quantiles $\{\hat{Q}_{x_i}(\tau)\}_{i=1,\dots,n'}$ on the test data, as follows: $ISE = \frac{1}{n'}\sum_{i=1}^{n'}\left(\hat{Q}_{x_i}(\tau) - Q_{x_i}(\tau)\right)^2$, where $x \longmapsto Q_x(\tau)$ represents the true quantile function for the probability level τ. By averaging over $R = 100$ replicas of the fitting and estimation process for $Q_x(\tau)$, we obtain the mean integrated squared error (MISE), which is also used by (Gnecco et al., 2024) and (Velthoen et al., 2023) for evaluating the performance of their methods. The second metric we use is the integrated bias (IBias), introduced by (Wang and Li, 2013), which is defined as the average: $IBias = \frac{1}{n'}\sum_{i=1}^{n'}\left(\hat{Q}_{x_i}(\tau) - Q_{x_i}(\tau)\right)$, calculated on the test data. Finally, the third metric we use is the median absolute error (MedAE), which represents the median of the absolute differences $\left|\hat{Q}_{x_i}(\tau) - Q_{x_i}(\tau)\right|$ on the test data $\{x_i\}_{i=1,\dots,n'}$.

The first scenario of this simulation study aims to demonstrate the ability of our method to provide accurate estimates when the predictor space is of high dimension and to test its robustness against noise. This scenario is similar to those presented by (Gnecco et al., 2024) and (Velthoen et al., 2023) in their simulation studies. The second scenario seeks to evaluate the performance of our method in complex situations, where the quantile function exhibits a highly nonlinear shape and where the covariate space is also large, thus addressing the issues raised in the introduction. In Appendix 5.C, we include an additional simulation scenario to demonstrate the ability of erf_Pen to provide accurate estimates when $\xi > 1$, which corresponds to a critical case in extreme value modeling.

- **Scenario 1:** We assume that $Y|X = x \sim \gamma(x)\mathcal{T}_{\nu(x)}$, where $\mathcal{T}_{\nu(x)}$ represents the Student's

t-distribution with $\nu(x)$ degrees of freedom. The dependence functions are defined by

$$\gamma(x) = 1 + \mathbb{1}_{x_1>0} \quad \text{and} \quad \nu(x) = 4 - (x_1^2 - x_2^2).$$

In this scenario, only two variables, X_1 and X_2, are considered signal variables, while the remaining variables $(p-2)$ are noise. This framework allows us to assess the performance of our method in a context where the predictor space is of high dimension, while also accounting for the presence of noise. The results obtained for this scenario are presented in section 5.4.1.

- **Scenario 2:** We assume here that $Y|X = x \sim \gamma(x)\mathcal{T}_{\nu(x)}$, where

$$\gamma(x) = 1 + 2\pi\varphi(2x_1, 2x_2) \quad \text{and} \quad \nu(x) = 3 + \frac{7}{1+\exp(4x_1+1.2)},$$

with φ represents the density of the centered bivariate normal distribution with unit variance and a correlation coefficient of 0.75. In this scenario, we consider more complex forms for the functions $\gamma(x)$ and $\nu(x)$, which depend on the covariates x. This scenario thus allows us to assess the robustness of our method against combined challenges, such as those mentioned in the introduction. The results of this analysis are presented in section 5.4.2.

The performance of the generalized random forests method can be sensitive to the choice of hyperparameters, among which is the parameter `min.node.size`, which determines the minimum number of observations that a leaf can contain, and `num.trees`. These hyperparameters, as well as the penalty parameters λ and α in the penalized log-likelihood defined in (5.10), are optimized using the cross-validation method described in the appendix 5.A. In our simulation study, we determined the optimal value of `min.node.size` from $\{5, 10, 20, 50, 100\}$ using our cross-validation procedure. The other hyperparameters of the grf method were set to their default values in our simulation study.

5.4.1 Scenario 1

In this study, we evaluate the performance of our method in a context where the predictor space is high-dimensional, and for different probability levels close to 1. To ensure a fair comparison between the methods, the learning and estimation process is carried out on the same data, consisting of the block maxima $\{(X_i, Z_i)\}_{i=1,\ldots,n}$. Table 5.1 illustrates the evolution of the logarithm of MISE as a function of the quantile level τ for five conditional quantile estimation methods (erf_Pen, Unpen_ERF, evgam, grf, qrf) in Scenario 1. The results are presented for four configurations of covariate space dimensions: $p \in \{10, 20, 30, 40\}$.

Methods						
erf_Pen	Unpen_ERF	evgam	qrf	grf	p	τ
3.544159	3.554158	3.559510	3.534113	3.528069	10	0.8000
3.798552	3.827854	3.825582	3.841730	3.831264	10	0.9000
4.726821	4.857544	4.808876	5.049189	5.019799	10	0.9900
5.003168	5.175082	5.104060	5.541602	5.513555	10	0.9950
5.020182	5.420251	5.766500	6.166143	6.237731	10	0.9990
5.873965	6.223029	6.040300	6.121009	6.219681	10	0.9995
3.476097	3.480686	3.455641	3.497357	3.506545	20	0.8000
3.703584	3.717873	3.689100	3.667943	3.669805	20	0.9000
4.522016	4.586443	4.583590	4.651087	4.670408	20	0.9900
4.760237	4.844819	4.855621	4.874497	4.903185	20	0.9950
5.279417	5.421024	5.466527	5.902284	5.895118	20	0.9990
5.487002	5.658167	5.718320	6.355174	6.384959	20	0.9995
3.531561	3.541192	3.540931	3.525474	3.529149	30	0.8000
3.794515	3.821840	3.836681	3.738508	3.741153	30	0.9000
4.748053	4.860781	4.980860	5.128974	5.143107	30	0.9900
5.031631	5.176482	5.337251	5.571242	5.440019	30	0.9950
5.665040	5.895358	6.159317	6.883641	6.927317	30	0.9990
5.925686	6.197717	6.508989	7.145184	7.259607	30	0.9995
3.469519	3.471978	3.474885	3.443556	3.439969	40	0.8000
3.681014	3.688681	3.698054	3.682495	3.673479	40	0.9000
4.432620	4.467006	4.542034	4.472557	4.462648	40	0.9900
4.645826	4.690889	4.793666	4.684846	4.685848	40	0.9950
5.095707	5.170992	5.348638	5.598266	5.659914	40	0.9990
5.267733	5.358748	5.572876	5.568182	5.640856	40	0.9995

Table 5.1: Log(MISE) for different methods under varying dimensions p and probability levels τ.

The log(MISE) error increases gradually with τ, reflecting the growing difficulty in estimating extreme quantiles. This trend is particularly pronounced for $\tau \geq 0.9$, where the error rises almost exponentially, indicating increased variance and instability in predictions for these quantiles, as discussed in the introduction. For moderate probability levels ($0.8 \leq \tau \leq 0.9$), all methods exhibit similar performance, with nearly identical errors. In contrast, for higher quantiles ($\tau > 0.9$), the erf_Pen method stands out with lower error, especially as τ approaches 1. The evgam, grf, and qrf methods have slightly higher errors than the unpenalized Unpen_ERF method, which itself is somewhat less accurate than erf_Pen. For moderate dimensions p of the covariate space, evgam and Unpen_ERF have similar performance, but as p increases, evgam becomes less accurate than Unpen_ERF. Although evgam remains competitive compared to erf_Pen across different probability levels, its accuracy decreases as p increases. Furthermore, as pointed out by (Youngman, 2019), evgam becomes increasingly computationally demanding as p grows. These results indicate that our similarity-weighted estimation method provides

better estimates than the other methods, and the integration of penalization further improves its performance. Increasing the dimension of covariates from $p = 10$ to $p = 40$ leads to a general increase in the log(MISE) error, confirming the challenge of estimating conditional quantiles in high dimensions. However, this increase remains relatively moderate for erf_Pen, which retains an accuracy advantage, particularly for extreme values of τ. Thus, erf_Pen appears to be the most robust method under these conditions, exhibiting a more controlled growth of error compared to Unpen_ERF, evgam, grf, and qrf, making it a preferred choice for estimating conditional quantiles in high-dimensional settings and at extreme levels. This conclusion remains unchanged when performance is evaluated using other error metrics, as shown in the results presented in table 5.2.

5.4.2 Scenario 2

In this scenario, we incorporate the unpenalize model (Unpen_ERF) to compare its performance with proposed penalize method. Figure 5.1 illustrates the comparison of logarithmic integrated squared errors (Log(ISE)) obtained by different conditional quantile estimation methods under Scenario 2 of our simulation study. It evaluates the robustness of the methods against extreme quantiles and the increase in the dimension p of the covariates. Due to the high computational cost of the evgam method when the covariate dimension p is large, we exclude it from the comparison in this first analysis of this scenario. However, we include it in the second analysis summarized in Table 5.2, for a fixed covariate dimension $p = 40$, to demonstrate that the results obtained in Scenario 1 remain valid for Scenario 2 across the different evaluation metrics considered. According to Figure 5.1, when $\tau = 0.99$, the Unpen_ERF, grf, and qrf methods yield almost similar performance, but the erf_Pen method performs better, showing relatively moderate errors across the different covariate spaces considered, $p \in \{30, 40, 60\}$. However, the Unpen_ERF method, which does not incorporate penalization, although it shows better performance than the grf and qrf methods, exhibits larger errors than erf_Pen, thereby confirming the usefulness of penalization for this type of estimation. As τ increases ($\tau = 0.99, 0.995, 0.9995$), we observe an increase in error for all methods. However, the erf_Pen method maintains more stable performance, with systematically lower error compared to grf and qrf, which are more affected by the extremity of the quantiles. This indicates that erf_Pen is better suited for extreme quantile estimation. As the dimension of the covariate space increases ($p = 30, 40, 60$), the error tends to grow for all methods, reflecting the increased difficulty of estimation in high dimensions. However, erf_Pen again stands out by showing a more controlled progression of error, while grf and qrf display more dispersed errors, indicating a loss of stability in their estimates. In all configurations tested, erf_Pen proves to be the most effective method. It exhibits significantly lower error than its competitors, particularly for extreme quantiles and in high dimensions. It is also more stable, with less

