Université du Québec
Institut national de la recherche scientifique
Centre Eau Terre Environnement

MÉTHODES D'ESTIMATION DES QUANTILES CONDITIONNELS EN HYDRO-CLIMATOLOGIE

Par

BOUCHRA NASRI

Thèse présentée pour l'obtention du grade de
de Philosophiae Doctor (Ph.D.)
en sciences de l'eau

Jury d'évaluation

Examinateur externe	Fahim Ashkar Université de Moncton
Examinateur externe	Mhamed Mesfioui Université du Québec à Trois-Rivières
Examinateur interne	Sophie Duchesne INRS-ETE
Co-directeur de recherche	Taoufik Bouezmarni Université de Sherbrooke
Co-directeur de recherche	Taha B.M.J. Ouarda INRS-ETE
Directeur de recherche	André St-Hilaire INRS-ETE

Remerciements

Soyons reconnaissants aux personnes qui nous donnent du bonheur; elles sont les charmants jardiniers par qui nos âmes sont fleuries.

Marcel Proust

Je tiens tout d'abord à remercier mon directeur de thèse, André St-Hilaire. Les mots me manquent pour exprimer ma gratitude. L'idée de voler de mes propres ailes est un peu effrayante, mais j'ai l'impression d'avoir grandi, d'avoir acquis une certaine confiance grâce à toi. J'ai énormément appris à tes côtés. Merci aussi pour toutes les révisions, les conseils, les congrès, les séminaires, etc. Bref, merci pour tout. Je remercie également mon co-directeur Taha B.M.J Ouarda, pour sa gentillesse, ses révisions et surtout de m'avoir permis de réaliser mon travail de paillasse dans les meilleures conditions possibles. Merci encore pour tout. Finalement, mes remerciements vont à mon co-directeur Taoufik Bouezmarni. Son aide et ses conseils avisés m'ont permis d'élargir mes connaissances en statistique et surtout en développement théorique. Merci pour tout.

J'exprime mes remerciements aussi à l'ensemble des membres de mon jury : madame Sophie Duchesne, monsieur Fahim Ashkar et monsieur Mhamed Mesfoui.

Je remercie toutes les personnes formidables que j'ai rencontrées par le biais de l'INRS, de l'université de Sherbrooke et de l'Association des statisticiennes et des statisticiens du Québec. Merci pour votre soutien et vos encouragements.

J'adresse toute ma gratitude à mes amis, plus spécialement, Sylvain, Yona, Gaêl, Martin, Malika, Alida, Fatiha et Hind pour les bons moments et le grand soutien moral.

Enfin, les mots les plus simples étant les plus forts, j'adresse toute mon affection à ma famille, et en particulier à ma maman. Malgré mon éloignement depuis de nombreuses années, leur intelligence, leur confiance, leur tendresse et leur amour me portent et me guident tous les jours. Merci pour avoir fait de moi ce que je suis aujourd'hui. Je vous aime.

Une pensée pour terminer ces remerciements pour toi, mon papa, qui n'a pas vu l'aboutissement de mon travail, mais je sais que tu aurais été très fier de ta fille !

Résumé

La vie n'est bonne qu'à étudier et à enseigner les mathématiques.

De Blaise Pascal

Cette thèse de doctorat a pour objectif de développer de nouvelles méthodes pour l'estimation des quantiles des événements hydrologiques extrêmes en présence des covariables climatiques (donc, il s'agit de l'estimation des quantiles conditionnels) dans le cas où la dépendance entre la variable d'intérêt et les covariables est de type non nécessairement linéaire ou inconnu. Dans cette thèse, nous proposons trois approches pour l'estimation des quantiles conditionnels : deux sont basées sur des dépendances de type B-Splines décrivant soit la relation entre les paramètres d'une fonction de répartition d'une variable d'intérêt et les covariables, soit le lien entre la variable d'intérêt et les covariables sous un modèle de régression des quantiles et une approche basée sur des dépendances de type copule décrivant le lien entre une fonction de répartition d'une variable d'intérêt et les covariables. Ces trois approches ont été tout d'abord comparées avec les approches classiques qui reposent sur des dépendances linéaires ou quadratiques et ensuite comparées entre elles afin de déterminer le meilleur estimateur. Aussi, elles ont été appliquées sur des bases de données hydro-climatiques afin d'estimer le risque des extrêmes dans certaines régions du monde, plus spécifiquement le nord de l'Afrique et le nord-est du Canada. Les résultats de nos travaux ont montré le grand avantage de nos approches par rapport aux méthodes classiques. En plus, l'estimation des quantiles conditionnels basée sur les copules a montré une grande performance par rapport aux

deux autres approches basées sur les fonctions B-Splines. Cette performance a été démontrée en se basant sur des simulations considérant différents modèles statistiques. En effet, ajouter un modèle de dépendance, par exemple une copule, aux modèles des quantiles conditionnels permet de capturer la structure de dépendance globale entre la variable d'intérêt et les covariables. Par conséquent, cela permet de diminuer largement le biais d'estimation, ce qui donne des estimations de risque de plus en plus précises pour la gestion des événements hydrologiques ou climatiques extrêmes.

Mots-clés: quantiles conditionnels, covariables, B-Splines, copules, hydro-climatologie.

Abstract

This PhD thesis aims to develop new methods for estimating the quantiles of extreme hydrological events in the presence of climatic covariates (i.e., the estimation of conditional quantiles) in the context where the dependence between the variable of interest and covariates is not necessarily linear or known. In this thesis, we propose three approaches for estimating the conditional quantiles: two approaches are based on the B-Splines functions which describe either the relationship between the parameters of a cumulative distribution function of a variable of interest and the covariables, or the link between the variable of interest and the covariates under a quantile regression model and an approach based on copula functions which describe the link between a cumulative distribution function of a variable of interest and the covariables. We first compared these three approaches with more classical non stationary approaches which are based on linear or quadratic dependency models and then compared the three proposed approaches to each other in order to determine the best estimators. Also, they have been applied to case studies where hydro-climatic data are used to estimate the risk of extremes in some regions of the world, specifically northern Africa and eastern Canada. The results of our work have shown the great advantage of using our approaches compared to the classical approaches and they showed the performance of estimating the conditional quantiles based on the copula approach. This performance is demonstrated by simulations which consider different statistical models. Indeed, using copula function to estimate conditional quantile models allows to capture the overall dependence structure between the variable of interest and covariates and then

provides an important estimation improvement.

Keywords: Conditional quantiles, covariates, B-Splines, copulas, hydro-climatology.

Table des matières

Liste des figures

Liste des tableaux

Première partie

Synthèse

Chapitre 1

Introduction

1.1 Introduction

Les débits dans un cours d'eau présentent des fluctuations saisonnières importantes pouvant avoir des impacts majeurs sur l'environnement et la population. En période de sécheresse, les débits d'étiage peuvent diminuer la capacité de dilution des polluants et augmenter la température de l'eau, ce qui a des conséquences dommageables sur les habitats aquatiques et la qualité de l'eau dans les bassins versants. Les débits importants qui résultent des fortes précipitations ou d'une fonte rapide de la neige sont, quant à eux, aptes à provoquer des crues, et dans le cas extrême, des inondations. Ces conséquences ont des impacts majeurs sur la population et les infrastructures. Les débits de crue et d'étiage ont une importance primordiale en hydrologie et leur impact socio-économique est déterminant. Ils représentent des variables clés pour les prévisions des risques, le dimensionnement des ouvrages (ponts, barrage, etc.), l'approvisionnement et la gestion des ressources (eau, électricité, etc.), l'occupation des sols et la stabilité des milieux aquatiques. La cause principale de la grande variabilité des débits d'une année à une autre demeure les aléas climatiques. En effet, le réchauffement du système climatique est sans équivoque. Ce qui n'était encore qu'une hypothèse

il y a quelques décennies semble aujourd'hui un fait incontestable. L'augmentation évidente de la concentration des gaz à effet de serre dans l'atmosphère a engendré des répercussions sur un certain nombre de variables climatiques. Les changements climatiques (CC) pourraient se traduire par l'augmentation ou la diminution des amplitudes et/ou des fréquences des évènements extrêmes. Les CC pourraient même amener des changements dans les conditions hydrométéorologiques qui conduisent à des évènements extrêmes. Une vaste revue sur l'impact possible des CC est trouvée dans le rapport d'évaluation du groupe intergouvernemental sur l'évolution du climat [GIEC, 2007] et dans [Bates *et al.*, 2008]. De nombreuses études ont conclu à l'existence d'une tendance temporelle et/ou à la non-stationnarité dans les séries chronologiques des débits ou des précipitations dans différentes régions du monde [e.g., Déry & Wood, 2005; Ehsanzadeh & Adamowski, 2007; Ouarda & Adlouni, 2011; Cunderlik & Burn, 2002]. Comprendre la variabilité temporelle des processus hydrologiques et leurs statistiques connexes, dans ce contexte, est essentiel pour une meilleure gestion des ressources en eau.

1.2 Problématique

L'estimation des quantiles des évènements extrêmes a fait l'objet de plusieurs études scientifiques [e.g., Olsen *et al.*, 1999; Vrac & Naveau, 2007; Cannon, 2011]. L'analyse fréquentielle (AF) est une des méthodes statistiques pour la modélisation des quantiles des extrêmes. Elle repose sur la définition et la mise en œuvre d'un modèle fréquentiel, qui se résume dans la description du comportement statistique d'un processus. Une des méthodes les plus utilisées dans la littérature est l'estimation des quantiles inconditionnels, qui est aussi appelée l'AF classique. Cette méthode a pour but l'estimation des quantiles par le biais de l'estimation de la fonction de répartition. La fonction des quantiles, par la suite, n'est que l'inverse de cette dernière. Des méthodes paramétriques et

non-paramétriques ont été utilisées, en général, pour estimer une fonction de répartition. En ce qui concerne l'estimation paramétrique, la fonction de répartition est supposée appartenir à une famille de distribution avec des paramètres connus (par exemple, gamma, lognormale, Weibull, etc.). Par ailleurs, pour l'estimation non paramétrique, il existe plusieurs méthodes d'estimation de la fonction de répartition, telles que la fonction de répartition empirique [e.g., Van der Vaart, 1998] et les fonctions de distribution basées sur les fonctions de type noyaux (exemple: Kernel, Bernstein etc.) [e.g., Yamato, 1973]. L'estimation des quantiles inconditionnels a fait l'objet de plusieurs études en hydrologie dans les années 80 et 90 [e.g., Buishand, 1984, 1989, 1991; Carter & Challenor, 1981; Cunnane, 1989; Grehys, 1996; Madsen *et al.*, 1997; Lang *et al.*, 1999]. L'estimation des quantiles, dans ce cadre, est basée sur l'hypothèse que les observations doivent être indépendantes et identiquement distribuées (i.i.d). Autrement dit, la distribution de probabilité de la variable d'intérêt ne doit pas changer dans le temps (stationnaire). Par contre, comme il est indiqué dans l'introduction et à cause des CC, les distributions des évènements hydro-climatiques semblent changer dans le temps. La modélisation des quantiles, dans ce sens, peut se faire en introduisant dans le modèle le temps ou des covariables pour bien expliquer ces changements, ce qui mène à la mise en place des approches pour l'estimation des quantiles conditionnels à des covariables. Dans ce cadre, deux approches peuvent être utilisées:

— La première approche repose sur l'utilisation de la fonction de répartition en introduisant les covariables dans les paramètres de la distribution, ce qui donne une distribution de probabilité avec des paramètres non constants. Cette approche a été largement utilisée dans la littérature, surtout dans le cadre où la dépendance entre la variable d'intérêt et les covariables est sous forme d'une fonction polynomiale (linéaire ou quadratique) [e.g., Coles, 2001; Aissaoui-Fqayeh *et al.*, 2009; El Adlouni & Ouarda, 2009; Cannon, 2011].

— La deuxième approche consiste plutôt à l'utilisation de la régression des quantiles [Koenker & Bassett, 1987]. Contrairement à la régression ordinaire qui fournit une estimation de la moyenne conditionnelle de la variable d'intérêt étant donné certaines valeurs des covariables, la régression des quantiles estime les quantiles conditionnels à des valeurs des covariables. Récemment, certaines études en hydro-climatologie ont porté sur l'utilisation de cette approche pour l'estimation des quantiles conditionnels [e.g., Friederichs & Hense, 2007; Tareghian & Rasmussen, 2013]. Pourtant, dans ces travaux, les auteurs sont limités à l'utilisation du modèle de régression des quantiles linéaires.

Bien que les modèles linéaires, dans les deux approches, soient largement utilisés pour leur simplicité, il reste que leur domaine d'applicabilité demeure très restreint. En effet, généralement la structure de la dépendance entre les variables d'intérêt et les covariables en hydro-climatologie est plus complexe. On retrouve rarement de fortes dépendances linéaires entre ces variables et généralement la dépendance est de type non-linéaire ou inconnue. Dans ce cas, l'utilisation des modèles linéaires peut engendrer un biais considérable.

1.3 Objectifs

Le but de cette thèse de doctorat est la généralisation des approches citées dans la problématique dans le cadre où la dépendance entre la variable d'intérêt et les covariables n'est pas nécessairement linéaire ou connue.

Les objectifs spécifiques du projet de recherche sont les suivants :

— Utilisation des fonctions de lissage (B-Splines) afin de décrire la dépendance entre la variable d'intérêt et les covariables dans l'approche de l'inverse de fonction de répartition. La fonction

de répartition utilisée ici est la distribution généralisée des valeurs extrêmes (Generalized extreme value, GEV).

— Utilisation des fonctions de lissage (B-Splines) dans le cadre du modèle de la régression des quantiles.

— Utilisation des fonctions de dépendance (copules) dans l'approche de l'inverse de fonction de répartition. La fonction copule décrit la forme générale de la dépendance entre la variable d'intérêt et les covariables.

1.4 Organisation du document

Cette thèse de recherche est divisée en deux parties:

— La partie I contiendra, outre le chapitre 1, le chapitre 2 qui sera composé de deux sections. La premiere section présentera une revue de littérature sur les extrêmes en hydrologie, la deuxième section contiendra une revue de littérature sur les méthodes classiques d'estimation des quantiles des évènements hydrologiques extrêmes. La méthodologie proposée et ses détails seront l'objectif du chapitre 3. Le chapitre 4 sera consacré aux résultats et conclusions.

— La partie II présentera les articles issus de cette thèse de recherche.

Chapitre 2

Revue de littérature

2.1 Extrêmes hydrologiques

2.1.1 Crues

Une crue est caractérisée par l'augmentation significative du niveau d'eau en écoulement dans un cours d'eau. Cette augmentation est souvent attribuable aux apports provenant de la fonte de la neige (crues printanières) ou de fortes précipitations sous forme liquide (e.g. crues d'automne au Québec). Une crue peut se définir par plusieurs paramètres (débit de pointe, durée, temps de montée, volume, etc.). Le débit de pointe correspond au débit maximum d'une crue. Les valeurs annuelles des débits de pointe sont généralement utilisées pour définir le classement des débits. Une probabilité est attribuée à chaque valeur de débit (en se basant sur des modélisations statistiques) [Meylan *et al.*, 2012]. Un débit de crue de récurrence de 2 ans (correspond à une probabilité de 0.5, i.e. médiane) représente un débit qui se répète en moyenne une fois aux deux ans. Le temps de montée définit la nature de crue. En effet, lorsque le temps de montée d'une crue est très court, par exemple inférieur à quelques heures, on parle d'une crue-éclair. Dans le cas où le temps de montée

est compris entre 2 h et 12 h, on parle d'une crue rapide. Dans le cas où le temps de montée dépasse 12 h on parle d'une crue lente [e.g., Anctil *et al.*, 2012; Roche *et al.*, 2012].

2.1.2 Étiages

Les définitions du mot « étiage » diffèrent selon les domaines scientifiques (hydrologie, écologie, agriculture, etc.). Abi-Zeid & Bobée [1999] ont proposé plusieurs définitions de l'étiage, dépendamment de l'utilisation et du domaine de l'application. En général, un étiage représente un évènement extrême qui se caractérise par une pénurie d'eau sur une période de temps significative. En hydrologie, l'étiage se définit comme une baisse prolongée des eaux de surface ou encore des eaux souterraines due à des températures élevées (évapotranspiration) et / ou un manque de précipitations [Smakhtin, 2001]. Contrairement aux crues qui sont caractérisées uniquement par une récurrence, les débits d'étiage peuvent être caractérisés par une récurrence et une durée. Un débit d'étiage de récurrence de 2 ans et de durée de 7 jours représente un débit faible qui se répète en moyenne chaque deux ans sur 7 jours consécutifs.

2.2 Estimation des quantiles pour les extrêmes hydrologiques

Dans cette section, une revue de littérature sur les modèles statistiques pour l'estimation des quantiles sera présentée. Cette section est divisée en deux parties. La première partie porte sur la modélisation des quantiles en utilisant l'inverse de la fonction de répartition et la deuxième partie est consacrée à la régression des quantiles.

2.2.1 Quantiles basés sur l'inverse de la fonction de répartition

La fonction quantile d'une variable aléatoire est l'inverse de sa fonction de répartition. Soit Y une variable aléatoire à valeurs dans $\mathbb{R}$, et F_Y sa fonction de répartition. On appelle fonction quantile de Y la fonction, notée Q_Y, de $]0,1[$ dans $\mathbb{R}$, qui associe à $p \in]0,1[$ (p est la probabilité du quantile):

$$Q_Y(p) \equiv F_Y^{-1}(p) = \inf\{y : F_Y(y) \geq p\}, \tag{2.1}$$

avec F_Y^{-1} est l'inverse de la fonction de répartition F_Y.

L'estimation de la fonction quantile, Q_Y, repose sur l'estimation de la fonction de répartition F_Y. Cette dernière peut être estimée d'une façon non paramétrique, en utilisant la fonction empirique ou les méthodes de lissage, par exemple la méthode à noyaux ou celle basée sur les polynômes de Bernstein. Par contre, en hydro-climatologie, l'estimation de la fonction de répartition est souvent faite par le biais d'une fonction de répartition paramétrique. Plusieurs distributions paramétriques peuvent être employées dans ce cadre, la log-normale, la gamma, la GEV, les lois de Halphen, etc. Bien que le choix des distributions paramétriques soit vaste, les distributions les plus utilisées pour la modélisation des extrêmes hydrologiques se résument aux distributions résultantes de la théorie des valeurs extrêmes (TVE). En effet, la TVE se base sur la description du comportement asymptotique des valeurs extrêmes. Plus formellement, considérons $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)$ un vecteur de n variables aléatoires i.i.d de fonction de répartition F_{Y_i} définie par:

$$F_{Y_i}(y_i) = P(Y_i \leq y_i) \qquad i = 1 \ldots n. \tag{2.2}$$

Pour approfondir le comportement des évènements extrêmes, on considère la variable aléatoire $\boldsymbol{M_n} = \max(Y_1, Y_2, \ldots, Y_n)$ [1]. Comme les variables aléatoires Y_i sont i.i.d, alors la fonction de répartition de $\boldsymbol{M_n}$ est définie par:

$$F_{\boldsymbol{M_n}}(y) = P(\boldsymbol{M_n} \leq y) = (F(y))^n. \tag{2.3}$$

Dans la pratique, il est difficile de calculer la fonction de répartition dans la formule (2.3). Le théorème de Fisher & Tippett [1928] donne une solution asymptotique pour le calcul de cette fonction de répartition. S'il existe deux suites de constantes $a_n > 0$ et $b_n \in \mathbb{R}$ et une distribution non dégénérée [2] G telle que:

$$\lim_{n\to\infty} P\left\{\left(\frac{\boldsymbol{M_n} - \boldsymbol{a_n}}{b_n}\right) \leq y\right\} \to G(y) \tag{2.4}$$

Alors G est de la forme:

$$G_{\mu,\sigma,\xi}(y) = \begin{cases} \exp\left[-\left(1+\xi\left(\frac{y-\mu}{\sigma}\right)_+\right)^{-\frac{1}{\xi}}\right] & \text{si} \quad \xi \neq 0 \\ \exp\left[-\exp\left(-\left(\frac{x-\mu}{\sigma}\right)_+\right)\right] & \text{si} \quad \xi = 0, \end{cases} \tag{2.5}$$

où $y_+ = \max(0, y)$ et μ, σ et ξ sont respectivement les paramètres de position (ou location), de dispersion (ou d'échelle) et de forme de la GEV. La distribution G s'appelle la loi généralisée des valeurs extrêmes (GEV: Generalized Extreme Value) et représente la première distribution issue de la TVE. La figure 2.1 représente la fonction de densité de la loi GEV dépendamment du paramètre de forme : le cas $\xi = 0$ correspond à la loi Gumbel, $\xi > 0$ à la loi de Fréchet et $\xi < 0$ à la loi Weibull.

1. Tous les résultats développés par le maximum peuvent être transposés pour le minimum en utilisant la formule mathématique suivante: $\min(Y) = -\max(-Y)$.

2. Non centrée sur une valeur.

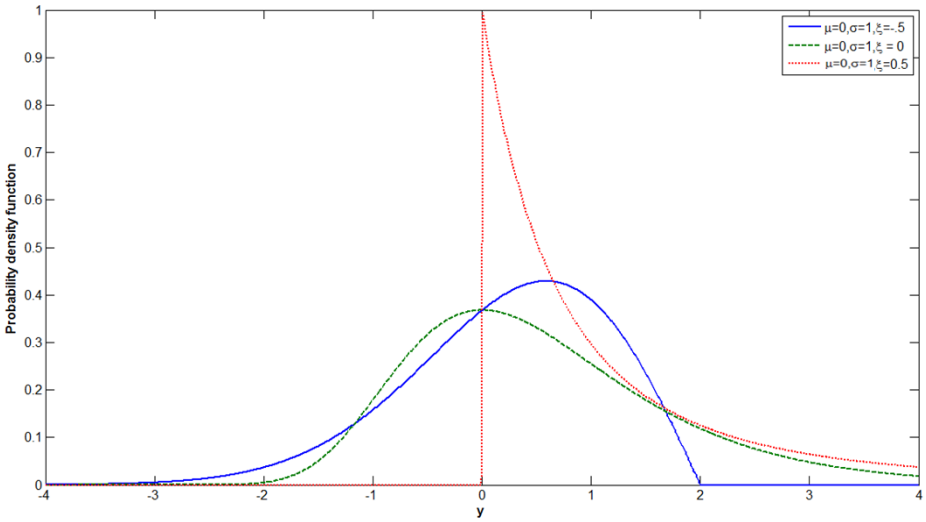

Figure 2.1 – Illustration de la fonction de densité de la loi GEV selon différentes valeurs de ξ.

L'approche basée sur la GEV a été critiquée dans la mesure où l'utilisation d'un seul maximum par année conduit à une perte d'information contenue dans les grandes valeurs observées dans un échantillon d'une variable aléatoire. Pour surmonter ce problème, Pickands [1975] a proposé la méthode des séries de durée partielle, ou excès au-delà d'un seuil (POT : Peaks over thresholds). La méthode POT consiste à utiliser, non seulement un maximum par variable aléatoire [1], mais toutes les observations qui dépassent un certain seuil prédéfini et plus particulièrement les différences entre ces observations et le seuil fixé. Formellement, considérons $u \in \mathbb{R}$, $N_u = card\{i;\ i = 1, \ldots, n;\ Y_i > u\}$ et $Z_j = Y_j - u > 0 \quad$ pour tout $j = 1, \ldots, N_u$ où N_u représente le nombre des dépassements après le seuil u et Z_j sont les nouvelles variables. Le but, ici, est de définir à partir de la loi des Y_i, $i = 1, \ldots, n$, une loi conditionnelle par rapport au seuil u pour les variables Y_i, $j = 1, \ldots, N_u$ qui est définie par:

$$F_u(z) = F(Y - u \leq z | Y > u) = \frac{F(u+z) - F(u)}{1 - F(u)}; \quad z > 0. \tag{2.6}$$

1. Nommé aussi, maximum par bloc

Pickands [1975] a proposé le résultat limite à la loi F_u. Ce résultat affirme que si F appartient à l'un des trois domaines d'attraction de la loi des valeurs extrêmes (Fréchet, Gumbel ou Weibull) et lorsque le seuil u tend vers le point terminal z_F, alors il existe une fonction $\sigma(u)$ strictement positive et un réel ξ tels que

$$\lim_{u \to z_F} \sup_{0 \leq z \leq z_F - u} \left| F_u(z) - H_{\sigma(u),\xi}(z) \right| = 0 \tag{2.7}$$

où $H_{\sigma(u),\xi}$ est la fonction de répartition de la loi de Pareto Généralisée (GPD: Generalized Pareto Distribution) et F_u est la fonction de répartition des excès au-delà du seuil u. Ainsi, pour u grand, la loi des excès est approchée par une loi GPD.

$$F_u \approx H_{\sigma(u),\xi}$$

La distribution généralisée de Pareto s'écrit sous la forme :

$$H_{\sigma(u),\xi}(z) = \begin{cases} 1 - \left(1 + \xi \dfrac{z}{\sigma(u)}\right)^{-1/\xi} & \text{si} \quad \xi \neq 0 \\ 1 - \exp\left(-\dfrac{z}{\sigma(u)}\right) & \text{si} \quad \xi = 0, \end{cases} \tag{2.8}$$

où $z \geq 0$ si $\xi \geq 0$ et $0 \leq z \leq \frac{-\sigma(u)}{\xi}$ si $\xi < 0$. σ est le paramètre d'échelle et ξ est le paramètre de forme.

La figure 2.2 montre une illustration de la fonction de densité de la loi de GPD pour différentes valeurs de ξ et de σ:

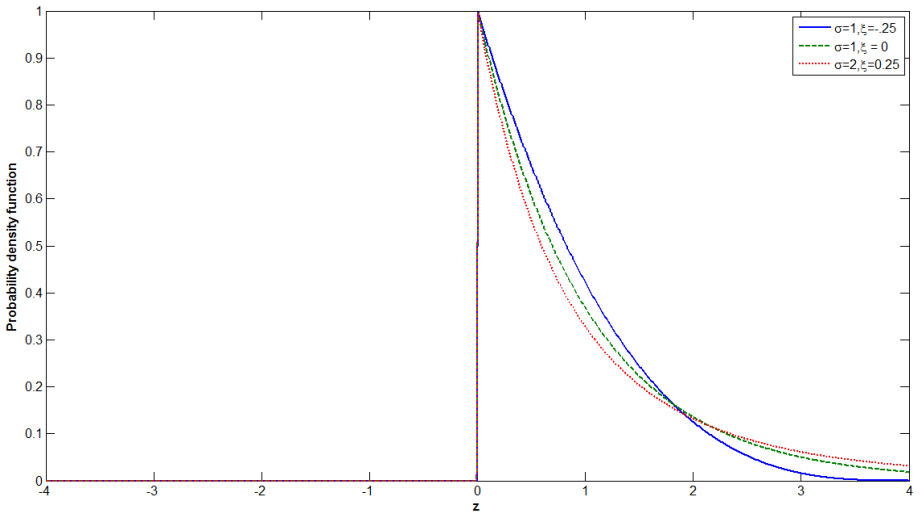

Figure 2.2 – Illustration de la fonction de densité de la distribution GPD pour différentes valeurs de σ et ξ

Plusieurs méthodes ont été fournies pour l'estimation des paramètres de la loi GEV et de la loi GPD. On trouve la méthode du maximum de vraisemblance [Smith, 1985], la méthode des moments [Christopeit, 1994] et la méthode bayésienne [Christopeit, 1994]. Des estimateurs non paramétriques ont été aussi développés, comme l'estimateur de Pickands [1975] et l'estimateur de Hill [1975]. L'estimation des quantiles inconditionnels (stationnaires) pour les lois GEV et GPD est donnée par les formules suivantes:

$$\widehat{Q}_Y^{GEV}(p) = G^{-1}_{\mu,\sigma,\xi}(p) = \begin{cases} \widehat{\mu} - \dfrac{\widehat{\sigma}}{\widehat{\xi}}\left[1 - (-\log(p))^{-\widehat{\xi}}\right] & \text{si} \quad \xi \neq 0 \\ \widehat{\mu} - \widehat{\sigma}\log(-\log(p)) & \text{si} \quad \xi = 0 \end{cases} \tag{2.9}$$

$$\widehat{Q}_Y^{GPD}(p) = H^{-1}_{\sigma(u),\xi}(p) = \begin{cases} u - \dfrac{\widehat{\sigma(u)}}{\widehat{\xi}}\left(1 - p^{\widehat{\xi}}\right) & \text{si} \quad \xi \neq 0 \\ u - \widehat{\sigma(u)}\log(p) & \text{si} \quad \xi = 0 \end{cases} \tag{2.10}$$

Dans le cadre non stationnaire (conditionnel), il y a en général deux méthodes permettant d'inclure la non-stationnarité dans l'estimation des quantiles hydro-climatiques, soit directement en

introduisant le temps, soit en introduisant une covariable qui est elle-même fonction du temps. L'introduction des covariables peut être effectuée au niveau de n'importe quel paramètre ou même à deux ou à trois paramètres à la fois dans la loi de probabilité choisie (dans notre cas c'est la loi GEV ou GPD). L'effet d'une covariable sur la variable d'intérêt peut être pris en compte dans une forme polynomiale (linéaire ou quadratique) [Coles, 2001; Ouarda & Adlouni, 2011; Cannon, 2010] ou dans d'autres formes non paramétriques (Splines, etc) [Chavez-Demoulin & Davison, 2005; Nasri *et al.*, 2013].

Considérons Y une variable aléatoire liée à une covariable X, en supposant que Y est distribuée selon la loi $\text{GEV}_{\mu,\sigma,\xi}$ (ou $\text{GPD}_{\sigma,\xi}$). La dépendance entre la variable d'intérêt Y et la covariable X peut se traduire dans les paramètres de la distribution GEV (GPD). Ici, nous allons nous contenter de donner l'exemple avec le paramètre de position μ. Donc, μ peut s'écrire comme une fonction de la covariable X: $\mu_X = f(X)$ où la fonction f est soit linéaire, quadratique ou autre. Dans ce cas, au lieu d'avoir une seule valeur du quantile pour chaque p, on aura n valeurs de quantile pour chaque p. La figure 2.3 donne une illustration de l'estimation de la médiane inconditionnelle versus la médiane conditionnelle dans le cadre linéaire et quadratique en se basant sur un échantillon de 1000 observations tirées de la distribution GEV avec $(\mu = (X/300) + (X/300)^2, \sigma = 1, \xi = 0)$ avec X une covariable qui prend des valeurs entre $[1, 1000]$.

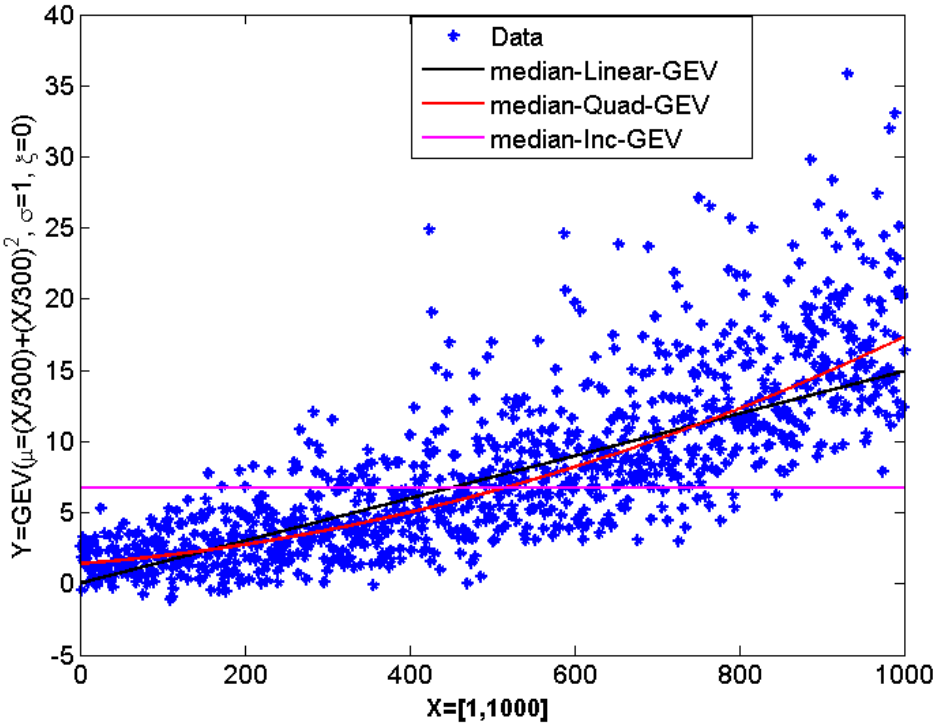

Figure 2.3 – Illustration de l'estimation de la médiane inconditionnelle versus la médiane conditionnelle dans le cadre linéaire et quadratique en se basant sur un échantillon de 1000 observations tirées de la distribution GEV avec $\left(\mu = (X/300) + (X/300)^2, \sigma = 1, \xi = 0\right)$ **avec** X **une covariable qui prend des valeurs entre** $[1, 1000]$

2.2.2 Quantiles basés sur la régression des quantiles

La régression des quantiles est une méthode statistique qui permet d'étudier l'impact de différentes covariables sur l'ensemble de la distribution de la variable d'intérêt. Contrairement à la régression ordinaire qui se rapproche des moyennes conditionnelles de la variable d'intérêt par rapport aux valeurs des covariables, la régression des quantiles donne une estimation des quantiles. Dans la régression ordinaire, le coefficient de régression représente le changement opéré dans la variable d'intérêt par unité de changement dans la covariable associée à ce coefficient. Dans la régression des quantiles, le coefficient de régression fournit une estimation du changement d'un quantile spécifique

de la variable d'intérêt par unité de changement de la covariable. Considérons une variable aléatoire Y (liée à une covariable X) de fonction de répartition $F_Y(y) = P(Y \leqslant y)$. Le quantile d'ordre p est défini par : $Q_p(Y) = \inf\{y : F_Y(y) \geq p\}$. La régression des quantiles tente d'évaluer comment les quantiles conditionnels $Q_p(Y|X) = \inf\{y : F_{Y|X}(y) \geq p\}$ changent lorsque la covariable X varie. Dans le cas de la régression des quantiles linéaires, les quantiles conditionnels prennent la forme suivante :

$$Q_p(Y|X) = X'\beta_p, \tag{2.11}$$

où à chaque valeur de p correspond un coefficient β_p. L'expression (2.11) peut s'écrire d'une manière équivalente :

$$Y = X'\beta_p + \varepsilon \quad \text{avec} \quad Q_p(\varepsilon|X) = 0. \tag{2.12}$$

Pour bien comprendre le principe de la régression des quantiles, il est bien utile de détailler comment on peut estimer les quantiles à partir du modèle de l'équation (2.12). En effet, l'estimation des régressions quantiles part de l'observation cruciale que le quantile d'ordre p est le résultat de la minimisation suivante (voir [Koenker & Bassett, 1987] pour la preuve):

$$Q_p(Y) = \underset{\beta}{\operatorname{argmin}} E\left[\rho_p\left(Y - X'\beta\right)\right], \tag{2.13}$$

où ρ_p est une fonction de perte définie par $\rho_p(u) = u\left(p - 1\{u < 0\}\right)$. Cette estimation peut sembler moins intuitive que l'approche directe, qui utilise la statistique d'ordre $Y_{(1)} < \ldots < Y_{(n)}$ en estimant $Q_p(Y)$ par $\widehat{Q}_p(Y) = Y_{[np]}$ où $[np]$ est le plus petit entier supérieur ou égal à np. L'intérêt de cette méthode est qu'elle peut s'étendre facilement à un cadre conditionnel où on modélise le quantile

conditionnel de la variable d'intérêt Y comme une fonction explicative des covariables X:

$$Q_p\left(Y|X=x\right) = \underset{\beta}{\operatorname{argmin}} E\left[\rho_p\left(Y - X'\beta\right)|X=x\right]. \tag{2.14}$$

Dans la régression des quantiles, les coefficients régresseurs peuvent se définir comme suit:

$$\beta_p = \underset{\beta}{\operatorname{argmin}} E\left[\rho_p\left(Y - X'\beta\right)\right]. \tag{2.15}$$

On peut noter l'analogie avec le modèle de régression ordinaire, qui modélise l'espérance conditionnelle de Y par une forme linéaire en X: $E\left(Y|X\right) = X'\beta$. Un estimateur de l'espérance d'une variable aléatoire Y conditionnel à X pouvant être obtenu par la fonction de perte quadratique $\underset{\beta}{\operatorname{argmin}} E\left[\left(Y - X'\beta\right)^2|X=x\right]$. La fonction de perte quadratique est donc remplacée, dans la régression quantile, par la fonction de perte ρ_p. La figure 2.4 illustre la différence entre la régression des quantiles et la régression ordinaire.