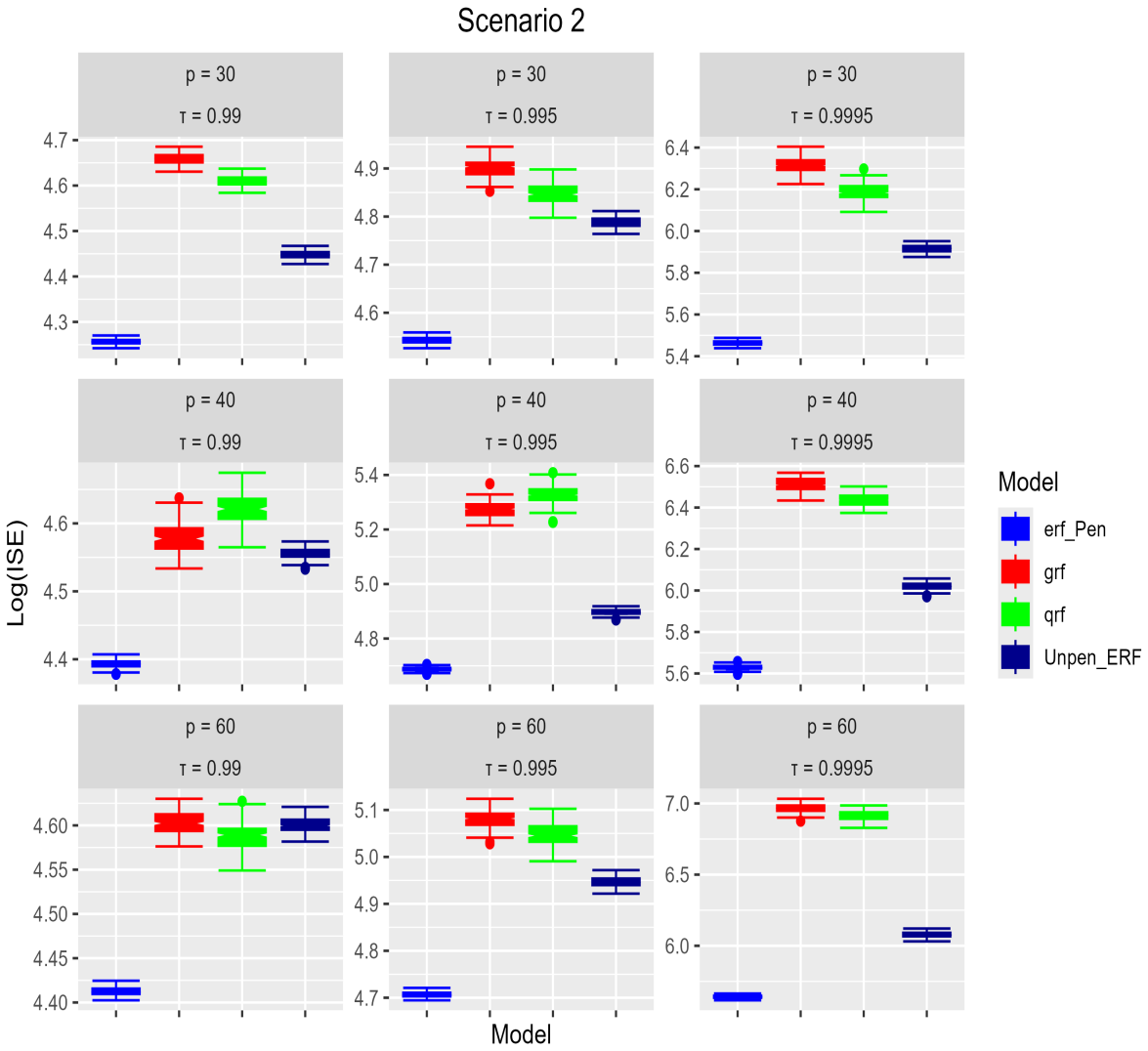

Figure 5.1: Boxplots of Log(ISE) over 100 simulations for different values of p and extreme probability levels.

variability in error compared to methods that show greater variability in their results. These results show that the erf_Pen method is the most robust for extreme quantile estimation, even in the presence of numerous covariates. Its effectiveness is particularly notable for high values of τ, where it greatly outperforms the grf method of (Athey et al., 2019) and the qrf method of (Meinshausen and Ridgeway, 2006). The increase in covariate dimension affects all methods, but erf_Pen is better able to withstand this complexity. In summary, erf_Pen stands out as the most suitable method for applications requiring reliable extreme quantile estimation, particularly in high-dimensional contexts. We also observe that our model remains effective in the presence of noise. The results presented in table 5.2 further confirm the superiority of the erf_Pen method by using other error evaluation metrics. This table shows that in terms of MedAE and IBias, the erf_Pen method outperforms the Unpen_ERF, evgam, grf and qrf meth-

Method	MedAE	IBias	MISE	τ	Scenario
erf_Pen	6.20	6.31	3.68		
Unpen_ERF	6.22	6.33	3.69		
evgam	6.20	6.23	3.70	**0.9**	
grf	6.23	6.30	3.68		
qrf	6.26	6.31	3.67		
erf_Pen	9.87	10.1	4.64		
Unpen_ERF	10.1	10.2	4.69		
evagm	10.1	10.4	4.79	**0.995**	**1**
grf	10.3	10.4	4.69		
qrf	10.3	10.4	4.68		
erf_Pen	13.2	13.6	5.27		
Unpen_ERF	13.8	14.2	5.36		
evagm	14.7	15.4	5.56	**0.9995**	
grf	16.7	16.4	5.64		
qrf	16.4	15.7	5.57		
erf_Pen	5.48	5.42	3.43		
Unpen_ERF	5.68	5.60	3.47		
evgam	5.58	5.57	3.46	**0.9**	
grf	5.76	5.64	3.49		
qrf	5.70	5.63	3.48		
erf_Pen	9.06	9.13	4.63		
Unpen_ERF	10.9	10.9	4.83		
evgam	11.2	11.5	5.02	**0.995**	**2**
grf	12.0	12.6	5.18		
qrf	11.7	12.6	5.19		
erf_Pen	11.8	11.7	5.49		
Unpen_ERF	15.6	15.6	5.80		
evgam	16.7	17.3	5.82	**0.9995**	
grf	17.2	17.8	5.91		
qrf	17.0	17.4	5.84		

Table 5.2: Table of errors according to various metrics for each method and different scenarios.

ods across different high probability levels, when the covariate dimension is set to $p = 40$.

5.5 Real dataset

We apply our methodology to predict the extreme quantiles of wage data from the 1980 U.S. Census. This dataset was used by (Angrist et al., 2006) to illustrate a quantile regression method in non-extreme contexts, and more recently by (Gnecco et al., 2024) in the analysis of extreme quantiles. The dataset consists of $65,023$ Black and White men, aged 40 to 49 years, with 5 to 20 years of education, and positive annual incomes and hours worked in the year prior to the census. In this study, we use weekly wages from 1979 as the response variable Y, expressed in U.S. dollars, calculated as the annual income divided by the number of weeks worked. The explanatory variables include a vector containing numerical variables representing age and years of education, as well as a categorical variable called Black, which is equal to 1 if the respondent is Black and 0 if the respondent is White. To increase the dimension of the predictor space and better assess the performance of our methodology in cases where the covariate space is large, we add 12 random predictors, generated independently and uniformly within the interval $(-1,1)$. This brings the total dimension of the predictor vector to $p = 15$. As done by (Gnecco et al., 2024), we split the dataset into two equal parts: the first part contains $32,511$ observations, used for exploratory analysis, and the other part contains $32,512$ observations, which is used for fitting and quantitatively evaluating the methods, including our proposed method, erf_Pen, as well as the three other methods used in the simulation study (evgam from (**?**), grf from (Athey et al., 2019) and qrf from (Meinshausen and Ridgeway, 2006)).

For the exploratory analysis, we fit the erf_Pen model to 40% of the data, which corresponds to a total of $13,004$ observations, and estimate the parameters of the conditional generalized extreme value (GEV) distribution, $\hat{\theta}(x) = (\hat{\mu}(x), \hat{\sigma}(x), \hat{\xi}(x))$, using the remaining 60% of the data, applying the algorithm 2. Since, as shown in Appendix 5.5, our method is sensitive to the block size, we used blocks of size $m = 5$ observations in the exploratory analysis to improve estimation in the tail of the distribution. However, for the quantitative analysis, we applied the test statistic described in Appendix 5.B and found that block sizes between 5 and 50 provide satisfactory goodness-of-fit. Therefore, we selected a block size of $m = 20$ for the quantitative analysis. To determine the optimal values of the hyperparameters α and λ, which appear in the expression of the penalized likelihood given in (5.10), as well as *min.node.size* for the grf method, we apply cross-validation as described in appendix 5.A. This procedure is performed for $(\alpha, \lambda) \in (0,5] \times (0,4]$ and *min.node.size* $\in \{5, 10, 30, 50\}$, keeping the default values for the other hyperparameters of the grf method. Finally, we proceed with the estimation of the parameters θ. Figure 5.2 shows the variation of the estimated parameters $\hat{\mu}(x)$, $\hat{\sigma}(x)$, and $\hat{\xi}(x)$ as a function of the number of years of education. It is observed that $\hat{\mu}(x)$ and $\hat{\sigma}(x)$