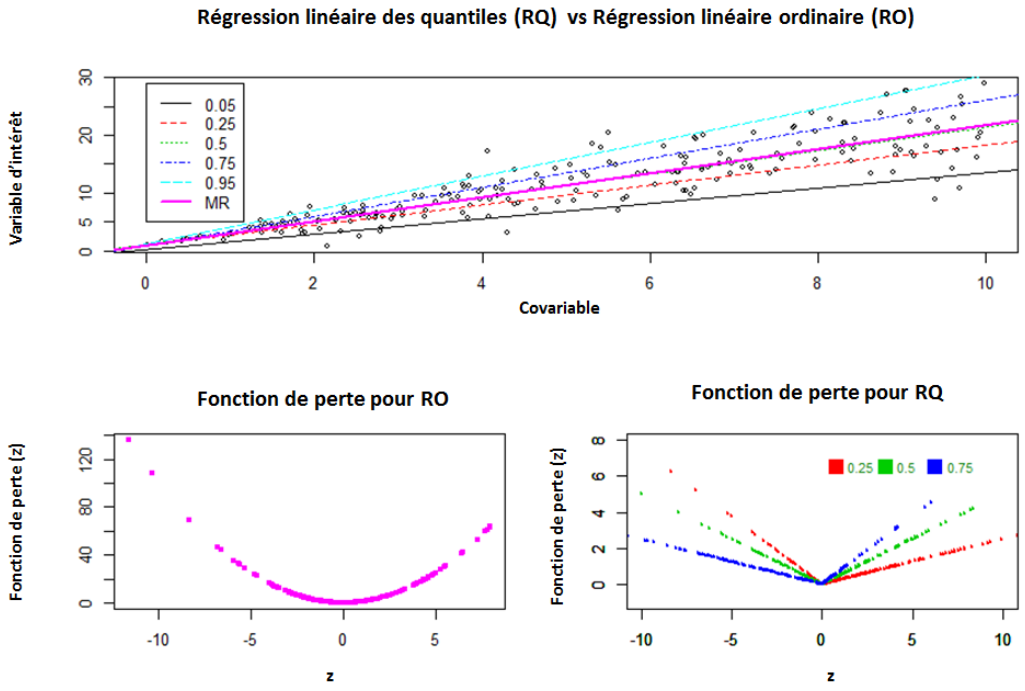

Figure 2.4 – Illustration de la différence entre la régression des quantiles et la régression ordinaire. z **est la différence entre la variable d'intérêt et son estimé.**

Contrairement à la forme quadratique de la fonction de perte de la régression ordinaire, la fonction de perte ρ_p pénalise moins les grand écarts, ce qui explique la robustesse de la régression des quantiles aux valeurs extrêmes et aux observations aberrantes.

Quelques récentes études en science de l'environnement ont utilisé la régression des quantiles linéaires pour l'estimation des quantiles conditionnels, par exemple [Friederichs & Hense, 2007; Tareghian & Rasmussen, 2013].

Chapitre 3

Objectifs et Méthodologie

3.1 Objectif 1: Modélisation des quantiles conditionnels en se basant sur la fonction de répartition avec des dépendances non nécessairement linéaires ou quadratiques

3.1.1 Description de l'objectif

Dans ce travail, nous proposons une approche pour l'estimation des quantiles conditionnels basée sur l'inverse de la fonction de répartition de la distribution généralisée des valeurs extrêmes (GEV) dans le cadre où les paramètres de la distribution dépendent des covariables. Ici, la dépendance est exprimée par des fonctions de lissage, nommées les B-Splines. L'estimation des paramètres du modèle proposé est faite dans un cadre bayésien et l'estimation de la loi a posteriori est effectuée en utilisant l'algorithme de Metropolis Hasting (M-H)[Metropolis *et al.*, 1953; Hastings, 1970].
Cette approche a été appliquée pour l'estimation des quantiles des précipitations annuelles maximales pour six stations en Afrique du Nord en présence de 14 covariables climatiques issues des

données réanalysées de NCEP-NCAR [1]. Ces covariables donnent des informations sur la vélocité du vent, la température, l'humidité et la hauteur géopotentielle pour chaque station étudiée. L'ensemble des covariables utilisées est donné dans la partie 1.2 de l'article 1. La figure 1.1 dans l'article 1 montre la position géographique de chaque station et le tableau 1.1 présente une description plus détaillée des données utilisées pour chaque station.

La section suivante résume les aspects théoriques de l'approche. L'article 1 présente plus de détails sur cette méthodologie.

3.1.2 Méthodologie

Le modéle GEV généralisé: Soit Y une variable aléatoire qui suit une loi GEV(μ_X, σ_X, ξ_X), avec X qui est la covariable associée à Y. Pour simplifier le modèle, nous supposons que seul le paramètre de location dépend de la covariable X. Ce paramètre s'écrit donc comme une fonction de cette covariable et prend la forme suivante:

$$\mu_X = f(X), \tag{3.1}$$

avec f, une fonction inconnue.

Dans la littérature, plusieurs méthodes ont été proposées afin d'approximer une fonction inconnue à partir d'un échantillon de données. La méthode la plus évidente est l'interpolation polynomiale, qui est une opération mathématique permettant de construire une fonction polynomiale (quadratique, cubique, etc.) à partir d'un nombre fini des données. Cependant ce type d'interpolation présente certains inconvénients, comme le phénomène de Runge [De Boor, 2001] (voir annexe 0.1 pour plus de détails). Une des solutions à ces inconvénients est l'utilisation des fonctions splines qui sont des

1. National Centers for Evironmental Prédiction/ National Center for Atmospheric Reserch

fonctions définies en utilisant des polynômes par morceaux. Il y a plusieurs types de fonctions splines [De Boor, 2001]. Les plus populaires sont les splines cubiques, les splines naturelles et les B-Splines. Dans notre cas, nous utilisons les fonctions B-Splines pour les avantages suivants: Un lissage à base B-spline est indépendant de la variable d'intérêt et dépend seulement des informations suivantes: (i) l'étendue de la variable explicative; (ii) le nombre et la position des nœuds et (iii) le degré du B-spline.

Les fonctions B-Splines: Soit $x_0 \leq x_1 \leq x_2 \cdots \leq x_m$ une suite de $(m+1)$ valeurs réelles. Ces valeurs sont appelées nœuds et l'élément $(x_0, x_1, \ldots, x_m)$ est appelé un vecteur de nœuds. Pour définir les fonctions B-splines, on a besoin de préciser le degré k et la $i^{ème}$ fonction B-spline $B_{i,k}(x)$ qui est définie par récurrence comme suit:

$$B_{i,0}(x) = \begin{cases} 1 \text{ si } x_i \leq x < x_{i+1} & \text{pour tout} \quad i = 0, \ldots, (m-1) \\ 0 \quad \text{ailleurs} \end{cases} \tag{3.2}$$

Et

$$B_{i,k}(x) = \frac{x - x_i}{x_{i+k} - x_i} B_{i,k-1}(x) + \frac{x_{i+k+1} - x}{x_{i+k+1} - x_{i+1}} B_{i+1,k-1}(x) \qquad \text{pour tout} \quad i = 0, \ldots, (m-k-1). \tag{3.3}$$

$B_{i,k}(x)$ est un polynôme de degré k sur chaque intervalle semi-ouvert $[x_i, x_{i+1}[$. Autrement dit, m représente le nombre de ces intervalles dans un échantillon de données et dans chaque intervalle, un polynôme de degré k est construit. Le modèle B-Splines avec $k = 1$ et $m = 1$ correspond au modèle linéaire et le modèle B-Spline avec $k = 2$ et $m = 1$ correspond au modèle quadratique. La figure 3.1 montre une illustration d'un lissage à base de fonctions B-Splines avec $k = 3$ et $m = 7$. Entre chaque deux noeuds, une fonction polynomiale de degré $k = 3$ est construite.

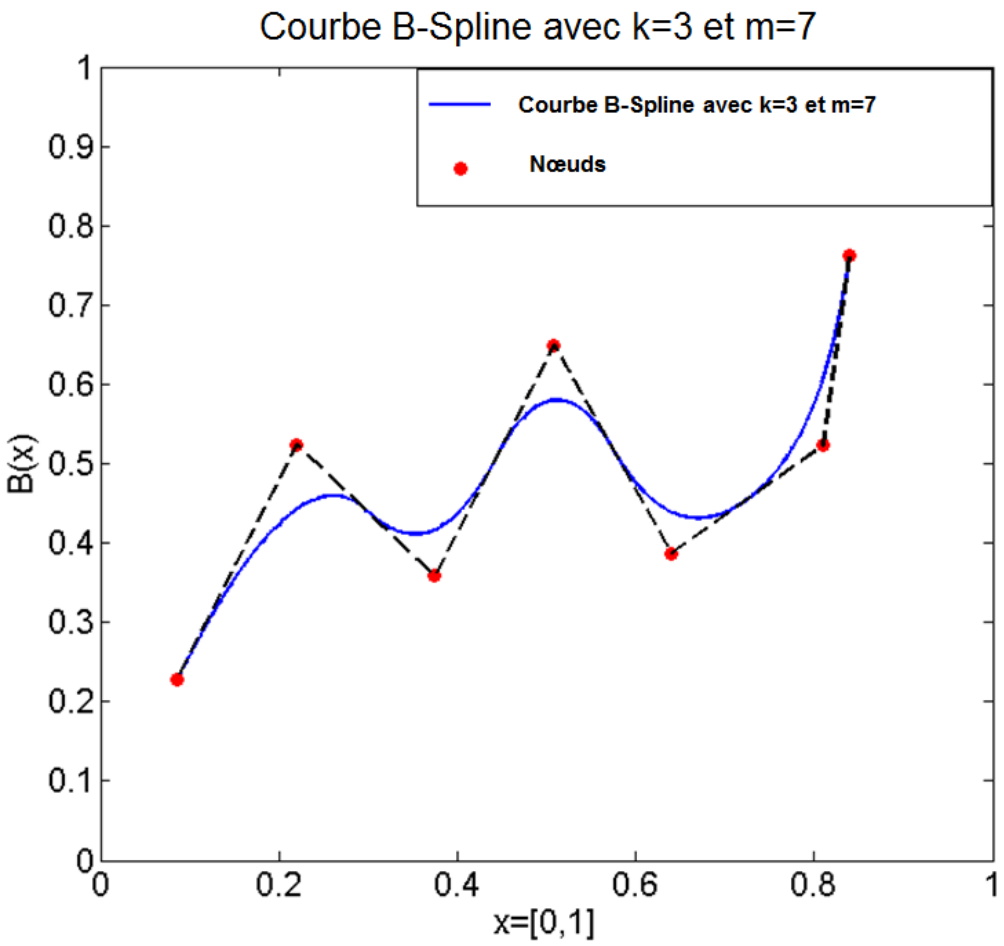

Figure 3.1 – Illustration d'un lissage à base de B-Splines

Le modèle GEV-B-Splines: On suppose que le paramètre de location μ_X dans le modèle GEV généralisé est donné par :

$$\mu_X = \sum_{i=1}^{m} \beta_i B_{i,k}(x). \tag{3.4}$$

Les bases $B_{i,k}$ seront calculées à partir de l'échantillon de la covariable X et β_i sont des paramètres à estimer.

Nous pouvons de la même manière faure varier les deux autres paramètres de la GEV, toutefois nous risquons d'avoir un problème de malédiction de la dimensionnalité.

Estimation des paramètres: L'estimation des paramètres du modèle GEV-B-Splines, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_m)$, σ et ξ, se fera dans un cadre bayésien.

Dans l'approche bayésienne, les paramètres ne sont pas des valeurs constantes inconnues mais des variables aléatoires admettant une distribution a priori. Toute l'inférence bayésienne est basée sur la loi a posteriori des paramètres et donc des quantiles, qui combine l'information tirée des données à travers la vraisemblance et celle de la loi a priori. L'estimation de la loi a posteriori est effectuée en utilisant l'algorithme de Metropolis-Hasting (M-H). Ce dernier est un algorithme de type Monte Carlo par Chaîne de Markov (MCMC) qui permet de simuler une distribution à l'aide d'une chaîne de Markov [1]. Il s'agit de l'algorithme le plus général parmi la famille des méthodes MCMC dans la mesure où il impose moins de conditions possibles à la densité cible. Cet algorithme fut d'abord publié par Metropolis *et al.* [1953] puis généralisé par Hastings [1970]. Il permet de générer une chaîne de Markov d'une loi de densité stationnaire à partir d'une densité appelée loi instrumentale (voir tableau 1.3).

L'approche bayésienne suppose que l'on connait les quantités suivantes:

— La fonction de vraisemblance de la variable aléatoire Y conditionnellement aux paramètres du modèle GEV-B-Splines $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma, \xi)$. Cette fonction sera notée $L(y|\boldsymbol{\theta})$.

— La loi a priori du vecteur de paramètres $\boldsymbol{\theta}$ notée par $\pi(\boldsymbol{\theta})$ qui résume l'information dont on dispose sur les paramètres du modèle à estimer.

On en déduit à l'aide du théorème de Bayes la loi a posteriori du vecteur $\boldsymbol{\theta}$:

$$\pi(\boldsymbol{\theta}|y) \propto L(y|\boldsymbol{\theta})\, \pi(\boldsymbol{\theta}). \tag{3.5}$$

1. Un processus de Markov est un processus stochastique possédant la propriété de Markov : l'information utile pour la prédiction du futur est entièrement contenue dans l'état présent du processus et n'est pas dépendante des états antérieurs (le système n'a pas de « mémoire »). En effet, Une suite $(X_i)_{i\geq 0}$ de v.a. discrètes est appelée chaîne de Markov si elle vérifie la propriété de Markov, qui caractérise les processus sans mémoire: $P(X_{i+1} = x_{i+1}|X_i = x_i, X_{i-1} = x_{i-1}, \ldots, X_0 = x_0) = P(X_{i+1} = x_{i+1}|X_i = x_i)$

La loi a posteriori s'interprète comme un résumé (en un sens probabiliste) de l'information disponible sur $\boldsymbol{\theta}$, une fois y observé. L'approche bayesienne $\pi(\boldsymbol{\theta}|y)$ réalise l'actualisation de l'information a priori $\pi(\boldsymbol{\theta})$ par l'observation y.

La formule suivante décrit la fonction de vraisemblance de la distribution GEV avec les fonctions B-Splines. La fonction de vraisemblance dans ce cas n'est plus que le produit de la densité de la distribution GEV, avec un paramètre de location qui dépend d'une covariable, évaluée pour un échantillon d'observation et elle s'écrit comme suit:

$$L(y|\boldsymbol{\theta}) = \prod_{i=1}^{n_1} \frac{1}{\sigma} \exp\left[-\left(1-\xi\left(\frac{y_i - \sum_{j=1}^{m}\beta_j B_{j,k}(x_i)}{\sigma}\right)\right)^{-\frac{1}{\xi}}\right]\left[1-\xi\left(\frac{y_i - \sum_{j=1}^{m}\beta_j B_{j,k}(x_i)}{\sigma}\right)\right]^{-1+\frac{1}{\xi}}$$
$$* \prod_{i=n_1+1}^{n} \frac{1}{\sigma} \exp\left[-\left(\frac{y_i - \sum_{j=1}^{m}\beta_j B_{j,k}(x_i)}{\sigma}\right)\right] \exp\left[-\exp\left(-\left(\frac{y_i - \sum_{j=1}^{m}\beta_j B_{j,k}(x_i)}{\sigma}\right)\right)\right] \quad (3.6)$$

avec n_1 est le nombre d'observations telles que $\xi \neq 0$.