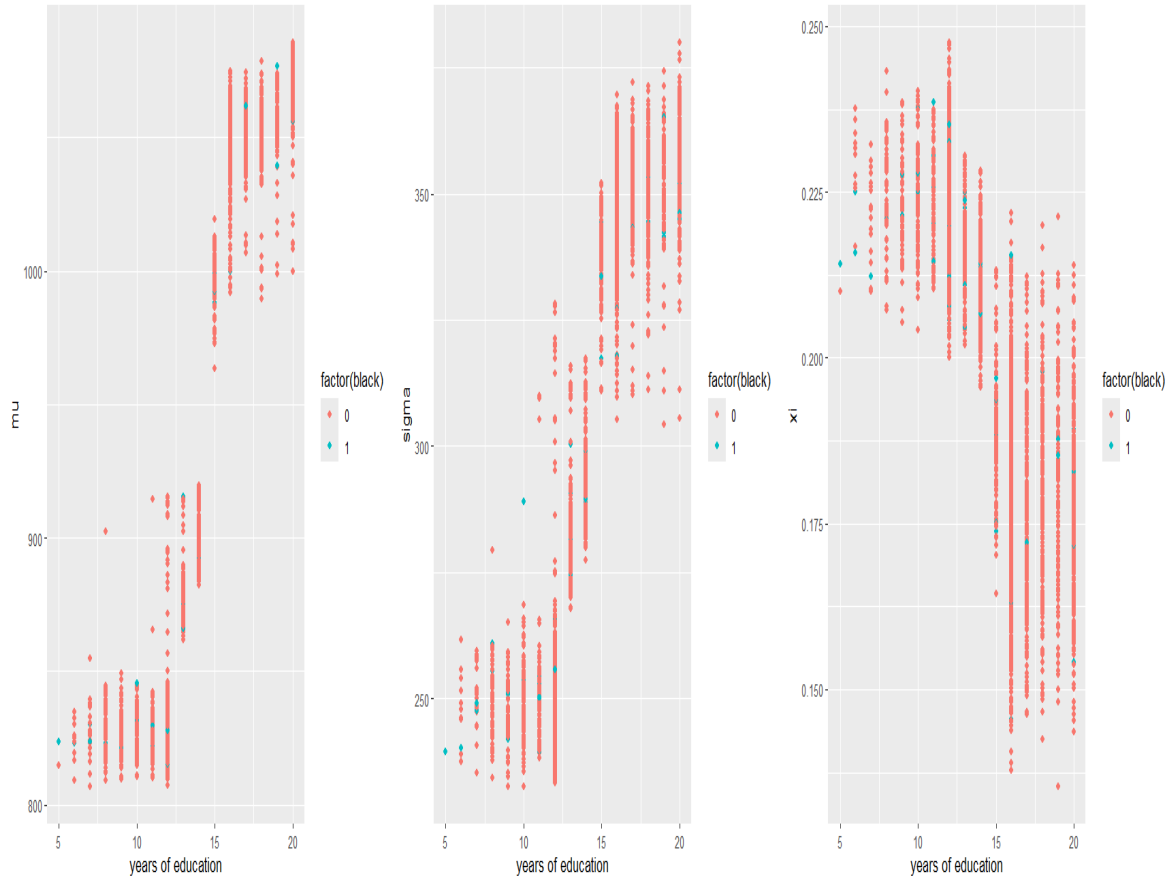

Figure 5.2: Variations of the parameters $\hat{\mu}(x)$, $\hat{\sigma}(x)$, and $\hat{\xi}(x)$ as a function of the number of years of education.

increase with the level of education, while the shape parameter $\hat{\xi}(x)$ follows an opposite trend, decreasing as the level of education increases. It is also noted that $\hat{\xi}(x)$ takes positive values between 0.1 and 0.25, indicating heavy tails in the predictor space, thus confirming the analysis conducted by (Gnecco et al., 2024). Furthermore, the study highlights a balanced distribution of wages between Black and White Americans and shows that the parameters of the conditional GEV distribution are not influenced by age, as illustrated in figure 5.7 in appendix 5.D. Figure 5.3 presents the estimates of the conditional quantiles of weekly income ($\hat{Q}_x(\tau)$) as a function of years of education, obtained by the three methods: erf_Pen, grf, and qrf. The results are displayed for three quantile levels ($\tau = 0.8$, $\tau = 0.9$, and $\tau = 0.995$). The red and blue points correspond to individuals for whom the variable `black` is 0 or 1, respectively. An increase in income is observed with the number of years of education, a trend that becomes more pronounced at the highest quantiles, indicating that individuals with the highest income have spent significantly more years in education. When comparing the methods, our extrapolation method maintains a good shape of the quantile function even for high probability levels, which is not the case for the grf and qrf methods, where the shape deteriorates as τ increases

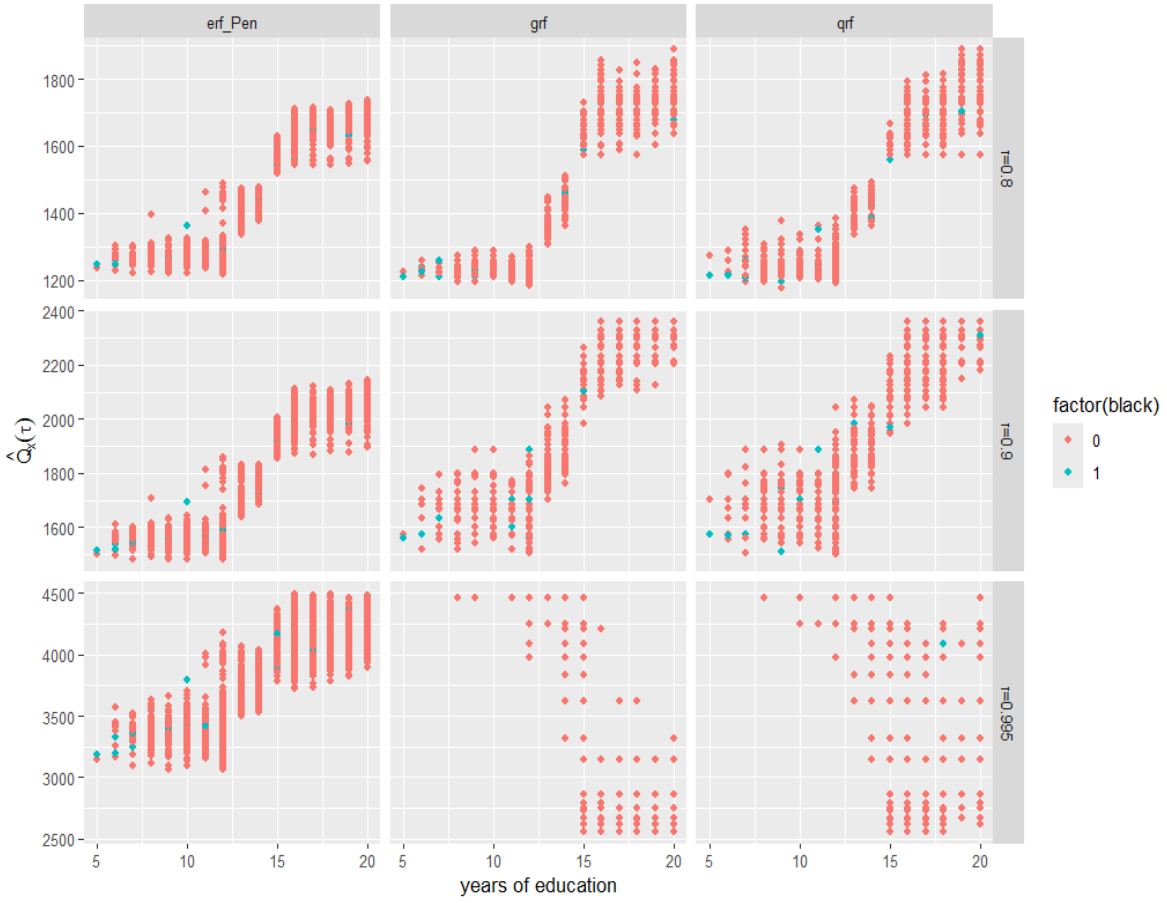

Figure 5.3: Predicted conditionnal quantiles at level $\tau = 0.8, 0.9, 0.995$ as fonction of years of eduction for erf_Pen, grf and qrf method.

(particularly for $\tau = 0.995$). This highlights the ability of our method to capture the complex structure of the quantile function. This analysis underscores the growing impact of education on income.

After the exploratory analysis, we assess the quantitative performance of our method as well as the three other methods using the metric proposed by (Wang and Li, 2013), which is also employed by (Gnecco et al., 2024) to evaluate the performance of their own method. This metric is defined as follows:

$$R_{n'}\left(\hat{Q}_x(\tau)\right) = \frac{\sum_{i=1}^{n'} \mathbb{1}\{Y_i < \hat{Q}_{X_i}(\tau)\} - n'\tau}{\sqrt{n'\tau(1-\tau)}}. \tag{5.13}$$

The function $\hat{Q}_x(\tau)$ represents the estimated conditional quantile on the test sample $\{(x_i, y_i)\}_{i=1,\cdots,n'}$. The metric $R_{n'}(\cdot)$ relies on the principle that, for a quantile at probability level τ, the proportion of observed values y_i below the predicted quantile $\hat{Q}_{x_i}(\tau)$ should be close to τ. Since $\mathbb{1}_{\{y_i < \hat{Q}_{x_i}(\tau)\}}$ has expectation τ and variance $\tau(1-\tau)$, this metric is asymptotically normal according to the central limit theorem. A value of $R_{n'}(\cdot)$ close to zero indicates that the method provides pre-

dictions consistent with the theoretical definition of the quantile, while a large value reveals an underestimation or overestimation bias. This measure is particularly useful for the evaluation of extreme quantiles, as it captures the model's ability to correctly reproduce the expected frequency of rare events, beyond standard criteria such as bias or mean squared error. For this purpose, we use the second part of the data, which was not used for the exploratory analysis, i.e., $32,512$ observations and consider the block size $m = 20$. We partition these data into 5

Model	$\left\|R_{n'}\left(\hat{Q}_x(\tau)\right)\right\|$			
	$\tau = 0.8$	$\tau = 0.9$	$\tau = 0.995$	$\tau = 0.9999$
erf_Pen	3.951039	9.014595	59.755834	467.221579
Unpen_ERF	4.205856	9.335311	60.925626	475.764389
evgam	4.256064	9.200778	61.227295	483.950181
grf	3.210531	10.292044	68.164030	488.939327
qrf	3.210713	10.302165	68.364223	488.282943

Table 5.3: Performance of models based on the metrics defined in (5.13).

parts and, for each part $i \in \{1,2,3,4,5\}$, we train the model on $4/5$ of the observations, representing the set of observations excluding the i-th part. We then calculate the absolute value of $R_{n'}\left(\hat{Q}_x(\tau)\right)$ on the $1/5$ of the observations contained in the i-th part. After calculating this absolute error for each $i \in \{1,2,3,4,5\}$, we compare the average of these absolute errors for the different models and for different quantile orders. In this quantitative analysis, we included in our comparisons the unpenalized weighted likelihood estimator and evgam method. The table 5.3 presents the mean values of these absolute errors over the ten repetitions for different quantile levels. From this table, it is clear that our method provides good estimates compared to the methods of (Youngman, 2022), (Athey et al., 2019) and (Meinshausen and Ridgeway, 2006). The results show that the weighted likelihood estimator provides satisfactory estimates, whose accuracy improves when penalization is introduced. The evgam method remains competitive with the unpenalized version and, for high quantile levels ($\tau = 0.995$), achieves performance comparable to that of erf_Pen. Nevertheless, our approach remains stable at extreme quantile levels.

Although the real data application does not fall within the $\xi > 1$ regime, the additional simulation study presented in Appendix 5.C highlights the limitations of the PMLE in the presence of very heavy tails. This is reflected in the larger values of $\log(\text{MISE})$ compared to Scenarios 1 and 2. Figure 5.6, which reports the results for this critical setting, shows that our method provides the most accurate estimates across the different high-dimensional covariate configurations considered. Overall, these findings confirm the robustness and reliability of our approach across varying degrees of tail heaviness, including the challenging case $\xi > 1$ examined in the simulation study.

5.6 Conclusion

In this paper, we propose a new flexible method, named erf_Pen, for the reliable estimation of conditional quantiles at extreme probability levels. This method is also robust when the shape of the quantile function is complex and when predictors are potentially high-dimensional. Our methodology leverages the flexibility of generalized random forests to capture the complex structure of the quantile function, as well as the asymptotic results of extreme value theory (Fisher and Tippett, 1928), (Gnedenko, 1943) to facilitate extrapolation in the tail of the distribution via the conditional GEV distribution. The estimation of the latter's parameters, using the penalty function of (Coles and Dixon, 1999), improves the estimation of the extreme value index, a key component in conditional quantile estimation. Our approach stands out for its robustness and its ability to effectively capture the complex structure of the data while enabling reliable estimation in the tail of the distribution. Through a simulation study, we demonstrate that erf_Pen outperforms other existing methods, notably the generalized random forests method of (Athey et al., 2019), the Generalized Additive Extreme Value Model of (Youngman, 2022) and the quantile regression forests method of (Meinshausen and Ridgeway, 2006). It offers greater stability and improved accuracy for high quantile levels ($\tau > 0.9$), even in the presence of noise. Our results also show that erf_Pen is more resistant to increasing covariate dimensionality, making it a particularly suitable alternative for high-dimensional applications. An application to the 1980 U.S. Census wage data confirmed these empirical findings, demonstrating that our method preserves a good estimation of the quantile function, even for extreme quantiles ($\tau = 0.995$). In conclusion, erf_Pen represents a significant methodological advancement in the estimation of extreme conditional quantiles. Its performance and robustness make it a promising tool for applications in risk management, finance, economics, and other fields requiring reliable estimation of extreme conditional quantiles. Future work may include extending this approach to multivariate contexts as well as further theoretical analysis.

Appendix

5.A Cross-validation method used to obtain α and λ and the hyperparameters of grf.

In this section, we present the method used to determine the tuning parameters of our model. Specifically, we address the selection of the penalization parameters λ and α, as well as the hyperparameters specific to the generalized random forest (such as *min.node.size* and *num.trees*). These parameters are chosen using a cross-validation procedure, adopting an approach different from the classical method described in section 7.10 of (Hastie, 2017). We rely on the

approach proposed by (Gnecco et al., 2024), which uses the likelihood function of the GPD as an error measure. However, in our case, we use the likelihood of the GEV distribution as the performance evaluation metric.

To illustrate this method, let's consider a sample $\mathcal{D}_n = \{(x_i, z_i)\}_{i=1,\cdots,n}$, available for our study. Suppose we wish to apply K-fold cross-validation. The first step is to partition the sample into K approximately equal-sized sub-samples, denoted $\mathcal{D}^j$, with $j = 1, \cdots, K$. A set of potential values is pre-defined for the penalization parameters λ and α, as well as for the hyperparameters specific to the generalized random forest (grf), such as *min.node.size* and *num.trees*. The cross-validation procedure then involves systematically evaluating all possible combinations of these parameters to identify those that optimize the model's performance. To do this, a search grid is constructed, denoted $\{\rho_1, \cdots, \rho_S\}$, where each ρ_s represents a vector containing a particular combination of the grf hyperparameters and the penalization parameters λ and α. For each parameter configuration $\rho_s \in \{\rho_1, \cdots, \rho_S\}$, the following steps are performed for each fold $j \in \{1, \cdots, K\}$:

1. **Model training:** The model is trained on the $K-1$ sub-samples, that is, on the entire dataset $\mathcal{D}_n \setminus \mathcal{D}^j$, excluding the sub-sample $\mathcal{D}^j$. This training is carried out using the *Pen-erf-fit* algorithm, with ρ_s as the configuration for the tuning parameters and the hyperparameters of the generalized random forest (grf).

2. **Estimation of conditional parameters:** the parameters $\theta(x)$ associated with the conditional GEV distribution are estimated using the data from the sub-sample $\mathcal{D}^j$.

3. **Performance Evaluation:** A performance metric is calculated on the sub-sample $\mathcal{D}^j$. In this context, the chosen metric is the negative log-likelihood of the GEV distribution.

Once these steps are applied to the K sub-samples, the average error across the K partitions is calculated. This average error, called the cross-validation error (*CV-error*), provides an overall estimate of the model's performance for a given configuration ρ_s. It is defined by:

$$CV(\rho_s) = \frac{1}{K} \sum_{j=1}^{K} \sum_{(x_i, z_i) \in \mathcal{D}^j} \ell_{(\hat{\theta}(x_i), \rho_s)}(z_i),$$

where $\ell_{(\hat{\theta}(x), \rho_s)}(z_i) = -\log\left(\frac{dG(z_i;\theta)}{dz_i}\right)$ represents the negative log-likelihood of the GEV distribution, with the parameters $\hat{\theta}(x)$ estimated using ρ_s as the parameter configuration. This approach allows us to identify the optimal configurations of tuning parameters and hyperparameters by minimizing the cross-validation error. The optimal parameter ρ_{optim} is the one that minimizes this error among all the configurations available in the set $\{\rho_1, \cdots, \rho_S\}$. Formally, it is defined as:

$$\rho_{optim} = \arg \min_{s \in \{1, \cdots, S\}} CV(\rho_s).$$

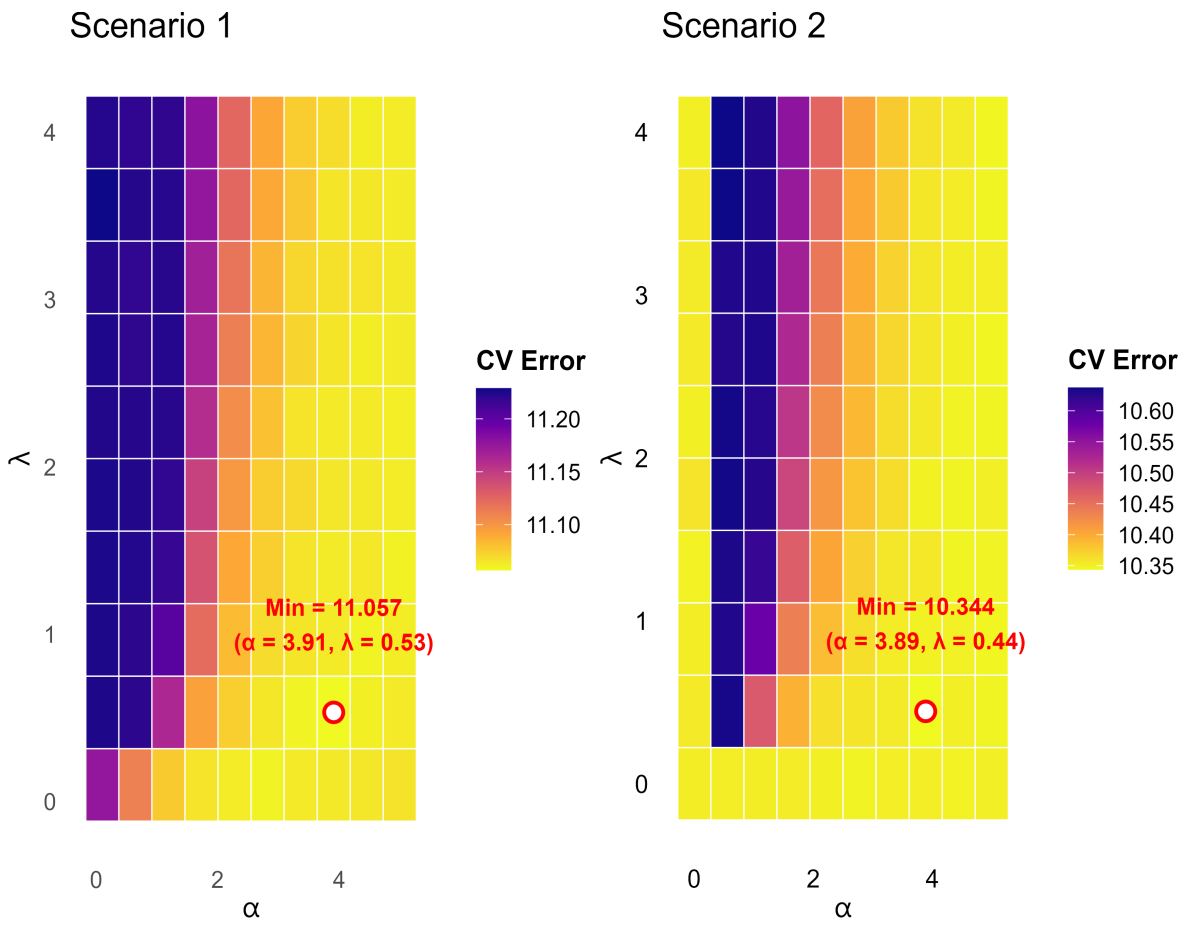

Figure 5.4: Cross-validation Error across Scenarios as a function of α and λ

This methodology ensures a robust and objective evaluation of the different parameter configurations, while identifying those that minimize the cross-validation error. This allows for reliable optimization of the model parameters, tailored to the specific characteristics of the data being studied. The figure 5.4 illustrates how the cross-validation error varies with $\alpha \in (0,5]$ and $\lambda \in (0,4]$ across the two scenarios considered in this work. As shown in this figure, the optimal values of these parameters depend on the dataset used. However, the cross-validation error varies only moderately across the different parameter grids considered. This indicates that while satisfactory performance is obtained when $\alpha = \lambda = 1$, the proposed cross-validation procedure can still assist in selecting appropriate optimal values depending on the data.