Maintenant, il reste à définir la distribution a priori du vecteur $\boldsymbol{\theta}$. Nous supposons dans ce travail que les paramètres de la distribution GEV sont indépendants [1] donc la distribution a priori de $\boldsymbol{\theta}$ s'écrira comme le produit des distributions a priori des paramètres $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_m)$, σ et ξ et elle prend la forme suivante:

$$\pi(\boldsymbol{\theta}) = \pi_1(\boldsymbol{\beta})\,\pi_2(\sigma)\,\pi_3(\xi). \quad (3.7)$$

Martins & Stedinger [2000] ont proposé la distribution Beta $\mathscr{B}(6,9)$ comme distribution a priori pour le paramètre de forme dans le cadre de la distribution GEV stationnaire (paramètres constants) afin d'éviter des estimateurs irrationnels (dans l'ensemble des nombres irrationnels).

Pour la densité a priori de $\boldsymbol{\beta}$, nous n'avons pas un grand choix de distributions. Dans la littérature,

1. l'indépendance est supposée pour faciliter le modèle et aussi parce que ne nous possédons aucune information sur les distributions jointes des paramètres.

quatre distributions paramétriques multivariées sont considérées: les distributions Student, normale, gamma et lognormale. Les distributions gamma et lognormale sont restreintes aux paramètres positifs, ce qui n'est pas le cas des paramètres des fonctions B-Splines qui peuvent prendre des valeurs négatives aussi. Par conséquent, il semble naturel de considérer la distribution normale ou Student. Nous considérons ici la distribution normale multivariée (voir Green & Silverman [1994], p. 51-52, pour une discussion sur l'utilisation de la densité normale multivariée comme loi a priori dans ce contexte).

Pour le paramètre d'échelle, nous considérons la distribution a priori non informative suivante: $\pi_2(\sigma) = \frac{1}{\sigma}$.

L'estimation de la distribution a posteriori s'effectuera en utilisant l'algorithme de M-H décrit par les étapes du tableau 3.1:

Table 3.1 – Description de l'algorithme de Métropolis-Hasting

l'algorithme M-H
Initialiser les paramètres $(\boldsymbol{\theta^0} \sim \Phi(.))$ (Φ est appelée une distribution instrumentale)
pour les itérations $i = 1, 2, ...$
faire
Proposer: $\boldsymbol{\theta^*} \sim \Phi\left(.|\boldsymbol{\theta^{i-1}}\right)$ ($\boldsymbol{\theta^*}$ est appelé un candidat)
La procédure d'acceptation-rejet du candidat $\alpha\left(\boldsymbol{\theta^*}|\boldsymbol{\theta^{i-1}}\right) = \min\left\{1, \frac{\Phi\left(\boldsymbol{\theta^{i-1}}|\boldsymbol{\theta^*}\right)\pi(\boldsymbol{\theta^*})}{\Phi\left(\boldsymbol{\theta^*}|\boldsymbol{\theta^{i-1}}\right)\pi\left(\boldsymbol{\theta^{i-1}}\right)}\right\}$
$u \sim$ Uniforme $(., 0, 1)$
si $u < \alpha$
on accepte le candidat $\boldsymbol{\theta^i} \leftarrow \boldsymbol{\theta^*}$
sinon
on le rejette $\boldsymbol{\theta^i} \leftarrow \boldsymbol{\theta^{i-1}}$
fin si
fin pour

3.2 Objectif 2: Modélisation des quantiles conditionnels en se basant sur la régression des quantiles avec des dépendances non nécessairement linéaires ou quadratiques

3.2.1 Description de l'objectif

Cet objectif de recherche consiste à estimer des quantiles conditionnels en se basant sur le modèle généralisé de la régression des quantiles. Les B-Splines seront utilisées afin de décrire le lien entre la variable d'intérêt et la covariable. L'estimation des paramètres se fera dans un cadre bayésien en utilisant l'algorithme M-H. Cette approche a été appliquée pour estimer les quantiles conditionnels des débits maximaux et minimaux annuels en Ontario en présence des covariables climatiques suivantes : l'indice de l'oscillation décennale du Pacifique (PDO: Pacific Decadal Oscillation)[Nathan & Hare, 2002] et l'indice de l'oscillation multidécennale de l'Atlantique [Teegavarapu *et al.*, 2013](AMO: Atlantic-Multi-decadal Oscillation). Le PDO et l'AMO sont deux indices qui caractérisent, respectivement, les variations de la température de surface dans l'océan Pacifique et l'océan Atlantique. Ces deux indices ont été choisis parmi de nombreux autres indices testés dans cette étude en raison de leurs dépendances significatives avec les variables d'intérêt. Nous avons sélectionné cinq stations pour estimer les quantiles conditionnels des débits maximaux et cinq autres stations pour estimer les quantiles conditionnels des débits minimaux. Ces stations sont été choisies parmi 139 stations en Ontario parce qu'elles répondaient aux conditions suivantes :

— Elles avaient plus de 30 ans de données journalières complètes par station,

— Les séries chronologiques de données des minimas ou des maximas annuels présentaient une tendance temporelle (Cette tendance est détectée en appliquant le test original de Mann Kendall [Mann, 1945; Kendall, 1975],

— Les séries chronologiques de données des minimas ou des maximas annuelles ont au moins une dépendance significative avec un indice climatique.

La figure 2.2 dans l'article 2 montre la position géographique de chaque station étudiée et le tableau 2.1 présente une description plus détaillée des données utilisées.
La section suivante décrit brièvement les aspects théoriques de l'approche. L'article 2 présente plus de détails sur cette méthodologie.

3.2.2 Méthodologie

Régression des quantiles généralisée: Le modèle généralisé de la régression des quantiles relie la variable d'intérêt et les covariables par une fonction de lien. Il s'agit d'une généralisation du modèle linéaire qui suppose que cette fonction de lien est linéaire. En effet, le but de la régression des quantiles généralisée est d'identifier les meilleures fonctions possibles qui décrivent le lien entre la variable d'intérêt et les covariables selon la distribution des données plutôt que d'estimer les paramètres à partir d'un modèle spécifique. Considérons le cas de régression des quantiles généralisés avec une seule covariable ainsi:

$$y = f(x) + \varepsilon. \tag{3.8}$$

Dans la régression quantile généralisée, la fonction f n'est pas spécifiée et on suppose généralement que les erreurs sont indépendantes et identiquement distribuées. Plusieurs méthodes ont été proposées dans la littérature [Koenker, 2005] pour estimer la fonction f. Dans ce travail, nous proposons les fonctions B-Splines pour l'estimation de cette fonction.

Régression des quantiles avec les fonctions B-Splines: Le modèle de regression généralisée avec les B-Splines s'écrit comme suit:

$$y = \sum_{i=1}^{m} \beta_i B_{i,k}(x) + \varepsilon, \tag{3.9}$$

avec β_i des paramètres à estimer, m le nombre de noeuds et k le degré de la fonction B-Spline.

Estimation des paramètres: L'approche classique pour l'estimation des paramètres des modèles de régression des quantiles est basée sur les méthodes du simplex ou les méthodes de point intérieur décrites dans [Koenker, 2005]. Dans cet objectif, nous proposons l'utilisation de la méthode bayésienne pour l'estimation des paramètres. Nous proposons la loi normale multivariée comme distribution a priori pour les paramètres β_i pour les mêmes raisons que celles mentionnées pour le premier objectif. Il ne nous reste alors qu'à définir la fonction de vraisemblance $L(y|\boldsymbol{\beta})$. Pour ce faire, nous nous sommes basés sur l'observation donnée par [Yu & Moyeed, 2001] qui fait le lien entre la minimisation de la fonction de perte utilisée dans la régression des quantiles et la maximisation de la fonction de vraisemblance de la distribution asymétrique de Laplace.
Soit U une variable aléatoire. On dit que U suit la distribution asymétrique de Laplace si sa fonction de densité de probabilité s'écrit comme suit:

$$L_p(u) = p(p-1)\exp\{-\rho_p(u)\}; \quad -\infty < u < +\infty \quad \text{et} \quad p \in]0,1[. \tag{3.10}$$

La maximisation de la fonction de densité de Laplace est similaire à la minimisation de la fonction de perte utilisée pour estimer les paramètres du modèle de la régression des quantiles. Donc, dorénavant on utilise la fonction (3.10) au lieu de la fonction de perte. Par conséquent, la fonction de

vraisemblance peut s'écrire comme suit:

$$L\left(y|\boldsymbol{\beta}\right) = p^n(1-p)^n \exp\left\{-\sum_{j=1}^{n} \rho_p\left(y_j - \sum_{i=1}^{m} \boldsymbol{\beta}_i B_{ik}(x_j)\right)\right\}. \tag{3.11}$$

Nous utilisons également l'algorithme M.-H. pour l'estimation de la distribution a posteriori dans cet objectif.

3.3 Objectif 3: Modélisation des quantiles conditionnels en se basant sur les copules

3.3.1 Description de l'objectif

Dans cet objectif, nous considérons une nouvelle approche pour la modélisation des quantiles conditionnels basée sur les copules; des fonctions qui définissent la structure de dépendance entre plusieurs variables. L'idée clé de cette approche consiste à exploiter le lien entre la copule et une distribution conditionnelle comme le montre le travail de Bouyé & Salmon [2002]. Nous allons donc utiliser ce lien pour estimer les quantiles conditionnels. Nous proposons dans cet objectif deux estimateurs pour les quantiles conditionnels basés sur les copules et nous les comparons avec les estimateurs proposés dans la littérature et ceux proposés dans les objectifs 3.1 et 3.2. Ces deux estimateurs seront utilisés pour estimer les quantiles conditionnels pour les deux cas d'étude suivants: (i) l'estimation des quantiles du débit maximum annuel pour une station en Ontario en présence de l'indice d'oscillation multidécennale de l'Atlantique (AMO), (ii) l'estimation des quantiles conditionnels des précipitations annuelles maximales pour une station en Californie en présence de l'oscillation décennale du Pacifique (PDO) et l'indice de l'oscillation australe (SOI).

Le choix des covariables ici est basé sur nos travaux antérieurs qui ont proposé les mêmes jeux de données (voir [Nasri *et al.*, 2013] et article 2). Les figures 3.3 et 3.5 montrent les positions géographiques des stations ainsi que les séries de données des variables étudiées dans les deux cas d'étude.

Le paragraphe suivant donne d'abord une revue de littérature sur les copules et ensuite explique brièvement l'approche théorique de cet objectif. D'autres détails sur cette méthodologie sont fournis dans la partie 3.

3.3.2 Méthodologie

Les copules: définitions et propriétés

Définition 1 *Une copule de dimension q est une fonction de distribution multivariée, notée C, avec des distributions marginales uniformes sur* $[0,1]$*:*

- *C : $[0,1]^q \rightarrow [0,1]$;*
- *C est bornée et q-croissante;*
- *C possède des distributions marginales uniformes C_i, c'est-à-dire:*

$$C_i(u) = C\left(1, \ldots, u, 1, \ldots, 1\right) = u \qquad \textit{pour tout} \qquad u \in [0,1]$$

Le théorème de Sklar (1959) a mis en lumière le grand potentiel des copules pour la construction des distributions mutltivariées:

Théoreme 1 *Soit* $\boldsymbol{F}$ *une fonction de répartition q-dimensionelle avec des marginales* $F_1, \ldots, F_q$*, alors il existe une copule* C *telle que pour tout* $\boldsymbol{x} = (x_1, \ldots, x_q) \in \mathbb{R}^q$*:*

$$\boldsymbol{F}(x_1, \ldots, x_q) = C(F_1(x_1), \ldots, F_q(x_q)). \tag{3.12}$$

Selon le théorème de Sklar, pour $(X_1, \ldots, X_q)$ un vecteur de variables aléatoires continues admettant $F_1, \ldots, F_q$ comme fonctions de répartitions marginales et $\boldsymbol{F}$ comme fonction de répartition jointe, alors il existe une copule C qui vérifie l'équation (3.12). Si les marginales $F, \ldots, F_q$ sont continues, alors C est unique.

La densité d'une copule, notée c, si elle existe, est définie comme suit:

$$c(u_1, \ldots, u_q) = \frac{\partial^q C}{\partial u_1 \ldots \partial u_q}(u_1, \ldots, u_q). \tag{3.13}$$

Les types de copules

— **Copules archimédiennes:** Soit φ une fonction convexe, continue, strictement décroissante de $[0, 1]$ dans $[0, \infty[$ telle que $\varphi(1) = 0$ et $\varphi(0) = \infty$ alors $C(u_1, \ldots, u_q) = \varphi^{-1}(\varphi(u_1) + \ldots + (u_q))$ est une copule archimédienne stricte et φ est appelée générateur strict de C. Les exemples les plus communs de copules archimédiennes comprennent les familles de Frank [Frank, 1979], de Gumbel [Frank, 1960] et de Clayton [Clayton, 1978]. Ici, on donne les générateurs de ces familles dans le cadre bivarié.

(a) Clayton: $\varphi(t) = \theta^{-1}\left(t^\theta - 1\right), \theta \in [-1, +\infty[\setminus 0$.

(b) Gumbel: $\varphi = (-\log(t))^\theta, \theta \in [1, +\infty[$.

(c) Frank: $\varphi(t) = -\log\left(\frac{\exp(-\theta t)-1}{\exp(-\theta)-1}\right), \theta \in]-\infty, +\infty[\setminus 0$.

Le paramètre θ mesure le degré de dépendance entre X_1 et X_2.

La figure 3.2 montre une illustration des fonctions de densité et des fonctions cumulatives des copules Clayton, Gumbel et Frank pour un $\theta = 5$.

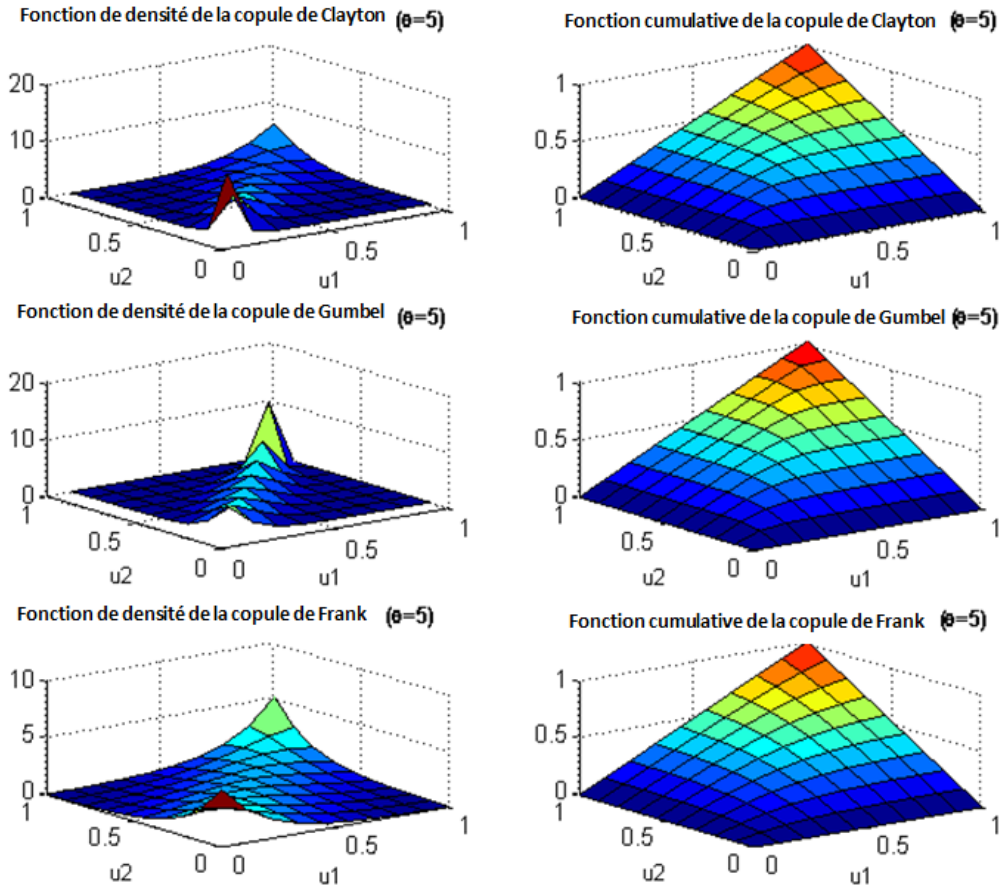

Figure 3.2 – Illustration des fonctions de densité et des fonctions cumulatives des copules Clayton, Gumbel et Frank ($\theta = 5$).

La copule de Gumbel ne saisit que des dépendances positives et ne peut représenter que les variables dont la structure de dépendance est plus accentuée sur la queue supérieure.

À l'inverse de la copule de Gumbel, la copule de Clayton suppose une forte dépendance dans la queue inférieure des événements. La copule de Frank permet de modéliser les dépendances symétriques aussi bien positives que négatives.

— **Copules elliptiques:** Une distribution est dite elliptique si elle peut s'écrire sous la forme:

$$Z_{q+1} = \mu_{q\times 1} + R\boldsymbol{A}_{q\times q}\boldsymbol{u}_{\boldsymbol{q}\times\boldsymbol{1}} \sim \mathcal{E}\left(\mu, \Sigma, g\right) \tag{3.14}$$

où $\mu_{q\times 1}$ est un vecteur de localisation, R est une variable aléatoire positive, $\boldsymbol{A}$ est une matrice $q \times q$ telle que $\boldsymbol{A}\boldsymbol{A}^{\boldsymbol{T}} = \Sigma_{q\times q}$, $\boldsymbol{u}$ est un vecteur uniformément distribué sur la sphère de dimension q et g est une fonction d'échelle, appelée générateur. Les distributions elliptiques les plus célébres sont la loi normale, de Student, exponentielle et Cauchy.