5.B Sensitivity analysis

In this section, we investigate the sensitivity of the erf_Pen method with respect to the choice of block size. This parameter is crucial in the block maxima framework: a block size that is too small induces bias in the estimation, while a block size that is too large increases the variance of the estimator. The choice of m must therefore achieve a bias–variance trade-off. To the best

of our knowledge, there is no universally optimal procedure for selecting the block size, and practical applications often rely on natural choices such as monthly, quarterly, semi-annual, or annual blocks. We propose here an alternative approach for selecting the block size based

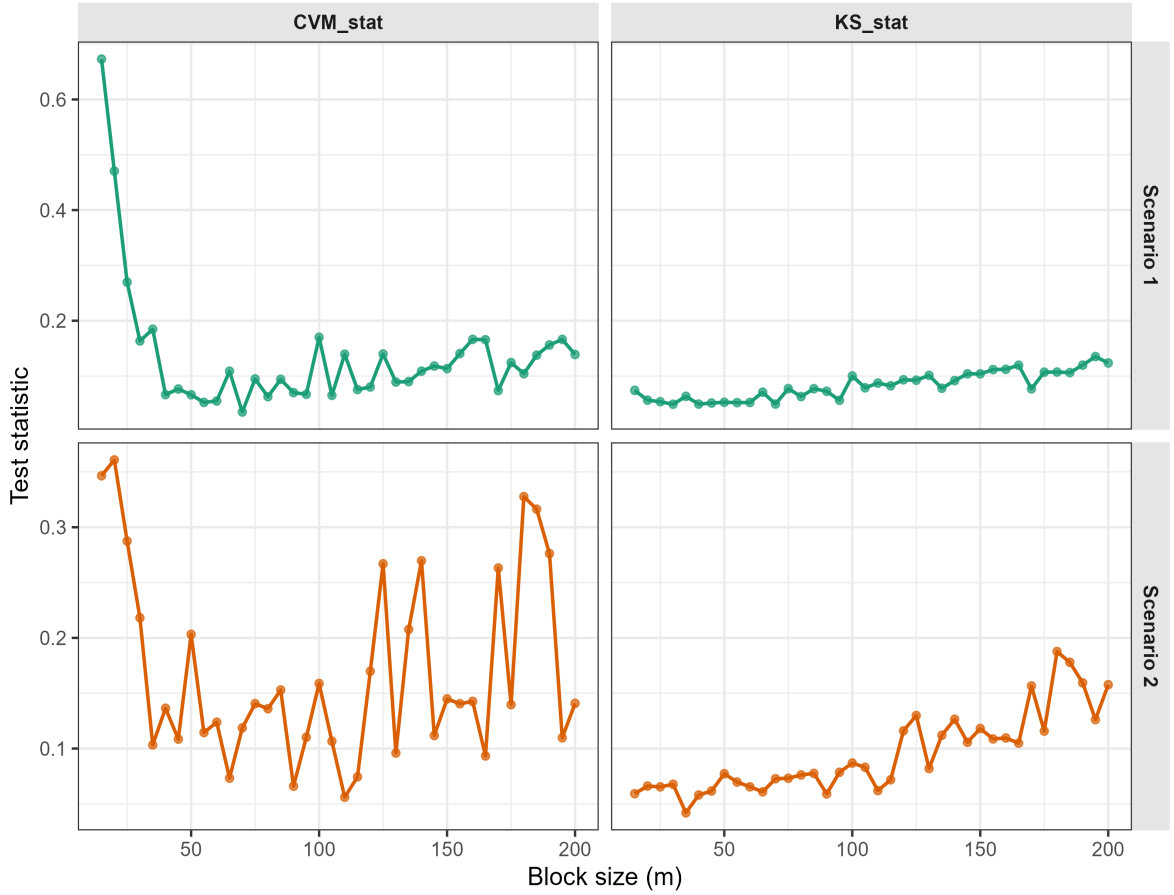

Figure 5.5: Evolution of test statistics as a function of block size under different scenarios

on goodness-of-fit test statistics, as done by (Wang et al., 2016) in selecting block sizes for estimating extreme loads in engineering vehicles. Specifically, we fit the GEV model with different block sizes and evaluate the goodness-of-fit of the probability integral transform data against the uniform $(0,1)$ distribution using the Kolmogorov–Smirnov (KS) and Cramér–von Mises (CVM) tests. The KS test measures the maximum deviation between the empirical and theoretical distributions, while the CVM test integrates the squared deviations over the entire support, thus providing a more global assessment. Their joint use yields complementary and robust evidence of goodness-of-fit, which is particularly useful for guiding the choice of block size.

To this end, we consider a training sample $\mathcal{D}_N^{train} = \{(x_i, y_i)\}_{i=1}^N$ and an independent test sample $\mathcal{D}_l^{test} = \{(x_i, y_i)\}_{i=1}^l$. We also consider a range of block sizes m between 15 and 200, in increments of 5. For each value of m, the model is fitted on the training sample $\mathcal{D}_N^{train}$, and the parameters of the GEV distribution $(\mu(x), \sigma(x), \xi(x))$ are then estimated using the test sample

$\mathcal{D}_l^{test}$. Let l' denote the number of blocks of size m formed from the test data. The probability integral transforms $\{G_{(\hat{\mu}(x_i),\hat{\sigma}(x_i),\hat{\xi}(x_i))}(z_i)\}_{i=1}^{l'}$ should follow a uniform distribution if the estimation is adequate, where z_i denotes the maximum associated with the i-th block. In this analysis, the penalization parameters are fixed at $\lambda = \alpha = 1$, following the recommendations of (Coles and Dixon, 1999), and we adopt the default values of min.node.size and num.trees for the generalized random forest method proposed by (Athey et al., 2019). Figure 5.5 displays the evolution of the KS and CVM test statistics for the two scenarios considered. In Scenario 1 (top panels), the CVM statistic decreases sharply for block sizes smaller than $m = 30$, then stabilizes at a relatively low and constant level for m between 35 and 100. The KS statistic remains globally low but exhibits a slight upward trend as m increases, suggesting a gradual loss of fit for very large block sizes. In Scenario 2 (bottom panels), the results are more variable. The CVM statistic shows strong instability for small block sizes ($m < 35$), followed by marked fluctuations thereafter, indicating increased sensitivity of the estimation. The KS statistic exhibits a generally increasing trend with larger m, highlighting that excessively large block sizes deteriorate the goodness-of-fit.

Overall, these results confirm the method's sensitivity to block size and suggest an optimal range of $30-70$ for Scenario 1 and $35-110$ for Scenario 2, ensuring good agreement between the estimated GEV distribution and the block maxima. In practice, this type of sensitivity analysis can guide the empirical choice of m, although natural block sizes remain preferable. For our analyses, we selected $m = 40$ in both scenarios, in order to balance model adequacy with a sufficient number of block maxima for the learning process.

5.C Additional Simulation Study

In this section, we conduct an additional simulation study to assess the ability of erf_Pen to more accurately estimate conditional quantiles in a heavy-tailed regime where $\xi > 1$. This setting is particularly challenging, as some moments of the distribution may be infinite, making the estimation of extreme quantiles notoriously difficult. We consider a conditional distribution of the form $Y \mid X = x \sim \gamma(x)\,\mathcal{T}_{\nu(x)}$, where $\mathcal{T}_{\nu(x)}$ denotes a Student's t distribution with $\nu(x)$ degrees of freedom. As in Scenario 1, we set $\gamma(x) = 1 + \mathbb{1}_{\{x_1>0\}}$, thereby introducing heterogeneity in the conditional scale. In this study, we assume a constant shape parameter given by

$$\xi(x) = \frac{1}{\nu(x)} = 1.5, \quad \text{for all } x \in [-1,1]^p,$$

which corresponds to a heavy-tailed regime. This design allows us to evaluate the robustness of the proposed approach under a critical scenario in which extreme quantile estimation is particularly demanding. The covariates $X = (X_1, \ldots, X_p)$ are generated independently from a uniform distribution on $[-1,1]^p$, with $p \in \{30, 50\}$, consistent with the two scenarios consid-

ered in our simulation study. We evaluate the `log(MISE)` as a function of the probability level $\tau \in [0.8, 1)$ for both covariate dimensions. We do not include the evgam model in this comparison because it becomes computationally expensive as p increases. Moreover, although its performance is superior to that of the grf and qrf methods for moderate values of p, it becomes comparable to that of Unpen_ERF, grf, and qrf for larger values of p, as evidenced by the results reported in Tables 5.1 and 5.2.

The figure 5.6 shows the evolution of log(MISE) as a function of probability levels τ. The results shown in this figure indicate that erf_Pen remains stable and delivers more accurate estimates than competing methods in this strongly heavy-tailed regime. In particular, the beneficial effect of penalization becomes more pronounced as both the probability level τ and the dimension p increase, suggesting that the proposed approach is especially well suited for extreme quantile estimation when $\xi > 1$. We also observe that the non-penalized method performs similarly to qrf and grf for moderate dimensionality ($p = 30$), whereas for higher dimensionality ($p = 50$), it outperforms both competitors at high probability levels. These findings confirm that erf_Pen remains applicable and effective beyond the standard setting $\xi < 1$, extending its relevance to heavy-tailed contexts with $\xi \geq 1$.