Les copules elliptiques sont définies à partir des familles des lois elliptiques. On en considère ici deux cas particuliers, la copule gaussienne et la copule de Student.

(a) Gaussienne: la copule gaussienne est définie sur le cube unitaire $[0,1]^q$. Elle est construite à partir d'une distribution normale multivariée sur $\mathbb{R}^q$. Pour une matrice de corrélation donnée $R \in [-1,1]^{q\times q}$, la copule gaussienne peut s'écrire:

$$C\left(u_0, u_1, \ldots, u_q\right) = \Phi_R\left(\Phi^{-1}\left(u_0\right), \Phi^{-1}\left(u_1\right), \ldots, \Phi^{-1}\left(u_q\right)\right),$$

où Φ^{-1} est l'inverse de la fonction cumulative de la ditribution normale standardisée et Φ_R est la distribution normale multivariée avec une moyenne $\mathbf{0}$ et une matrice de variance-covariance R.

(b) Student: la copule de Student est définie sur le cube unitaire $[0,1]^q$. Elle est construite à partir d'une distribution de Student multivariée sur $\mathbb{R}^q$. Pour une matrice de corrélation donnée $R \in [-1,1]^{q\times q}$ et d un degré de liberté. La copule de Student avec la matrice de paramètres R et d peut s'écrire:

$$C\left(u_0, u_1, \ldots, u_q\right) = \boldsymbol{t}_{d,R}\left(t_d^{-1}\left(u_0\right), \ldots, t_d^{-1}\left(u_q\right)\right),$$

où t_d^{-1} est l'inverse de la distribution univariée de Student et $\boldsymbol{t}_{d,R}$ est la distribution multivariée de Student avec les paramètres R et d.

La figure 3.3 montre une illustration des fonctions de densité et des fonctions cumulatives des copules gaussiennes et Student.

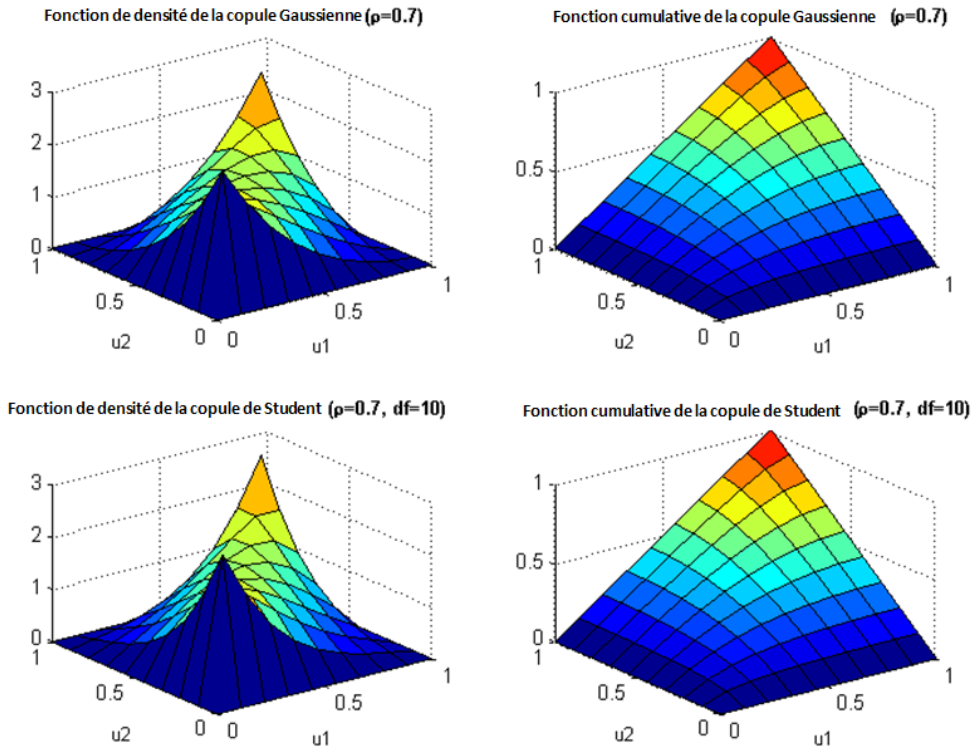

Figure 3.3 – illustration des fonctions de densité et des fonctions cumulatives des copules Gaussienne et de Student.

Copules et mesure de dépendance: Schweizer & Wolff [1981] ont montré que les copules sont capables de prendre en compte toute la dépendance entre deux variables aléatoires X_1 et X_2 . En effet, soient g_1 et g_2 deux fonctions strictement croissantes sur le domaine de définition de X_1 et X_2. Alors les variables obtenues par les transformations g_1 et g_2 ont la même copule que X_1 et X_2. Schweizer & Wolff [1981] ont démontré aussi que les mesures non paramètriques de dépendance (tau

de Kendall et rho de Spearman) peuvent s'exprimer à l'aide des copules. Le tableau 3.2 donne la relation entre le paramétre d'une copule et le tau de Kendall dans le cas bivarié.

Table 3.2 – Relation entre le paramètre d'une copule et le tau de Kendall dans le cas bivarié

Copule(paramétre)	**Tau de kendall**
Gaussienne(ρ)	$\frac{2}{\pi}\arcsin(\rho)$
Student (ρ, d)	$\frac{2}{\pi}\arcsin(\rho)$
Clayton (θ)	$\frac{\theta}{\theta+2}$
Gumbel (θ)	$1-\frac{1}{\theta}$
Frank (θ)	$1-\frac{4}{\theta}+4\frac{D_1(\theta)}{\theta}$; avec $D_1(\theta)=\int_0^\theta \frac{\frac{x}{\theta}}{\exp x-1}dx$

La relation entre les copules et les quantiles conditionnels: L'expression de la distribution conditionnelle d'une variable d'intérêt Y sachant un vecteur de covariables $\boldsymbol{X}=(X_1,\ldots,X_q)$ en terme de la copule et des distributions marginales est établie dans le travail de [Bouyé & Salmon, 2002] et donnée comme suit:

$$F_{Y|\boldsymbol{X}}(y|\boldsymbol{x})=\tilde{C}\left(F_Y(y),\boldsymbol{F}(\boldsymbol{x})\right) \tag{3.15}$$

où $F_{Y|\boldsymbol{X}}$ est la distribution conditionnelle de Y sachant $\boldsymbol{X}$ et

$$\tilde{C}(u_0,u_1,\ldots,u_q)=\frac{f_1(x_1)\ldots f_q(x_q)}{f(x_1,\ldots,x_q)}\frac{\partial^q C(u_0,u_1,\ldots,u_q)}{\partial u_1\ldots\partial u_q},$$

où f_j (resp. F_j) est la densité (resp. la fonction cumulative) de X_j, $u_j = F_j(x_j)$, $j = 1, \ldots, q$ et $u_0 = F_Y(y)$.

Dans ce cas, la fonction des quantiles conditionnels n'est rien d'autre que l'inverse de la distribution conditionnelle et s'écrit comme suit:

$$Q_Y(p|\boldsymbol{X} = \boldsymbol{x}) = F_Y^{-1}\left[\Gamma\left(\boldsymbol{F}(\boldsymbol{x}), p\right)\right], \tag{3.16}$$

avec Γ est l'inverse partiel de $\tilde{C}$ en respectant le second argument et F_Y^{-1} est l'inverse de F_Y.

Estimateurs proposés et estimations de paramètres: L'expression donnée par l'équation (3.16) permet une grande flexibilité pour estimer la fonction des quantiles conditionnels. Pour ce faire, on pourrait utiliser des approches d'estimation paramétriques, semi-paramétriques ou non paramétriques tout dépendemment de la méthode d'estimation des distributions marginales et de la fonction copule. Dans cet objectif, nous proposons deux approches, une paramétrique et l'autre semi-paramétrique.

Estimateur paramétrique: Cette approche suppose un modèle paramétrique pour la copule C, notée $C(.,.,\boldsymbol{\theta})$, et des modèles paramétriques pour les distributions marginales de $\boldsymbol{X}$ et Y, notés $F_Y(.,.,\boldsymbol{\alpha})$ et $\boldsymbol{F}(.,.,\boldsymbol{\beta})$ respectivement. L'estimateur paramétrique proposé est donc donné par:

$$\widehat{Q}_Y^p(\tau|\boldsymbol{X} = \boldsymbol{x}) = F_Y^{-1}\left[\Gamma\left(\boldsymbol{F}\left(\boldsymbol{x}, \widehat{\boldsymbol{\beta}}\right), \widehat{\boldsymbol{\theta}}, \tau\right), \widehat{\boldsymbol{\alpha}}\right], \tag{3.17}$$

où $\widehat{\boldsymbol{\alpha}}$, $\widehat{\boldsymbol{\beta}}$ et $\widehat{\boldsymbol{\theta}}$ sont les estimateurs des paramètres $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ et $\boldsymbol{\theta}$. Deux méthodes d'estimation des paramètres de la copule et de ses distributions marginales sont développées dans la littérature. La première repose sur l'utilisation de la fonction de vraisemblance complète [Shih & Louis, 1995;

Joe, 1997] pour estimer les paramètres $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ et $\boldsymbol{\theta}$. Le problème avec cette méthode est qu'elle peut engendrer des temps de calcul très longs dans le cas d'une grande dimension car elle nécessite d'estimer conjointement les paramètres des lois marginales et les paramètres de la structure de dépendance. De plus, l'estimation de la copule est sensible à une éventuelle erreur d'estimation des marginales car celles-ci interviennent dans le calcul de la vraisemblance. Comme solution à ce problème, Joe & Xu [1996] ont proposé une deuxième approche, appelée la méthode d'inférence pour les marginales (IFM: Inference Functions for Margins). Cette méthode estime dans une première étape les paramètres des fonctions marginales et ensuite elle estime les paramètres de la copule. La méthode IFM est fréquemment utilisée dans la littérature pour son efficacité. Pour plus de détails, voir Oakes [1982], Romano [2002] et Joe [2005]. Cette méthode sera utilisée pour estimer les paramètres dans notre approche paramétrique.

Estimateur semi-paramétrique: L'approche semi-paramétrique suppose un modèle paramétrique pour la copule C, notée $C\left(.,.,\boldsymbol{\theta}\right)$, et des modèles non paramétriques pour les distributions marginales de $\boldsymbol{X}$ et Y. L'estimateur semi-paramétrique proposé est donc donné par:

$$\widehat{Q}_{Y}^{sp}\left(\tau|\boldsymbol{X}=\boldsymbol{x}\right)=F_{Y}^{-1}\left[\Gamma\left(\widehat{\boldsymbol{F}}\left(\boldsymbol{x}\right);\widehat{\boldsymbol{\theta}},\tau\right)\right], \tag{3.18}$$

où $\widehat{\boldsymbol{\theta}}$ est un estimateur du paramètre de la copule $\boldsymbol{\theta}$ et $\widehat{F_{Y}}$, $\widehat{\boldsymbol{F}}$ sont les distributions empiriques de F_{Y} et $\boldsymbol{F}$ respectivement et elles sont définis ainsi:

$$\widehat{F_{Y}}\left(y\right)=\left(n+1\right)^{-1}\sum_{i=1}^{n}I\left(Y_{i}\leq y\right) \tag{3.19}$$

Et

$$\widehat{\boldsymbol{F}}\left(\boldsymbol{x}\right)=\left(n+1\right)^{-1}\sum_{i=1}^{n}I\left(\boldsymbol{X}_{i}\leq\boldsymbol{x}\right) \tag{3.20}$$

Pour estimer le paramètre de la copule, nous utilisons la méthode du maximum de vraisemblance canonique (CML: Canonical Maximum Likelihood) proposée par [Genest *et al.*, 1995]. Cette méthode est conceptuellement presque la même que l'IFM, mais elle ne pose aucune hypothèse sur les distributions marginales de la copule.

Chapitre 4

Résultats et conclusions générales

4.1 Résultats et conculsions de l'objectif 1

Dans ce travail, l'influence de 14 covariables climatiques sur les précipitations annuelles maximales pour six stations situées en Afrique du Nord a été étudiée en se basant sur le modèle GEV-B-Splines. L'avantage de ce modèle, par opposition à d'autres modèles non stationnaires, est qu'il prend en considération à la fois les fluctuations temporelles linéaires et non linéaires des covariables. Les modèles non stationnaires avec dépendance linéaire et quadratique prédéfinissent la forme de dépendance entre la variable d'intérêt et les covariables. Une comparaison a été effectuée entre le modèle GEV-B-Splines et les modèles GEV-linéaire et GEV-quadratique en se basant sur le test du rapport de la fonction de vraisemblance. Ce test permet la comparaison entre les fonctions de vraisemblance de chaque modèle tout en tenant compte du nombre de paramètres à estimer [2]. Le modèle GEV-B-Splines a été utilisé pour l'estimation des quantiles conditionnels des précipitations annuelles en présence de 14 covariables climatiques. Tout d'abord, chaque covariable a été introduite séparément dans le modèle GEV-B-Spline afin de détecter son effet sur la variable d'intérêt.

2. Test équivalent au critère d'information bayésien (BIC)

Ensuite, pour comprendre l'effet combiné de ces covariables, nous avons construit de nouvelles covariables, à partir de ces 14 covariables, en se basant sur des analyses en composantes principales. Les covariables sont considérées dans différents pas de temps (différents lag) afin de mieux comprendre leurs influences à court et à long termes (voir section 1.2 de l'article 1).

Les résultats ont montré que:

— Pour chaque station, il y a au moins une covariable qui explique significativement les variations des précipitations à long et à court termes.

— Les covariables les plus influentes sur la variable étudiée dans cette région sont les covariables décrivant la circulation à grande échelle (hauteur géopotentielle) et l'humidité.

— Les covariables décrivant la vélocité du vent influencent d'avantage les stations proches de la côte méditerranéenne.

Les covariables qui ont une influence significative sur les variables des précipiations étudiées ont été utilisées afin d'estimer les quantiles conditionnels en utilisant le modèle GEV-B-Splines. La Figure 1.6 donne un exemple de l'estimation de la médiane conditionnelle à partir du modèle GEV-B-Spline versus l'estimation de la médiane stationnaire pour chaque station étudiée. Nous remarquons que la médiane (temps de retour de 2 ans) estimée à partir du modèle GEV-B-Splines prend différentes valeurs tout dépendamenent de la valeur de la covariable. Nous constatons aussi que pour certaines valeurs de la covariable, le biais relatif entre les résultats de la médiane conditionnelle et la médiane stationnaire s'élève à plus de 30%. Ce biais augmente d'une façon exponentielle pour des temps de retour plus élevés. On peut conclure de ces résultats la grande utilité du modèle proposé pour l'estimation des quantiles non stationnaires et par conséquent pour l'estimation des risques d'inondation dans la région étudiée.

4.2 Résultats et conclusions de l'objectif 2

Dans ce travail, nous avons proposé une approche d'estimation pour les quantiles conditionnels liés aux extrêmes en présence de covariables climatiques basée sur la régression des quantiles et les fonctions de lissage (B-splines). Cette méthode a été utilisée pour estimer les quantiles conditionnels des débits maximaux et minimaux en Ontario en utilisant des covariables climatiques, en particulier l'AMO et le PDO. Une comparaison entre les modèles classiques de régression des quantiles et celui proposé est effectuée en se basant sur deux critères, le coefficient de determination pour les quantiles et le critère d'information bayésien (BIC). Les résultats ont montré que

— L'indice PDO influence plutôt le débit annuel minimum alors qu'il n'affecte pas le débit maximal annuel et inversement pour l'indice AMO. En regardant la série temporelle des oscillations PDO et AMO, on a constaté que la relation entre ces deux indices est négative pendant les périodes 1942-1965 et 1968-1998 [Rowan & Daniel, 2005]. Ceci peut expliquer l'influence de l'AMO pour le débit annuel maximal et le PDO pour les valeurs de débit annuel minimum.

— La forme de la relation entre l'AMO et les débits maximaux annuels est semblable pour les cinq stations étudiées, ce n'est pas le cas pour les débits minimaux et l'indice PDO. Deux des cinq stations situées au nord montrent une relation négative entre les débits minimaux et l'indice PDO, tandis que pour les trois stations situées dans le sud de l'Ontario, cette relation est positive. En observant l'ensemble de données quotidiennes des débits dans ces stations, nous avons constaté que les valeurs de débit minimal dans les stations du nord sont plus souvent observées en fin d'hiver, généralement entre mars et avril. Cependant, les valeurs de débit minimal dans les autres stations sont souvent observées pendant l'été ou l'automne, surtout entre juillet et novembre.