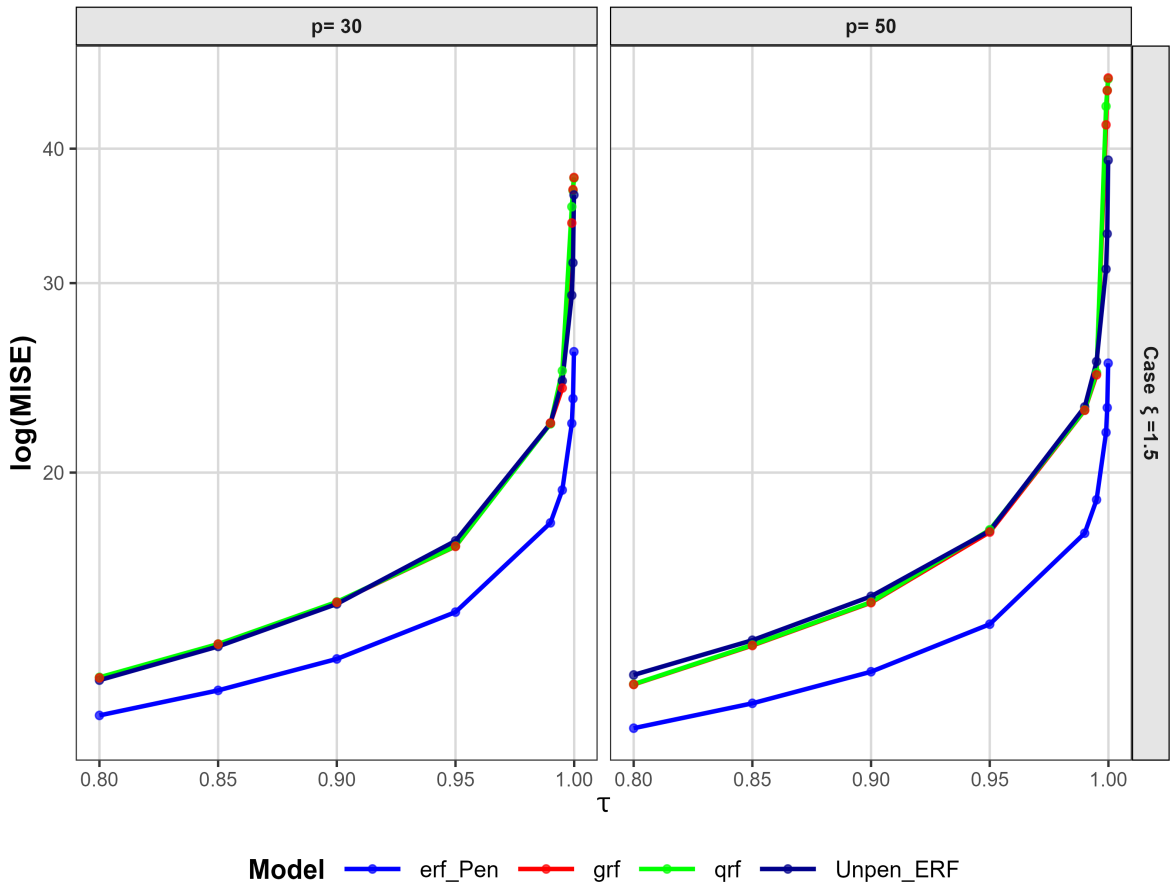

Figure 5.6: Evolution of log(MISE) as a function of τ for $p \in \{30, 50\}$

5.D Variation of the parameters $\hat{\mu}(x)$, $\hat{\sigma}(x)$, and $\hat{\xi}(x)$ as a function of age.

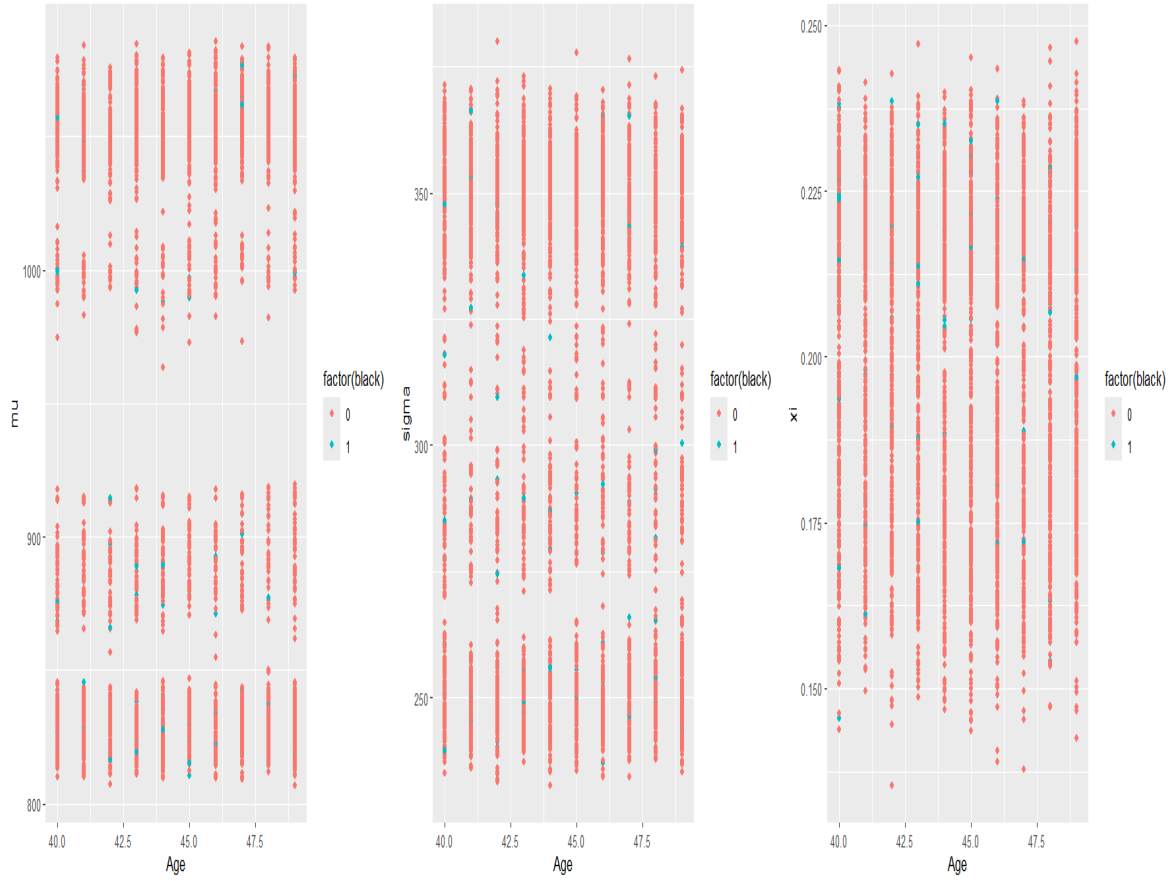

Figure 5.7: Variation of the parameters $\hat{\mu}(x)$, $\hat{\sigma}(x)$, and $\hat{\xi}(x)$ as a function of age.

Conclusion and Perspectives

This thesis lies within the field of statistical modelling of extreme values and asymptotic behaviours. Its objective was to develop robust and flexible methods for analysing and predicting the risk associated with rare events, which are of major importance in domains such as finance, insurance, environmental sciences, and engineering. To address these challenges, we proposed several methodological approaches based on extreme quantile regression, whose theoretical properties were studied, empirical performances validated, and applications illustrated using both simulated and real datasets.

In the first part, we introduced a method for estimating conditional extreme quantiles based on the conditional Generalized Extreme Value (GEV) distribution. The proposed estimator relies on a weighted version of the maximum likelihood estimator, where the weights are derived from generalized random forests. This approach overcomes several limitations of classical quantile regression methods, in particular the difficulty of estimating extreme quantiles when $\tau_n \to 1$, as well as the challenge posed by high-dimensional covariate spaces. We established the existence and convergence of the proposed weighted likelihood estimator, proving its consistency under realistic assumptions tailored to the context of extremes.

In the second part, we proposed a more application-oriented model by introducing an L_2-type penalisation on the shape index ξ of the extreme value distribution. Detailed in Chapter 4, this penalisation makes it possible to better capture the asymptotic behaviour of the variables under study, even in the presence of scarce, heterogeneous, or noisy data. It enhances the stability of high-quantile estimation while preserving desirable asymptotic properties. We further developed suitable estimation algorithms and proposed strategies for tuning the model parameters (block size, penalisation coefficient), allowing optimisation of the method's robustness. The approach was validated through simulation studies and illustrated using daily meteorological data from the Fort Collins weather station (Colorado, USA).

Finally, we proposed a weighted version of the penalised maximum likelihood estimator originally introduced by (Coles and Dixon, 1999), in which the penalisation function is specifically designed for the GEV distribution. This formulation simultaneously addresses: (i) the limitations encountered in quantile regression for the estimation of very high quantiles, and (ii) the need to improve estimation performance for small samples while maintaining strong efficiency for larger samples. We validated this method through extensive comparisons with standard approaches based on statistical learning techniques. The results show that the pro-

posed model provides improved performance in terms of accuracy, stability, and robustness. The application to real data confirms the practical relevance of the approach for forecasting rare events in various contexts, thereby offering reliable tools for risk management and decision support.

Perspectives

Several research directions naturally emerge from the work carried out in this thesis.

- **Extension to the multivariate setting.** Extending our methods to multivariate models would enable the study of co-occurring extreme events and allow better modelling of dependence structures.
- **Temporal data.** Adapting our procedures to time series would open the way to dynamic analysis of extreme risks, including the detection of structural changes or the study of persistent extremes.
- **Spatial and spatio-temporal models.** Extending our approaches to the spatial domain would constitute a major advancement for the analysis of geographically distributed phenomena (extreme precipitation, pollution, natural hazards, etc.).
- **Software development.** Creating an `R` package integrating the methods developed in this thesis would greatly facilitate their dissemination within the scientific community.
- **Interdisciplinary applications.** Fields related to natural disaster management, industrial safety, or the analysis of rare biomedical signals offer many opportunities for applying the tools developed in this work.

This thesis provides significant contributions to the statistical modelling of extreme values and asymptotic behaviours. The developed methods combine theoretical rigour, practical robustness, and flexibility, offering new perspectives for the analysis of rare events in complex environments. The obtained results pave the way for ambitious future research, with the potential to enrich the study of extreme risks and to strengthen the impact of modern statistical methods across numerous applied domains.

References

Abad, P., Benito, S., and López, C. (2014). A comprehensive review of Value at Risk methodologies. *The Spanish Review of Financial Economics*, 12(1):15–32.

Abdelaziz, R. (2013). Application de la théorie des valeurs extrêmes pour estimer quelques outils probabilistes dans l'hydrologie et l'actuariat.

Afroz, R., Johnson, F., and Sharma, A. (2021). The residual mass severity index – A new method to characterize sustained hydroclimatic extremes. *Journal of Hydrology*, 602:126724.

Al-Behadili, H., Grumpe, A., and Wöhler, C. (2016). Outlier detection based on confidence band and extreme value theory for semi-supervised learning of an incremental polynomial classifier. *International Journal of Simulation: Systems, Science and Technology*, 17:15.1–15.7.

Allouche, M., Girard, S., and Gobet, E. (2024). Estimation of extreme quantiles from heavy-tailed distributions with neural networks. *Statistics and Computing*, 34(1):1–17. Company: Springer Distributor: Springer Institution: Springer Label: Springer Number: 1.

Angrist, J., Chernozhukov, V., and Fernández-Val, I. (2006). Quantile regression under misspecification, with an application to the US wage structure. *Econometrica*, 74(2):539–563.

Arlot, S. (2018). Fondamentaux de l'apprentissage statistique.

Arnell, N. W. (1988). Unbiased estimation of flood risk with the GEV distribution. *Stochastic Hydrology and Hydraulics*, 2(3):201–212.

Athey, S., Tibshirani, J., and Wager, S. (2019). Generalized random forests. *The Annals of Statistics*, 47(2):1148–1178.