— Ces covariables ont été utilisées dans le modèle de régression des quantiles avec les B-Splines afin d'estimer les quantiles conditionnels. Pour la plupart des stations étudiées, un modèle de régression des quantiles avec des fonctions B-Splines ayant un nombre de nœuds et un degré supérieurs à 2 a été sélectionné par les critères utilisés. Les résultats de la comparaison sont montrés dans les tableaux 2.2, 2.3, 2.4 et 2.5 de l'article 2. Nous avons aussi remarqué qu'il est préférable d'employer le critére BIC au lieu du coefficient de détermination pour sélectionner le modèle avec les fonctions B-Splines. En effet, le coefficient de détermination semble surestimer l'ajustement du modèle. Les Figures 2.7 et 2.8 montrent les résultats des quantiles conditionnels estimés pour les probabilités 0.5 et 0.9 (temps de retour 2 et 10 ans) respectivement en fonction des covariables AMO et PDO. Nous avons constaté que les quantiles conditionnels peuvent prendre des valeurs beaucoup plus élevées que les quantiles stationnaires. Par exemple, pour la station 02AC001 les résultats montrent que la médiane stationnaire est égale à 48 (m^3/s), cependant la médiane non stationnaire peut atteindre 140 (m^3/s). Cette différence entre les valeurs estimées des quantiles stationnaires et des quantiles non stationnaires devient plus importante pour des périodes de retour plus élevées. En effet, pour la même station, le quantile stationnaire pour la probabilité 0.9 est égal à 96.9 (m^3/s), par contre le quantile non stationnaire peut atteindre $250(m^3/s)$. Les résultats des quantiles stationnaires sont donnés dans le tableau 2.1. Des résultats similaires ont été remarqués pour toutes les autres stations étudiées, ce qui confirme l'importance de l'utilisation des modèles complexes afin d'assurer une meilleure qualité d'estimation des événements hydrologiques extrêmes.

4.3 Résultats et conclusions de l'objectif 3

Dans ce travail, nous avons proposé une nouvelle approche pour l'estimation des quantiles conditionnels qui prend en considération une structure de dépendance donnée par la fonction copule. Deux estimateurs ont été suggérés: un estimateur paramétrique pour la fonction des quantiles conditionnels et qui suppose un modèle paramétrique pour la fonction copule et pour les distributions marginales et un estimateur semi-paramétrique pour la fonction des quantiles conditionnels en supposant un modèle paramétrique pour la fonction copule et un modèle non paramétrique pour les distributions marginales. Tout d'abord, nous avons démontré la convergence asymptotique de nos estimateurs dans le but de connaître leurs distributions asymptotiques et de construire les intervalles de confiance. Les résultats de la convergence sont donnés par les théorèmes 1 et 2 de l'article 3. Les preuves de ces théorèmes sont détaillées dans la section 0.6 de l'annexe III. Une comparaison entre nos estimateurs et ceux proposés en littérature a été réalisée en se basant sur des modèles de simulation. Ces modèles de simulation suggèrent différentes familles de copules archimédiennes et elliptiques ainsi que différentes distributions marginales pour la variable d'intérêt et les covariables afin d'évaluer la variabilité des estimateurs proposées. Les résultats des simulations ont montré l'efficacité de nos estimateurs et une bonne performance de notre estimateur paramétrique. Le tableau 3.1 donne plus de détails sur ces résultats. Deux études de cas ont été fournies afin de bien montrer l'utilité de nos estimateurs, une étude qui traite les précipitations annuelles dans une station en Californie en présence de deux indices climatique (SOI et PDO) et une autre qui étudie les débits maximaux annuels dans une station en Ontario en présence d'un seul indice climatique (AMO). Les figures 3.7 et 3.8 montrent les résultats de l'estimation des quantiles conditionnels. Nous avons également produit des résultats de quantiles conditionnels pour ces deux cas d'études en se basant sur les modèles GEV-B-Splines et la régression des quantiles avec les B-Splines (voir Figures 3.4 et

3.6). En comparant ces figures, nous avons remarqué que la copule donne une estimation plus lisse et moins bruitée par rapport aux méthodes basées sur le lissage B-Splines. Ceci est dû probablement au fait que la copule prend en considération la dépendance globale entre les données, par contre les fonctions B-Splines considèrent des dépendances par morceaux.

Chapitre 5

Conclusions et perspectives

L'étude de l'estimation des quantiles présente une grande importance dans le domaine de l'hydro-climatologie, parce qu'elle apporte de l'information sur les risques des extrêmes hydrologiques. Les deux dernières décennies ont vu un grand développement de la modélisation statistique de ces extrêmes. Cette modélisation est passée des modèles classiques stationnaires aux modèles non stationnaires généralement utilisés avec des dépendances linéaires ou quadratiques. Cette thèse de doctorat avait pour but de développer/adapter de nouveaux modèles non stationnaires en présence des dépendances non nécessairement polynomiales. Nous avons proposé alors trois approches, deux sont basées sur des dépendances de type B-Splines et une est basée sur les copules. Ces approches ont été comparées aux différents modèles proposés dans la littérature et aussi comparées entre elles afin de choisir le meilleur modèle. Les résultats de comparaison ont montré l'avantage de l'utilisation de nos approches pour l'estimation des quantiles non stationnaires et plus spécifiquement l'approche basées sur les copules.

Malgré la performance de nos approches proposées, elles restent limitées en terme de la dimension de la variable d'intérêt. En effet, dans les trois approches proposées, nous supposons que la variable d'intérêt est de dimension un et conditionnelle à plusieurs covariables. Pourtant en hydro-

climatologie, les évènements extrêmes sont généralement définis pour plusieurs variables d'intérêt. Par exemple, une crue peut se définir en fonction de la pointe (débit maximal), la durée et le volume. Donc, il serait important de développer des modèles qui prennent en considération plusieurs variables d'intérêt à la fois et plusieurs covariables afin de mieux estimer le risque des extrêmes hydrologiques tout en tenant compte des aléas climatiques. Ces modèles peuvent être tout simplement une généralisation aux approches proposées dans cette thèse, il suffit de bien définir la fonction des quantiles conditionnels multivariés qui vont dépendre de différentes probabilités liées aux différentes variables d'intérêt étudiées.

Aussi, ce qui pourrait être intéressant comme suite à nos approches proposées serait de développer des techniques qui permettent la sélection du meilleur modèle de prédiction en présence de plusieurs covariables. Plusieurs techniques peuvent être utilisées dans ce cadre, nous citons comme exemple la régression LASSO (Least absolute Shrinkage and selection operator) [Tibshirani, 1996] ou la méthode de sélection bayésienne [Scheipl, 2011].

Deuxième partie

Articles

Chapitre 1

Atmospheric Predictors for Annual Maximum Precipitation in North Africa

Titre traduit

Prédicteurs atmosphériques pour le maximum annuel des précipitations en Afrique du Nord.

Auteurs

Bouchra Nasri[1], Yves Tramblay[2], Salaheddine El Adlouni [3], Elke Hertig [4] et Taha B.M.J Ouarda [1,5]

[1] Institut national de recherche scientifique, Eau-Terre-Environnement, Quebec, Canada.

[2] IRD, UMR Hydrosciences-Montpellier, France.

[3] Université de Moncton, Département de Mathématique et de Statistique NB, Canada

[4] Institute of Geography, University of Augsburg, Germany.

[5] Masdar Institute of Science and Technology, Abu Dhabi, UAE.

Contribution des auteurs

Bouchra Nasri: rédaction théorique de la méthode, élaboration des codes pour l'estimation bayésienne, traitement des données, l'analyse des résultats et la réaction de l'article.

Yves Tramblay: proposition du projet, collecte des données, révision des résultats, révision de l'article.

Salaheddine El Adlouni: révision de la partie statistique de l'article, élaboration des codes du modèle GEV-B-Splines, révision de l'article.

Elke Hertig: révision de la partie climatologie de l'article.

Taha B.M.J. Ouarda: lecture et révision de l'article.

Remerciements

Je tiens à remercier mes co-auteurs pour la collaboration et la réussite de ce travail.

Article publié

Journal: Journal of Applied Meteorology and Climatology

Date de publication: Avril 2016

Résumé

La forte variabilité des précipitations en Afrique du Nord représente un défi majeur pour la population et les infrastructures de la région. Les dernières décennies ont vu de nombreuses inondations provoquées par les précipitations extrêmes dans ce secteur. Il existe, donc, un grand besoin d'identifier les prédicteurs atmosphériques les plus pertinents pour modéliser ces événements extrêmes. Dans le présent travail, on évalue l'effet de 14 différents prédicteurs calculés à partir des données réanalysées NCEP-NCAR, avec des échelles quotidiennes et saisonnières, sur les précipitations maximales annuelles (MAP) dans six stations côtières situées en Afrique du Nord (Larache, Tanger, Melilla, Alger, Tunis et Gabès). Le modèle GEV-B-spline a été utilisé pour détecter cette influence. Ce modèle considère toutes les formes de dépendance continue (linéaire, quadratique et autres) entre les

covariables et la variable d'intérêt, fournissant ainsi un cadre très flexible pour évaluer les effets des covariables sur les paramètres de la distribution GEV. Les résultats montrent qu'aucun ensemble unique de covariables n'est valide pour toutes les stations. Dans l'ensemble, une forte dépendance entre les prédicteurs NCEP-NCAR et MAP a été détectée, en particulier avec des prédicteurs décrivant la circulation à grande échelle (hauteur géopotentielle) ou l'humidité. Cette étude peut donc fournir des aperçus pour développer des modèles de réduction des précipitations extrêmes adaptés aux conditions de l'Afrique du Nord.

Abstract

The high precipitation variability over North Africa presents a major challenge for the population and the infrastructure in the region. The last decades have seen many flood events caused by extreme precipitation in this area. There is a strong need to identify the most relevant atmospheric predictors to model these extreme events. In the present work, the effect of 14 different predictors calculated from NCEP–NCAR reanalysis, with daily to seasonal time steps, on the maximum annual precipitation (MAP) is evaluated at six coastal stations located in North Africa (Larache, Tangier, Melilla, Algiers, Tunis, and Gabes). The generalized extreme value (GEV) B-Splines model was used to detect this influence. This model considers all continuous dependence forms (linear, quadratic, etc.) between the covariates and the variable of interest, thus providing a very flexible framework to evaluate the covariate effects on the GEV model parameters. Results show that no single set of covariates is valid for all stations. Overall, a strong dependence between the NCEP–NCAR predictors and MAP is detected, particularly with predictors describing large-scale circulation (geopotential height) or moisture (humidity). This study can therefore provide insights for developing extreme precipitation downscaling models that are tailored for North African conditions.

Keyword

GEV, B-Splines, NCEP-NCAR, North Africa, precipitations, covariates, extreme.

1.1 Introduction

Heavy precipitation events are causing extensive damage to the populations and infrastructure of the countries located in the southern part of the Mediterranean basin. The last decades saw several deadly flood events caused by extreme precipitation, including the 2001 flood near Algiers, Algeria, causing more than 700 fatalities [Argence *et al.*, 2008], the 1969 floods in the region of Kairouan, Tunisia, with 150–400 fatalities [Preisendorfer, 1988a], or the 1995 flood in the Ourika valley, Morocco, with more than 200 fatalities [Saidi *et al.*, 2003]. To better mitigate the impacts of these extreme events, there is a need to evaluate their predictability on different time scales. In particular, it is necessary to estimate their return periods in a climate change context since several countries experienced an increased vulnerability to these events during the last decade [Di Baldassarre *et al.*, 2010]. Several recent studies have focused on seasonal precipitation and its extremes in the Mediterranean region, with the objective of identifying the associated large-scale patterns and the relevant predictors [Knippertz *et al.*, 2003; Xoplaki *et al.*, 2004; Martin-Vide & Lopez-Bustins, 2006; Toreti *et al.*, 2010; Tramblay *et al.*, 2011; Kallache *et al.*, 2011; Hertig *et al.*, ges ; Tramblay *et al.*, 2012; Hertig *et al.*, 2013; Ouachani *et al.*, 2013; Donat *et al.*, 2014]. Indeed, to resolve the mismatch of scales between general circulation models and the locations of interest for impact studies, there is a need to develop downscaling techniques tailored for extreme precipitation [Fowler *et al.*, 2007; Maraun *et al.*, 2010]. To overcome the limitations of climate models in reproducing extremes [Sillmann *et al.*, 2013], several studies have used covariates in nonstationary extreme precipitation frequency analysis [Vrac & Naveau, 2007; Aissaoui-Fqayeh *et al.*, 2009; Beguería *et al.*, 2011; Friederichs, 2010; Kallache *et al.*, 2011; Tramblay *et al.*, 2011; Maraun *et al.*, 2010; Ouachani *et al.*, 2013; El Adlouni & Ouarda, 2009; Cannon, 2010; Ouarda & Adlouni, 2011]. However, even if several authors have shown the efficiency of atmospheric humidity and moisture flux as predictors

for daily rainfall modeling and downscaling [Cavazos & Hewitson, 2005; Bliefernicht & Bárdossy, 2007; Tramblay *et al.*, 2011, 2013], the best predictors may differ from one site to another [Kallache *et al.*, 2011; Hertig *et al.*, 2013; Chandran *et al.*, 2016]. In addition, it should be noted that the above studies have mostly applied polynomial dependence (linear or quadratic) between the covariate and the variable of interest. In the present work, the nonstationary generalized extreme value (GEV) model with B-Splines dependent function [Chavez-Demoulin & Davison, 2005; Nasri *et al.*, 2013] is applied. B-Splines functions are piecewise polynomial functions that have certain advantages. A smoothing B-Splines basis is independent of the response variable and depends only on the following information: (i) the extent of the explanatory variable, (ii) the number and position of the knots, and (iii) the degree of the B-Splines. These advantages make it a suitable option for use in the GEV model with covariates to explain the effect of covariates on the response variable. Therefore, the goal of this study is to identify relevant large-scale predictors influencing the annual maximum precipitation at coastal stations in the southern part of the Mediterranean region using the GEV B-Splines model. This model will help describe the predictors' influence on the precipitation records within the study period. A number of studies on the impact of climate variability on extreme precipitation in the Mediterranean region have employed atmospheric–oceanic teleconnection indices such as the North Atlantic Oscillation (NAO; Wanner *et al.* [2001], the Mediterranean oscillation [Conte *et al.*, 1989], or the western Mediterranean oscillation (WEMO) [Knippertz *et al.*, 2003; Vicente-Serrano *et al.*, 2009]. However, in Morocco Tramblay *et al.* [2012] observed a possible dependence of precipitation extremes with these indices only at 2 stations (Larache and Tangier) out of 10. In the present study, we tested the effect of different predictors measured with NCEP–NCAR reanalysis data to evaluate their influence on the maximum annual daily precipitation (MAP) time series. To our knowledge, no studies have previously examined these extreme precipitation events and their relationship to large-scale atmospheric influences in this area at the daily time step with rain gauge

data, mainly because of the limited access to the data. Since reanalysis data are available at a spatial scale similar to that of global circulation models (GCMs), this study provides the first step toward the development of extreme precipitation downscaling methods that are tailored for North African conditions.

1.2 Datasets

In the present study, we collected long daily precipitation time series maintained by the governmental hydrological services of Algeria [Agence Nationale des Ressources Hydrauliques (ANRH)], Morocco [Direction de la Recherche et de la Planification de l'Eau (DRPE)], and Tunisia [Direction Générale des Ressources en Eau (DGRE)]. The daily data of the Melilla station located in northern Morocco were obtained from the European Climate Assessment and Dataset (ECAD; `http://eca.knmi.nl`). The data from each station were carefully scrutinized, in particular to look for shifts, absurd values, and missing data [Tramblay *et al.*, 2013]. The stations that were subsequently selected had less than 5% missing days between September and May. The years with more than 5% missing days during this period were discarded. Figure 1.1 illustrates the geographic location of all stations, and Table 1.1 presents a description of the selected stations with long precipitation records. Reanalysis data from the National Centers for Environmental Prediction (NCEP; Kalnay *et al.* [1996]; Kistler *et al.* [2001]) are used to compute several large-scale predictors. Various variables were extracted to be tested in the model; the selection of covariates is based on the previous studies of Cavazos & Hewitson [2005], Kallache *et al.* [2011], Tramblay *et al.* [2011], and Hertig *et al.* [2013]. The NCEP–NCAR reanalysis data have been selected over more recent products such as ERA-Interim because the time span of the NCEP–NCAR reanalysis is larger and encompasses the whole period of observations, up to the present. The advantages of recent reanalysis products are

manifold, including new atmospheric and assimilation systems and finer grid spacing. However, they cover only the recent period (from 1979 to present for MERRA, CFS reanalysis, or ERA-Interim; Hofer *et al.* [2012]). Thus the advantage of the choice of the NCEP–NCAR reanalysis is to cover the whole period where observations are available using a single reanalysis product in order to identify possible associations with large-scale climate dynamics. It must be noted that our goal is not to check the adequacy of the particular NCEP–NCAR reanalysis product but to evaluate if the distribution of extreme precipitation can be related to large-scale predictors, if the same predictors are valid for different stations, and at which time scales the large-scale forcings are relevant. For these objectives, the bias of a given reanalysis product compared to other products is of little relevance, and because of the strong interannual variability of precipitation in North Africa, it is important to evaluate the relationships with large-scale dynamics over long time periods to obtain robust results. The gridded NCEP–NCAR data (six grid cells covering the study area) were interpolated by the inverse distance method to the station locations in order to provide individual descriptor sets for each station. The selected variables include the following:

— geopotential height at 500 and 850 hPa (geopot_500 and geopot_850),

— vertical velocity at 500 and 850 hPa and at surface (omega_500, omega _850, and omega _surf),

— potential temperature at surface (ptemp_surf),

— precipitable water content (pwater_surf),

— relative humidity at surface (rhum_surf),

— specific humidity at 500 and 850 hPa (shum_500 and shum_850),

— mean sea level pressure (slp_surf),

— surface temperature (temp_surf),

— zonal wind at surface (uwind_surf),

— meridional wind at surface (vwind_surf)

The homogeneity of these covariates has been assessed by following [Pettitt, 1988] and by using the modified version of the standard normal homogeneity test (SNHT) by [Khaliq & Ouarda, 2007]. Indeed, the gradual introduction of satellite data into reanalysis products can introduce an artificial change point leading to the false detection of trends or homogeneity breaks (Sterl 2004). The Pettitt and SNHT tests agree only on a significant changepoint at the 5% level in relative humidity, in 1957 for SNHT and in 1963 for the Pettitt test. Therefore, no changepoints are detected in the beginning of the 1980s following the introduction of satellite data. These covariates are considered at different time steps. In the first case, we considered the maximum observed daily precipitation during the extended winter season (October–March) and the simultaneous daily covariate in the reanalysis data associated with this extreme rainfall event. This gives one observation of maximum daily winter rainfall and its associated covariate for each year (hereafter case 1). In case 2, we considered the maximum winter precipitation and the average value of each covariate 5 days before the date of the annual maximum rainfall during the winter. These two cases can be considered as dealing with the shortterm effect of covariates on MAP. In case 3, we calculated the 30-day average of the covariate before the date of maximum winter precipitation. Finally, in case 4, we considered the maximum daily winter precipitation and the value of each covariate for the entire season (October–March average). These last two cases can be considered as dealing with the long-term effect of covariates on MAP.

1.3 Methods

For modeling extreme rainfall events, we used the GEV distribution [Coles, 2001]. The role of the GEV distribution is to describe a sample that follows a maximum of distributions introduced by Fisher & Tippett [1928]. The GEV distribution is flexible and has been the subject of several theoretical studies and applications for modeling extreme flood, precipitation, and wind events [El Adlouni *et al.*, 2007; Hundecha *et al.*, 2008]. The development of stationary GEV distribution models for univariate extreme value analysis can be found in the literature [Coles, 2001; Olsen *et al.*, 1999]. The use of this distribution in the frequency analysis of extreme events is based on a number of specific hypotheses concerning the variable of interest. Indeed, the observations must be independent and identically distributed. However, the stationarity assumption is often not met for observed hydroclimatic datasets [Khaliq *et al.*, 2006]. In this case, the distribution parameters and the distribution itself could be changing in time. Therefore, it is essential to develop the GEV model in the multivariate space, where extreme events can be associated with other variables. To model the relationship between the covariates and the extreme variable of interest, we can use the GEVB-Splines approach [Nasri *et al.*, 2013]. This approach has been developed to describe the association of an external covariate with the variable of interest. The estimation of the parameters of the GEV B-Splines model is done in a Bayesian framework to obtain the posterior distribution by applying Markov chain Monte Carlo (MCMC) algorithms.

1.3.1 The GEV distribution

The GEV distribution is characterized by three parameters: the location μ, scale σ, and shape ξ parameters. Depending on the value of the shape parameter we have three types of extreme value distributions namely, the Gumbel ($\xi = 0$), Fréchet ($\xi > 0$), and Weibull ($\xi < 0$). Considering

a sample $Y = (y_1, y_2, \ldots, y_n)$, the GEV distribution function is as follows

$$G_{\mu,\sigma,\xi}(y) = \begin{cases} \exp\left[-\left(1+\xi\left(\frac{y-\mu}{\sigma}\right)_+^{-\frac{1}{\xi}}\right)\right] & \text{if} \quad \xi \neq 0 \\ \exp\left[-\exp\left(-\left(\frac{x-\mu}{\sigma}\right)_+\right)\right] & \text{if} \quad \xi = 0 \end{cases}, \qquad y_+ = \max(0, y) \tag{1.1}$$

This classical GEV distribution is based on the stationarity assumption and does not consider the dependence of extreme events on other variables. In the following section, the nonstationary GEV approach is presented to consider the effect of a covariate on extreme values.

1.3.2 The nonstationary GEV-B-Splines model

In the nonstationary case of a GEV distribution, the parameters of the GEV distribution are assumed to change in time or depend on covariates. In the present form of the GEV, parameters σ and ξ are assumed to be constant. Having a random variable Y that follows the GEV($\mu_{\boldsymbol{X}}, \sigma, \xi$) and a vector of p covariates given by $\boldsymbol{X} = (X_1, X_2, \ldots, X_p)$, the location parameter of the GEV is written as follows:

$$\mu_{\boldsymbol{X}} = \sum_{i=1}^{p} f_i(X_i) = f_1(X_1) + \ldots + f_p(X_p). \tag{1.2}$$

where f_i is a function that represents the relationship between the parameter and the covariates X_i. This function can be described by the following B-Splines function

$$f_i(X_i) = \sum_{j=1}^{m} \beta_j B_{i,j,k}(X_i) \qquad i = 1 \ldots p \tag{1.3}$$

where $B_{j,d}(x)$ is a polynomial function of degree d and m is the number of control points (or knots) [Nasri *et al.*, 2013]. Therefore, 1.2 can be rewritten as follows:

$$\mu_{\boldsymbol{X}} = \sum_{i=1}^{p} f_i(X_i) = \sum_{i=1}^{p} \sum_{j=1}^{m} \beta_j B_{i,j,k}(X_i). \tag{1.4}$$

The predictors' interaction can be expressed in our model by using multivariate B-Splines functions [De Boor, 2001]. These functions allow considering the correlation between the predictors. In this study 14 predictors are used. Consequently, in order to simplify the model we did not consider the interaction between predictors.

1.3.3 Parameter estimation

In this study, the estimation of the parameters of the GEV-B-Splines model is carried out in a Bayesian framework. In the Bayesian approach, the unknown parameters are not constant and are considered as random variables with a prior distribution $\pi(\boldsymbol{\theta})$. Bayes's theorem therefore gives the following definition of the posterior distribution of these parameters:

$$\pi(\boldsymbol{\theta}|y) \propto L(y|\boldsymbol{\theta})\,\pi(\boldsymbol{\theta}). \tag{1.5}$$

where $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma, \xi)$. According to [Nasri *et al.*, 2013], we choose a multivariate normal distribution $f_{\mathscr{N}}$ as a prior for the location parameter $\boldsymbol{\beta} \sim \mathscr{N}\left(\mathbf{0}, \sum_{\beta} I\right)$, a noninformative prior for scale parameter $\frac{1}{\sigma}$, and a Beta distribution $f_{\mathscr{B}}$ as prior for the shape parameter $\mathscr{B}(6, 9)$. The posterior

distribution is written as follows

$$\pi\left(\boldsymbol{\theta}|y\right) \propto \prod_{i=1}^{n_1} \frac{1}{\sigma} \exp\left[-\left(1-\xi\left(\frac{y_i - \sum_{k=1}^{p}\sum_{j=1}^{m}\beta_j B_{j,k}\left(x_i\right)}{\sigma}\right)\right)^{-\frac{1}{\xi}}\right]\left[1-\xi\left(\frac{y_i - \sum_{k=1}^{p}\sum_{j=1}^{m}\beta_j B_{j,k}\left(x_i\right)}{\sigma}\right)\right]^{-1+\frac{1}{\xi}}$$
$$\times \prod_{i=n_1+1}^{n} \frac{1}{\sigma} \exp\left[-\left(\frac{y_i - \sum_{k=1}^{p}\sum_{j=1}^{m}\beta_j B_{j,k}\left(x_i\right)}{\sigma}\right)\right] \exp\left[-\exp\left(-\left(\frac{y_i - \sum_{k=1}^{p}\sum_{j=1}^{m}\beta_j B_{j,k}\left(x_i\right)}{\sigma}\right)\right)\right]$$
$$\times f_{\mathscr{N}}\left(\boldsymbol{\beta}; \mathbf{0}, \sum_{\beta} I\right) \times f_{\mathscr{B}}\left(\xi; 6, 9\right) \times \frac{1}{\sigma} \quad , \quad (1.6)$$

where n_1 is the number of observations such that $\xi \neq 0$.

The posterior distribution is estimated by the Metropolis–Hasting algorithm (see 0.2). To select the number of knots and the degree of B-Splines functions used in this study, we compared several combinations of degrees and knots using the maximum likelihood method. The following algorithm explains how these parameters are chosen:

— set $d \in [1, 10]$ and $m \in [1, 10]$,

— calculate Eq. 1.6 for all combinations of (d, m),

— choose values of (d, m) that maximize Eq. (1.6).

In this case, we apply the B-Splines functions with $d = 3$ and $m = 3$. This choice was found to be optimal for the majority of the stations' data.

1.3.4 Validity of the model with covariates

To validate the influence of covariates on the variable of interest, the log likelihood of the GEV B-Splines (M1) model and the stationary GEV (M0) model (without covariates) are compared using

the test of deviance:

$$D = 2\left[l\left(M_1\right) - l\left(M_0\right)\right], \tag{1.7}$$

where l is the maximum log likelihood function for model M. Large values of D indicate that model M_1 is more adequate at representing the data than model M_0. The D statistic is distributed according to a chi-square distribution χ^2, with v degrees of freedom, where v is the difference between the number of parameters of the M_1 and M_0 models. For a given confidence level, we reject H_0 hypothesis (H_0: M_1 and M_0 are similar) when $D > \chi^2_{1-\alpha}$. This statistic is often used to compare two models when one model is a special case of the other ($M_0 \in M_1$; [Coles, 2001; El Adlouni & Ouarda, 2009]). This test accounts for differences in model complexity to avoid overfitting.

1.3.5 Quantile estimation

The MCMC algorithm also produces the conditional quantile distribution for an observed value x_0 of the covariate X_i . Indeed, for each iteration t of the MCMC algorithm $t = 1, \ldots, N$, the quantiles corresponding to the nonexceedance probability $1 - p$; x^t_{p,x_0} and the parameter vector $\left[\mu^{(t)}_{x_0}, \sigma^{(t)}, \xi^{(t)}\right]$ are estimated using the inverse of the cumulative distribution function of the GEV distribution:

$$Q^{(t)}_p\left(Y|X_i = x_0\right) = \mu^{(t)}_{x_0} - \frac{\sigma^{(t)}}{\xi^{(t)}}\left(1 - \left(\log\left(1-p\right)\right)\right)^{\xi^{(t)}}. \tag{1.8}$$

where $\mu^{(t)}_{x_0}$ is the position parameter conditional on the particular value x_0 of X.

1.4 Results

1.4.1 Tests for independent and identically distributed random variables

In the first step of using the nonstationary GEV model (in this case GEV-B-Splines) we checked stationarity, homogeneity, and independence using the Mann–Kendall [Mann, 1945], Mann–Whitney [Wilcoxon, 1945], and Wald–Wolfowitz tests [Wald & Wolfowitz, 1940], respectively, for MAP series for each station. The results of these tests showed that all MAP series are nonstationary at the 5% level. However, all the time series of MAP respect the hypotheses of homogeneity and randomness. Figure 1.2 shows the variation of all MAP series versus time, and Figure 1.3 shows the monthly frequency of occurrence of annual maximum daily precipitation.

1.4.2 Predictors from reanalysis data

We selected 14 NCEP covariates extracted from reanalysis (see section 1.2) and developed our models with these covariates considering the four time scales (case 1 to 4). The negative log-likelihood and deviance between the model GEV-B-Splines and the stationary GEV model are analyzed to detect the influence of NCEP predictors on extreme rainfall for each of the four cases. To avoid overfitting, each covariate is considered separately in the GEV-B-Splines model, to evaluate if it provides a better fit than a stationary GEV model. As there are 14 covariates for each case, the results are presented in Table 1.2 only for the significant covariates on MAP at each station at the 5% and 10% significance levels, according to the test of Deviance. We note that all 14 covariates, depending on the station, are selected into non-stationary GEV models that better reproduce extreme precipitation than a standard stationary model, in both short-term and long-term association. Overall, a similar number of significant covariates is selected for the 4 cases tested

(i.e. daily to seasonal averages of covariates), with 11 covariates identified for case 1, 13 for case 2, 16 for case 3, 13 for case 4. This shows that all covariates tested may have an impact on extreme daily precipitation at different time steps, from daily values to seasonal averages.

It is observed that the geopotential height (geopot_500, geopot_850) usually affects rainfall at all stations excluding Melilla (station in Northern Morocco). For the two stations of Tangier and Larache, the geopotential heights have a short-term association with MAP (case 1 and case 2). In the opposite, for the Algiers station, these variables have an influence generally at the seasonal time scale (case 4), and for the stations of Tunisia (Gabes and Tunis) these variables influence MAP in both, short and long-term (cases 1, 2 and 4). The humidity predictors (rhum_surf, shum_500, shum_850) generally influence precipitation at all stations, excluding Algiers. For stations in Morocco, these predictors appear in almost all cases (cases 1 and 3 for the Tangier station, case 4 for Melilla station and cases 1, 2 and 3 for Larache station). For stations in Tunisia, these predictors have both short and long-term influence on MAP time series at Gabes station (cases 1, 3 and 4) and Tunis station (cases 1, 2 and 3). The velocity predictors (omega_500, omega_850, omega_surf) have more effects on precipitation in Morocco. We see a strong influence of these predictors on rainfall in Morocco. However, that is not observed for Algiers and Gabes stations. Wind predictors (uwind_surf, vwind_surf) influence the MAP only at the Tangier station. Overall, we note the small influence of wind covariates on precipitation extremes in all stations. The potential temperature at the surface (ptemp_surf) influences MAP of stations in Morocco in the long-term cases (cases 3 and 4). The surface temperature influences the MAP at stations in Morocco at different time scales (case 3 for Melilla, case 1 for Larache and case 4 for Tangier stations). The precipitable water content has an influence on MAP at all stations, usually only for the short-term cases (1 and 2) in all stations. The mean sea level pressure only influences MAP in the Tunis, Gabes and Algiers stations, generally, in the long-term cases 3 and 4.

1.4.3 Principal analysis of components for NCEP–NCAR predictors

After the analysis of the dependence of MAP with individual covariates, the possible relationships are also investigated in a multivariate context. Principal component analysis (PCA; [Preisendorfer, 1988b]) is used to that end. The reason for using PCA is to take into consideration the common signals in multivariate datasets. PCA represents a method for dimensionality reduction. PCA has been used for this purpose in many other studies [Wetterhall *et al.*, 2005; Maraun *et al.*, 2010]. The objective of this analysis is to summarize as much information as possible by transforming interrelated variables into new components (principal components) that are uncorrelated with each other. In this study, we first applied PCA on the 14 covariates for each station. Figures 1.4 and 1.5 show the results of the projections of the 14 covariates on the first and second components (F1 and F2, respectively) for the first and last case in each station. A number of criteria, such as the Kaiser criterion [Kaiser, 1960], can be used for the selection of the factorial axis. The Kaiser criterion lies on the factorial axis choices, where their eigenvalues are greater than 1. In the present study, we noticed that for all factors that have an eigenvalue greater than 1, those are generally factors 1 and 2. This justifies the choice of two factorial axes. It can be seen that at all stations, there are significant correlations between the covariates, depending on the case (1–4) considered for temporal aggregation. To avoid overfitting, each component is considered separately in the GEV B-Splines model to evaluate if it provides a better fit than a stationary GEV model. The results show, first, that most variables contribute to the formation of the components F1 and F2, with some covariates having a larger contribution, such as the geopotential height (geopot_500 and geopot_850), velocity (omega_500, omega_850, and omega_surf), and humidity (rhum_surf, shum_500, and shum_850). Predictors such as uwind_surf and vwind_surf contribute more in stations close to the Mediterranean coast such as Tangier, Tunis, and Gabes. We then applied the GEV-B-Splines

model of the MAP series for each station and each case using F1 and F2 as covariates. Next, we calculated the deviance between the results of the GEV B-Splines and GEV0 models to investigate the influence of these components on MAP data. Table 1.3 shows the results of deviance with a threshold of 5% and 10%. According to these results one can see that, at all stations, there is at least one component (F1 or F2) that influences the MAP series. For stations in Morocco (Larache, Melilla, and Tangier), we note that components that contains more information about the geopotential height, humidity, velocity, and wind has the largest influence on MAP. For the station in Algeria (Algiers), the MAP is more influenced by the geopotential height predictors rather than by others. For stations in Tunisia (Tunis and Gabes), we can see that the influence of geopotential height, velocity, temperature, and sea surface pressure on MAP is important.

1.4.4 Quantile estimation

We can also see the impact of the covariates on the estimated quantile level for each of the models. In the case of the GEV B-Splines model, quantile values depend not only on the nonexceedance probability $1 - p$ but also on the covariate values. This allows computing quantiles on a seasonal or annual basis, depending on the values of the covariates. To demonstrate the covariates' impact on quantile values, we show some quantile estimation examples for each station. Figure 1.6 displays a nonstationary quantile estimation example for each station for the 2-yr return period (nonexceedance probability $1 - p = 0.5$), which represents the median value of MAP. For each station, we observed different values of the 2-yr quantiles estimated with the GEV-B-Splines model since quantile values are dependent on covariates. In contrast, the GEV0 model provides just one estimate for the 2-yr quantile (e.g., For each station, we observed different values of the 2-yr quantiles estimated with the GEV-B-Splines model since quantile values are dependent on covariates. In contrast, the GEV0 model provides just one estimate for the 2-yr quantile. For instance, for the

Algiers station, the median precipitation value corresponding to the 2-yr quantile is 100mm for the GEV-B-Splines model and 64mm for the GEV0 model for values of covariates F1 defined in case 4, which essentially includes humidity and temperature covariates. For the Tunis station, the stationary quantile is equal to 50mm and the median of the nonstationary quantiles is equal to 70mm for values of F2 associated with case 3, which essentially includes wind velocity and temperature as covariates.). According to this figure, we notice that the covariate dependent quantile values are more flexible and allow reaching more extreme data values, unlike the stationary quantiles that do not take into consideration the interannual climatic variability. The estimated quantiles show the advantage of incorporating additional information into nonstationary models.