Azencott, C.-A. (2022). *Introduction au Machine Learning-2e éd.* Dunod.

Bagirov, A. M., Mahmood, A., and Barton, A. (2017). Prediction of monthly rainfall in Victoria, Australia: Clusterwise linear regression approach. *Atmospheric Research*, 188:20–29.

Balkema, A. A. and De Haan, L. (1974). Residual Life Time at Great Age. *The Annals of Probability*, 2(5):792–804.

Bücher, A., Lilienthal, J., Kinsvater, P., and Fried, R. (2021). Penalized quasi-maximum likelihood estimation for extreme value models with application to flood frequency analysis. *Extremes*, 24(2):325–348.

Bücher, A. and Segers, J. (2017). On the maximum likelihood estimator for the Generalized Extreme-Value distribution. *Extremes*, 20(4):839–872.

Beirlant, J., Goegebeur, Y., Segers, J., and Teugels, J. L. (2006). *Statistics of extremes: theory and applications*. John Wiley & Sons.

Bensalah, Y. (2000). Steps in applying extreme value theory to finance: a review.

Benziadi, F., Laksaci, A., and Tebboune, F. (2016). Recursive kernel estimate of the conditional quantile for functional ergodic data. *Communications in Statistics-Theory and Methods*, 45(11):3097–3113.

Beyerlein, A. (2014a). Quantile Regression–Opportunities and Challenges From a User's Perspective. *American Journal of Epidemiology*, 180(3):330–331.

Beyerlein, A. (2014b). Quantile regression—opportunities and challenges from a user's perspective. *American journal of epidemiology*, 180(3):330–331.

Bochenek, B. and Ustrnul, Z. (2022). Machine Learning in Weather Prediction and Climate Analyses—Applications and Perspectives. *Atmosphere*, 13(2):180. Number: 2.

Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence*. OUP Oxford.

Boudrissa, N., Cheraitia, H., and Halimi, L. (2017). Modelling maximum daily yearly rainfall in northern Algeria using generalized extreme value distributions from 1936 to 2009. *Meteorological Applications*, 24(1):114–119. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/met.1610.

Bousebata, M. (2022). *Statistical inference for extreme risk measures : Implication for the insurance of natural disasters*. Theses, Université Grenoble Alpes [2020-....]. Issue: 2022GRALM007.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24(2):123–140.

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

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). Introduction To Tree Classification. In *Classification and Regression Trees*. Routledge. Num Pages: 41.

Bremnes, J. B. (2004). Probabilistic Forecasts of Precipitation in Terms of Quantiles Using NWP Model Output. *Monthly Weather Review*, 132(1):338–347.

Buchinsky, M. (1998). Recent advances in quantile regression models: a practical guideline for empirical research. *Journal of human resources*, pages 88–126.

Buhai, S. (2005). Quantile regression: overview and selected applications. *Ad Astra*, 4(4):1–17.

Butler, J. C., Dowell, S. F., and Breiman, R. F. (1998). Epidemiology of emerging pneumococcal drug resistance: implications for treatment and prevention. *Vaccine*, 16(18):1693–1697.

Cade, B. S. and Noon, B. R. (2003). A gentle introduction to quantile regression for ecologists. *Frontiers in Ecology and the Environment*, 1(8):412–420.

Calabrese, R. and Giudici, P. (2015). Estimating bank default with generalised extreme value regression models. *The Journal of the Operational Research Society*, 66(11):1783–1792.

Cannon, A. J. (2018). Non-crossing nonlinear regression quantiles by monotone composite quantile regression neural network, with application to rainfall extremes. *Stochastic environmental research and risk assessment*, 32:3207–3225.

Cervantes, M. n. N., Benito, S., and López-Martín, C. (2024). Assessing the Performance of the Block Maxima Method in Estimating Market Risk. ISSN: 2693-5015.

Chaudhuri, P. and Loh, W.-Y. (2002). Nonparametric estimation of conditional quantiles using quantile regression trees. *Bernoulli*, pages 561–576.

Chernozhukov, V. (2005). Extremal quantile regression. *Annals of Statistics*, pages 806–839.

Chernozhukov, V. and Fernandez-Val, I. (2011). Inference for Extremal Conditional Quantile Models, with an Application to Market and Birthweight Risks. *The Review of Economic Studies*, 78(2):559–589. arXiv:0912.5013 [econ, math, q-fin, stat].

Chernozhukov, V., Fernández-Val, I., and Kaji, T. (2017). *Extremal Quantile Regression: An Overview*. arXiv:1612.06850 [econ, stat].

Chernozhukov, V., Fernández-Val, I., and Melly, B. (2020). Fast Algorithms for the Quantile Regression Process. arXiv:1909.05782 [econ, stat].

Cohen Sabban, I. (2022). *Analyse statistique de l'évolution des sinistres graves pour une garantie risque corporel*. Theses, Sorbonne Université. Issue: 2022SORUS129.

Coles, S. (2001). *An Introduction to Statistical Modeling of Extreme Values*. Springer Series in Statistics. Springer, London.

Coles, S. G. and Dixon, M. J. (1999). Likelihood-Based Inference for Extreme Value Models. *Extremes*, 2(1):5–23.

Coles, S. G. and Tawn, J. A. (1996). A Bayesian Analysis of Extreme Rainfall Data. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 45(4):463–478.

Dabo-Niang, S. and Laksaci, A. (2012). Nonparametric quantile regression estimation for functional dependent data. *Communications in statistics-Theory and methods*, 41(7):1254–1268.

Dabrowska, D. M. (1992). Nonparametric quantile regression with censored data. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 252–259.

Daouia, A., Gardes, L., and Girard, S. (2013). On kernel smoothing for extremal quantile regression. *Bernoulli*, 19(5B):2557–2589.

Davison, A. C. and Smith, R. L. (1990). Models for Exceedances Over High Thresholds. *Journal of the Royal Statistical Society: Series B (Methodological)*, 52(3):393–425. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.2517-6161.1990.tb01796.x.

De Haan, L. and Ferreira, A. (2006). *Extreme Value Theory*. Springer Series in Operations Research and Financial Engineering. Springer, New York, NY.

De Paola, F., Giugni, M., Pugliese, F., Annis, A., and Nardi, F. (2018). GEV Parameter Estimation and Stationary vs. Non-Stationary Analysis of Extreme Rainfall in African Test Cities. *Hydrology*, 5(2):28. Number: 2.

Dkengne, P. S., Girard, S., and Ahiad, S. (2020). An automatic procedure to select a block size in the continuous generalized extreme value model estimation.

Dombry, C. (2015). Existence and consistency of the maximum likelihood estimators for the extreme value index within the block maxima framework. *Bernoulli*, 21(1):420–436.

Dombry, C. and Ferreira, A. (2019). Maximum likelihood estimators based on the block maxima method. *Bernoulli*, 25(3):1690–1723.

El Methni, J. and Stupfler, G. (2017). Extreme versions of Wang risk measures and their estimation for heavy-tailed distributions. *Statistica Sinica*, pages 907–930.

Embrechts, P., Klüppelberg, C., Mikosch, T., Emberchts, P., Klüppelberg, C., and Mikosch, T. (1997). Risk theory. *Modelling Extremal Events: for Insurance and Finance*, pages 21–57.

Fakoor, R., Kim, T., Mueller, J., Smola, A. J., and Tibshirani, R. J. (2023). Flexible model aggregation for quantile regression. *Journal of Machine Learning Research*, 24(162):1–45.

Fan, J. and Li, R. (2001). Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties. *Journal of the American Statistical Association*, 96(456):1348–1360.

Farkas, S., Heranval, A., Lopez, O., and Thomas, M. (2024). Generalized pareto regression trees for extreme event analysis. *Extremes*.

Ferreira, A. and De Haan, L. (2015). On the block maxima method in extreme value theory: PWM estimators. *The Annals of Statistics*, 43(1):276–298.

Fisher, R. A. and Tippett, L. H. C. (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample. *Mathematical Proceedings of the Cambridge Philosophical Society*, 24(2):180–190.

Galton, F. (1889). *Natural Inheritance*. Macmillan. Google-Books-ID: vc8oAAAAYAAJ.

Gardes, L., Guillou, A., and Roman, C. (2020). Estimation of extreme conditional quantiles under a general tail-first-order condition. *Annals of the Institute of Statistical Mathematics*, 72(4):915–943.

Gardes, L. and Stupfler, G. (2015). Estimating extreme quantiles under random truncation. *TEST*, 24(2):207–227.

Gardes, L. and Stupfler, G. (2019). An integrated functional Weissman estimator for conditional extreme quantiles. *REVSTAT-Statistical Journal*, 17(1):109–144.

Gilli, M. and këllezi, E. (2006). An Application of Extreme Value Theory for Measuring Financial Risk. *Computational Economics*, 27(2):207–228.

Gnecco, N., Terefe, E. M., and Engelke, S. (2024). Extremal Random Forests. *Journal of the American Statistical Association*, pages 1–24.

Gnedenko, B. (1943). Sur La Distribution Limite Du Terme Maximum D'Une Serie Aleatoire. *Annals of Mathematics*, 44(3):423–453.

Gumbel, E. J. (1941). Probability-interpretation of the observed return-periods of floods. *Transactions, American Geophysical Union*, 22(3):836.

Gumbel, E. J. (1963). Statistical Forecast of Droughts. *International Association of Scientific Hydrology. Bulletin*, 8(1):5–23. _eprint: https://doi.org/10.1080/02626666309493293.

Hallock, K. F. and Koenker, R. (2001). Quantile regression. *The Journal of Economic Perspectives*, 15(4):143.

Halton, J. H. (1964). Algorithm 247: Radical-inverse quasi-random point sequence. *Communications of the ACM*, 7(12):701–702.

Hastie, T. (2017). *The Elements of Statistical Learning*, volume 2 of *Springer Series in Statistics*. Springer, New York, NY, second edition edition.

Hill, B. M. (1975). A simple general approach to inference about the tail of a distribution. *The annals of statistics*, pages 1163–1174.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, 12(1):55–67.

Hu, G. and Franzke, C. L. E. (2020). Evaluation of Daily Precipitation Extremes in Reanalysis and Gridded Observation-Based Data Sets Over Germany. *Geophysical Research Letters*, 47(18):e2020GL089624. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1029/2020GL089624.

Jenkinson, A. F. (1955). The frequency distribution of the annual maximum (or minimum) values of meteorological elements. *Quarterly Journal of the Royal Meteorological Society*, 81(348):158–171. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/qj.49708134804.

Kallenberg, O. (2002). *Foundations of Modern Probability*. Probability and Its Applications. Springer, New York, NY.

Katz, R. W., Parlange, M. B., and Naveau, P. (2002). Statistics of extremes in hydrology. *Advances in Water Resources*, 25(8-12):1287–1304.

Kithinji, M. M., Mwita, P. N., and Kube, A. O. (2021). Adjusted Extreme Conditional Quantile Autoregression with Application to Risk Measurement. *Journal of Probability and Statistics*, 2021:1–10.

Koenker, R. (2011). Additive models for quantile regression: Model selection and confidence bandaids. *Brazilian Journal of Probability and Statistics*, 25(3):239–262.

Koenker, R. (2017). Quantile regression 40 years on. Technical report, The IFS.

Koenker, R. and Bassett, G. (1978). Regression Quantiles. *Econometrica*, 46(1):33.

Koenker, R. and Hallock, K. F. (2001). Quantile regression. *Journal of economic perspectives*, 15(4):143–156.

Laksaci, A. and Maref, F. (2009). Estimation non paramétrique de quantiles conditionnels pour des variables fonctionnelles spatialement dépendantes. *Comptes Rendus Mathematique*, 347(17-18):1075–1080.

Le, Q. V., Sears, T., and Smola, A. J. (2005). Nonparametric quantile regression. Technical report, Technical report, National ICT Australia, June 2005. Available at http://sml ....

Leadbetter, M. R. (1991). On a basis for 'Peaks over Threshold' modeling. *Statistics & Probability Letters*, 12(4):357–362.

Lin, Y. and Jeon, Y. (2006). Random forests and adaptive nearest neighbors. *Journal of the American Statistical Association*, 101(474):578–590.

Meinshausen, N. and Ridgeway, G. (2006). Quantile regression forests. *Journal of machine learning research*, 7(6).

Pasche, O. C. and Engelke, S. (2023). Neural Networks for Extreme Quantile Regression with an Application to Forecasting of Flood Risk. arXiv:2208.07590 [stat] version: 2.

Pasche, O. C. and Engelke, S. (2024). Neural networks for extreme quantile regression with an application to forecasting of flood risk. *The Annals of Applied Statistics*, 18(4):2818–2839.

Pickands III, J. (1975). Statistical inference using extreme order statistics. *the Annals of Statistics*, pages 119–131.

Reiss, R.-D., Thomas, M., and Reiss, R. D. (1997). *Statistical analysis of extreme values*, volume 2. Springer.

Resnick, S. I. (1987). *Extreme Values, Regular Variation and Point Processes*. Springer Series in Operations Research and Financial Engineering. Springer, New York, NY.

Resnick, S. I. (2007). *Heavy-Tail Phenomena*. Springer Series in Operations Research and Financial Engineering. Springer, New York, NY.

Saulo, H., Vila, R., Bittencourt, V. L., Leão, J., Leiva, V., and Christakos, G. (2022). On a new extreme value distribution: characterization, parametric quantile regression, and application to extreme air pollution events. *Stochastic Environmental Research and Risk Assessment*, pages 1–18.

Schaumburg, J. (2012). Predicting extreme value at risk: Nonparametric quantile regression with refinements from extreme value theory. *Computational Statistics & Data Analysis*, 56(12):4081–4096.

Sielenou, D. and Alain, P. (2020). FCwx.csv. 1.

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

Takeuchi, I., Le, Q., Sears, T., and Smola, A. (2006). Nonparametric quantile estimation.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288.

Tyralis, H., Papacharalampous, G., Burnetas, A., and Langousis, A. (2019). Hydrological post-processing using stacked generalization of quantile regression algorithms: Large-scale application over CONUS. *Journal of Hydrology*, 577:123957.

Van Der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge.

Vapnik, V. N. (2000). *The Nature of Statistical Learning Theory*. Springer, New York, NY.

Velthoen, J., Dombry, C., Cai, J.-J., and Engelke, S. (2023). Gradient boosting for extreme quantile regression. *Extremes*, 26(4):639–667.

Vidagbandji, L. M., Berred, A., Bertelle, C., and Amanton, L. (2025). Generalized random forest for extreme quantile regression. *Communications in Statistics - Simulation and Computation*, 0(0):1–24. _eprint: https://www.tandfonline.com/doi/pdf/10.1080/03610918.2025.2543854.

Vidagbandji, L. M., Berred, A., Bertelle, C., and Amanton, L. (2026). Penalized estimation of GEV parameters for extreme quantile regression. arXiv:2603.24153 [math].

Von Mises, R. (1936). La distribution de la plus grande de n valuers. *Rev. math. Union interbalcanique*, 1:141–160.

Wager, S. and Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242.

Wang, H. and Tsai, C.-L. (2009). Tail index regression. *Journal of the American Statistical Association*, 104(487):1233–1240.

Wang, H. J. and Li, D. (2013). Estimation of Extreme Conditional Quantiles Through Power Transformation. *Journal of the American Statistical Association*, 108(503):1062–1074.

Wang, H. J., Li, D., and He, X. (2012). Estimation of High Conditional Quantiles for Heavy-Tailed Distributions. *Journal of the American Statistical Association*, 107(500):1453–1464.

Wang, J., You, S., Wu, Y., Zhang, Y., and Bin, S. (2016). A Method of Selecting the Block Size of BMM for Estimating Extreme Loads in Engineering Vehicles. *Mathematical Problems in Engineering*, pages 1–9.

Yao, L., Lu, J., Zhang, W., Qin, J., Zhou, C., Tran, N. N., and Pinagé, E. R. (2022). Spatiotemporal Analysis of Extreme Temperature Change on the Tibetan Plateau Based On Quantile Regression. *Earth and Space Science*, 9(11):e2022EA002571.

Ye, S. S. and Padilla, O. H. M. (2020). Non-parametric Quantile Regression via the K-NN Fused Lasso. *Journal of Machine Learning Research*, 22:111:1–111:38.

Youngman, B. D. (2019). Generalized Additive Models for Exceedances of High Thresholds With an Application to Return Level Estimation for U.S. Wind Gusts. *Journal of the American Statistical Association*, 114(528):1865–1879.

Youngman, B. D. (2022). evgam: An R package for generalized additive extreme value models. *Journal of Statistical Software*, 103:1–26.

Yu, K. and Moyeed, R. A. (2001). Bayesian quantile regression. *Statistics & Probability Letters*, 54(4):437–447.

Özarı, i., Eren, z., and Erdoğan, E. (2018). A PROPOSAL METHOD TO SELECT THE OPTIMAL BLOCK SIZE: AN APPLICATION ON FINANCIAL MARKETS.

Özari, i., Eren, z., and Saygin, H. (2019). A new methodology for the block maxima approach in selecting the optimal block size. *Tehnički vjesnik*, 26(5):1292–1296.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2):894–942.

Zheng, W., Peng, X., Lu, D., Zhang, D., Liu, Y., Lin, Z., and Lin, L. (2017). Composite quantile regression extreme learning machine with feature selection for short-term wind speed forecasting: A new approach. *Energy Conversion and Management*, 151:737–752.

Zhou, C. (2009). Existence and consistency of the maximum likelihood estimator for the extreme value index. *Journal of Multivariate Analysis*, 100(4):794–815.

Zhu, H., Li, Y., Liu, B., Yao, W., and Zhang, R. (2022). Extreme quantile estimation for partial functional linear regression models with heavy-tailed distributions. *Canadian Journal of Statistics*, 50(1):267–286.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(2):301–320.

Résumé

Les méthodes classiques d'estimation des quantiles extrêmes présentent plusieurs limites : instabilité dans la queue de distribution, faible flexibilité en présence de covariables multidimensionnelles et difficulté des approches non paramétriques à capturer adéquatement les comportements asymptotiques. Pour répondre à ces défis, cette thèse développe de nouvelles méthodes de régression quantile extrême alliant de manière cohérente la théorie des valeurs extrêmes et des techniques modernes d'apprentissage statistique. La première contribution propose un estimateur du quantile conditionnel extrême fondé sur une version pondérée du maximum de vraisemblance pour la distribution GEV conditionnelle. Les poids issus des forêts aléatoires généralisées permettent de mieux capturer les relations non linéaires et les interactions complexes entre les covariables, tout en atténuant les effets liés à la forte dimensionnalité. L'existence et la convergence de l'estimateur sont établies, mettant en évidence son intérêt pour des quantiles élevés et des covariables de grande dimension. La seconde contribution introduit une pénalisation de type $L2$ sur l'indice de forme ξ, améliorant la stabilité de l'estimation des quantiles extrêmes. Enfin, une fonction de pénalité spécifiquement adaptée à la distribution GEV est proposée pour stabiliser davantage l'estimation de l'indice des valeurs extrêmes. Cette approche améliore les performances en petits échantillons tout en restant efficace sur de grands jeux de données. Les comparaisons avec des méthodes classiques d'apprentissage statistique montrent des gains substantiels en précision, stabilité et robustesse, confirmés par une application à des données salariales américaines de 1980.

Mots-clés : Distribution des valeurs extrêmes généralisée, régression quantile extrême, forêt aléatoire généralisée, estimateur du maximum de vraisemblance, méthode des maxima de blocs.

Abstract

Classical methods for estimating extreme quantiles present several limitations: instability in the tail of the distribution, limited flexibility in the presence of multidimensional covariates, and the difficulty of nonparametric approaches in adequately capturing asymptotic behaviors. To address these challenges, this thesis develops new extreme quantile regression methods that coherently combine extreme value theory with modern statistical learning techniques. The first contribution proposes an estimator of the conditional extreme quantile based on a weighted version of the maximum likelihood estimator for the conditional GEV distribution. The weights derived from generalized random forests allow for better capture of nonlinear relationships and complex interactions between covariates, while mitigating the effects related to high dimensionality. The existence and convergence of the estimator are established, highlighting its relevance for high quantiles and high-dimensional covariates. The second contribution introduces an $L2$-type penalization on the shape index ξ, improving the stability of extreme quantile estimation. Finally, a penalty function specifically adapted to the GEV distribution is introduced to further stabilize the estimation of the extreme value index. This approach improves performance in small samples while remaining effective for large datasets. Comparisons with classical statistical learning methods show substantial gains in accuracy, stability, and robustness, confirmed by an application to U.S. wage data from 1980.

Keywords: Generalized extreme value distribution, extreme quantile regression, generalized random forest, maximum likelihood estimator, block maxima method.