1.5 Conclusions

In this work, the influences of climatic variables such as geopotential height, pressure, or temperature on maximum annual daily precipitation have been studied at six stations located in North Africa with long precipitation time series. A total of 14 variables were computed from NCEP–NCAR reanalysis data. To study the influence of these covariates at the different stations, the GEV-B-Splines model [Nasri *et al.*, 2013] was used. The originality of this model, as opposed to other nonstationary models, is that it takes into consideration the nonstationary and the nonlinear temporal fluctuations of covariates. Nonstationary models, such as the GEV1 (linear dependence) and the GEV2 (quadratic dependence), define in advance the form of dependence between the variable of interest and the covariates. On the other hand, the GEV-B-Splines model takes into consideration all continuous dependence forms between the covariates and the variable of interest. The results of this study are divided into two parts. In the first part, the possible dependencies between the maximum annual precipitation and each of the individual climatic covariates were considered.

The GEV-B-Splines model was used to detect these dependencies, and the deviance likelihood ratio test was used to identify the nonstationary models with covariates that provide an improvement in comparison to stationary models in each station. In the second part, the combined dependencies were analyzed using principal component analysis of the different atmospheric predictors. From the results of the principal component analysis, we analyzed the influence of the combined variables using the two principal components (F1 and F2) for each station in the GEV-B-Splines model. Our results indicate that no single combination of atmospheric predictors is optimal for stations. The relevant covariates may vary from one station to another and also depend on the considered time scale, from daily to annual averages. These results are consistent with the fact that extreme precipitation is a process exhibiting a high spatiotemporal variability between different locations. Given this variability, it must be noted that the covariates describing the moisture flux in the atmosphere (relative or specific humidity) or in atmospheric circulation (pressure and geopotential heights) are often selected in the different stations as valid predictors. During winter, when most of the annual maximum precipitation occurs, geopotential height might be more important because of the southerly position of the extratropical westerlies. In other seasons thermodynamic predictors like humidity may gain significance because of the convective nature of precipitation in these seasons. The present work provides a first step prior to the development of statistical downscaling methods tailored for extreme precipitation in North Africa. The next step would be to use GCM outputs to first validate the method in the present climate, with the covariates that are correctly reproduced in historical climate simulations, and then to make future projections. However, in this case the use of the nonstationary GEV model with B-Splines functions would probably be less appropriate because of some limitations: (i) the introduction of several covariates within these types of models increases the number of hyperparameters, which increases the number of parameters to estimate as well as the estimation errors; (ii) the interactions between the predictors make the model much more complex

since we need to take into consideration multivariate spline functions [De Boor, 2001] or use some decisional model, such as an artificial neural network as in [Cannon, 2010]; and (iii) this type of model allows the description of the impact of covariates on the variable of interest and is not able to use them for prediction outside this period of study. Consequently, an alternative to this type of model is quantile regression methods [Buchinsky, 1998]. Unlike linear regression, which results in the estimation of the conditional mean for the response variable given certain values of predictor variables, quantile regression aims at estimating either the conditional median or other quantiles of the response variable. Quantile regression was considered by Jagger & Elsner [2006] for wind speed and by Friederichs & Hense [2008] for precipitation, based on several climatic covariates. Future work can focus on the comparison of extreme value models and the quantile regression approach to distinguish the relative benefits of the use of these two types of models for downscaling purposes.

Acknowledgments

The datasets were provided by the Agence Nationale des Ressources Hydrauliques (Algeria), Direction de la Recherche et de la Planification de l'Eau (Morocco), Direction Générale des Ressources en Eau (Tunisia), and European Climate Assessment and Dataset. Special thanks are given to H.Ben-Mansour,R.Bouaicha, L. Behlouli,K.Benhattab, R. Taibi, and K. Yaalaoui for their helpful contribution to database collection. The authors are indebted to editor Thomas Mote and to two anonymous reviewers whose comments helped considerably improve the quality of the manuscript.

Tables

Table 1.1 – Description of the selected stations with long records of precipitation.

Station	Country	Lat	Lon	Alt (m)	Record length (yr)	Starting year	Ending year
Algiers	Algeria	36.748	3.068	140	47	1951	2005
Larache	Morocco	35.188	26.158	5	51	1942	2011
Tangier	Morocco	35.778	25.808	5	33	1972	2006
Melilla	Morocco	35.298	22.948	47	46	1907	2009
Gabes	Tunisia	33.888	10.108	4	57	1950	2009
Tunis	Tunisia	36.838	10.238	66	58	1950	2009

Table 1.2 – The significant covariates at 5% and 10% significance levels for each station

Station	Predictors significant at 5% (case)	Predictors significant at 10% (case)	Station	Predictors significant at 5% (case)	Predictors significant at 10% (case)
Algiers	geopot_500 (case4) Slp_surf (case4)	pwater_surf (case1) geopot_500 (case4) Slp_surf (case4)	Mellila	Omega-surf (case1) omega_500 (case2) omega_850 (case2) rhum_surf (case2) ptemp_surf (case3) slp_surf (case3)	rhum_surf (case4) Omega-surf (case1) omega_500 (case2) omega_850 (case2) rhum_surf (case2) ptemp_surf (case3) slp_surf (case3)
Gabes	geopot_500 (case1) geopot_850 (case1,2,4) pwater_surf (case3) Slp_surf (case1,4) shum_500 (case3) shum_850 (case1)	geopot_850 (case2) rhum_surf (case4) geopot_500 (case1) geopot_850 (case1,2,4) pwater_surf (case3) Slp_surf (case1,4) shum_500 (case3) shum_850 (case1)	Tangier	pwater_surf (case1,2) shum_850 (case1) omega_850 (case2) uwind_surf (case2) ptemp_surf (case3) shum_850 (case3) vwind_surf (case3) geopot_500 (case4)	geopot_850 (case1) temp_surf (case4) uwind_surf (case4) pwater_surf (case1,2) shum_850 (case1) omega_850 (case2) uwind_surf (case2) ptemp_surf (case3) shum_850 (case3) vwind_surf (case3) geopot_500 (case4)
Larache	temp_surf (case2) Omega-500 (case2) shum_850 (case2) geopot_500 (case3) ptemp_surf (case3) pwater_surf (case3) rhum_surf (case3)	pwater_surf(case4) temp_surf (case2) Omega-500 (case2) shum_850 (case2) geopot_500 (case3) ptemp_surf (case3)	Tunis	pwater_surf (case2) rhum-surf (case2,3) shum_500 (case2) shum_850 (case2) omega_500 (case3) slp_surf (case3) omega_850 (case4) omega_surf (case4) ptemp_surf (case4)	rhum-surf (case1) shum_850 (case1) geopot_850 (case2) pwater_surf (case2) rhum-surf (case2,3) shum_500 (case2) shum_850 (case2) omega_500 (case3) slp_surf (case3) omega_850 (case4) omega_surf (case4) ptemp_surf (case4)

Table 1.3 – The results of the deviance for PCA

Station	Significant component	Case
Algiers	F1 at 5%	4
Larache	F1 at 5%	1
Tangier	F1 at 5%	1
	F2 at 5%	1
Mellila	F2 at 5%	2
	F1 at 10%	3
	F2 at 5 %	4
Tunis	F1 at 10%	3
	F2 at 5%	3
Gabes	F1 at 10%	1
	F2 at 5%	1
	F1 at 5%	2
	F2 at 5%	2
	F2 at 5%	4

Figures

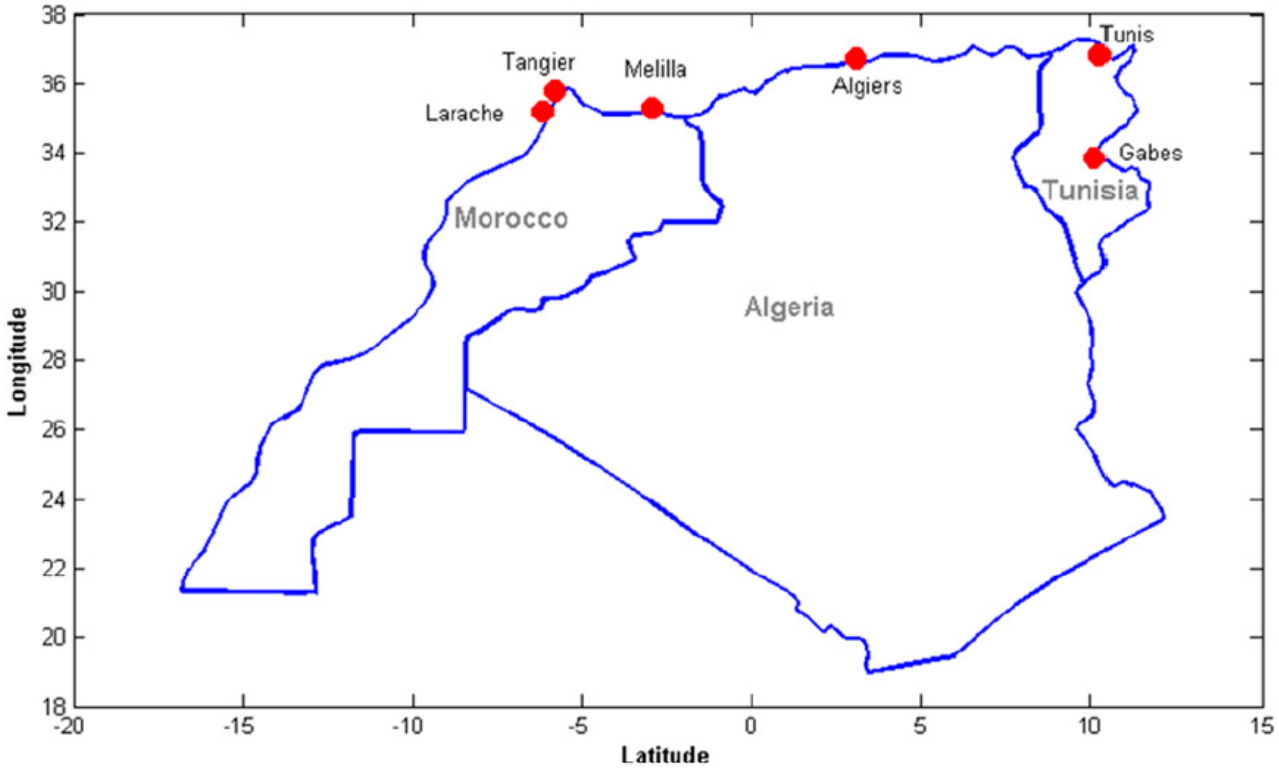

Figure 1.1 – Geographic location of all stations (three selected stations in Morocco, one in Algeria, and two in Tunisia).

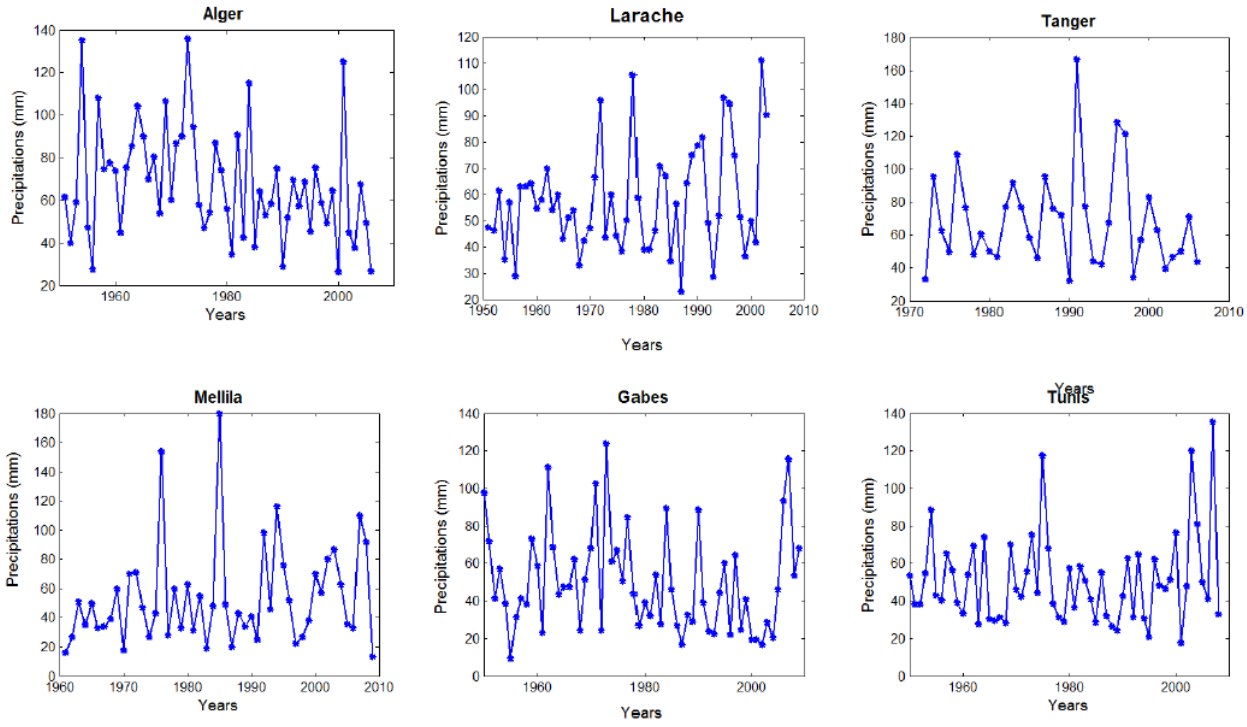

Figure 1.2 – Variation of all MAP series vs time for selected stations.

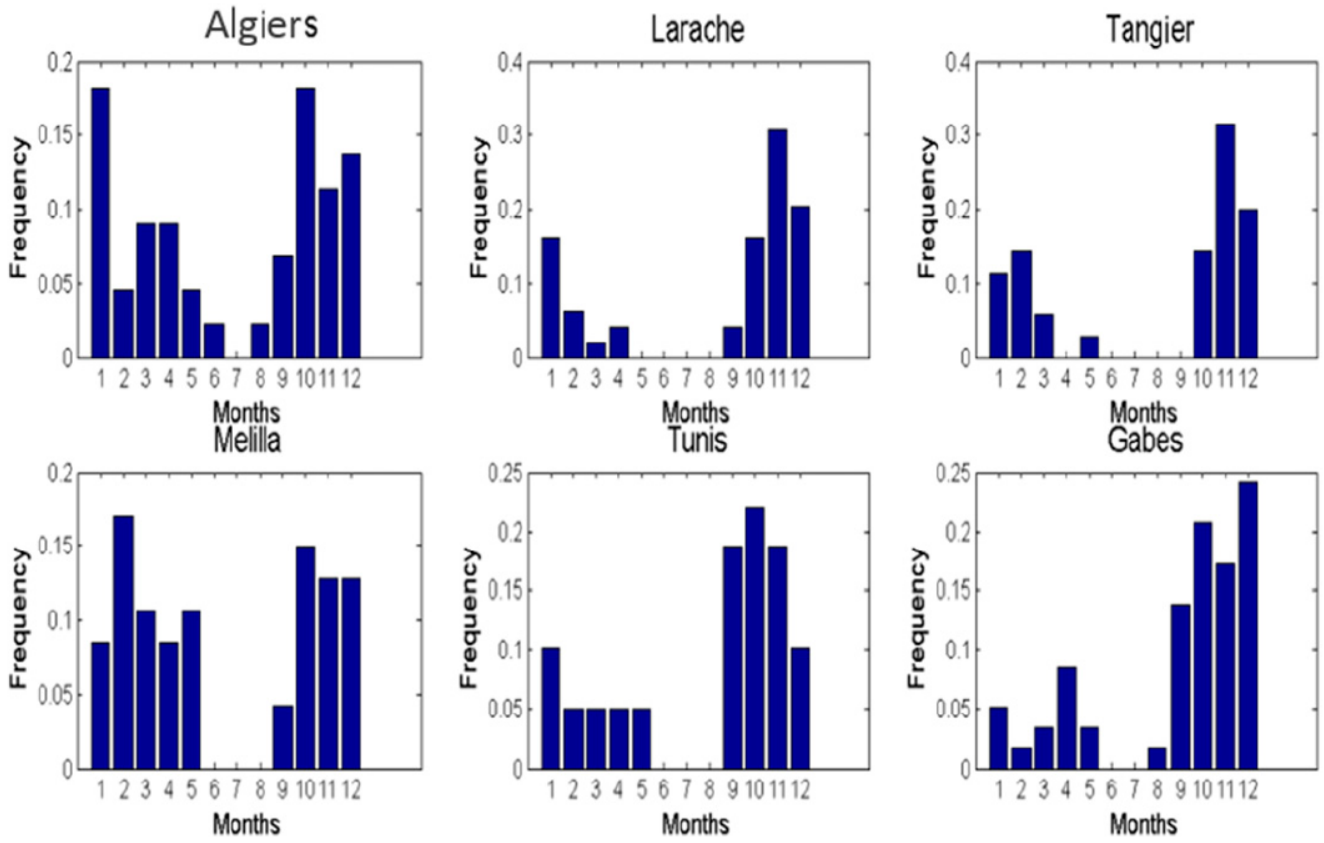

Figure 1.3 – Monthly frequencies of occurrence for daily MAP in each selected station.

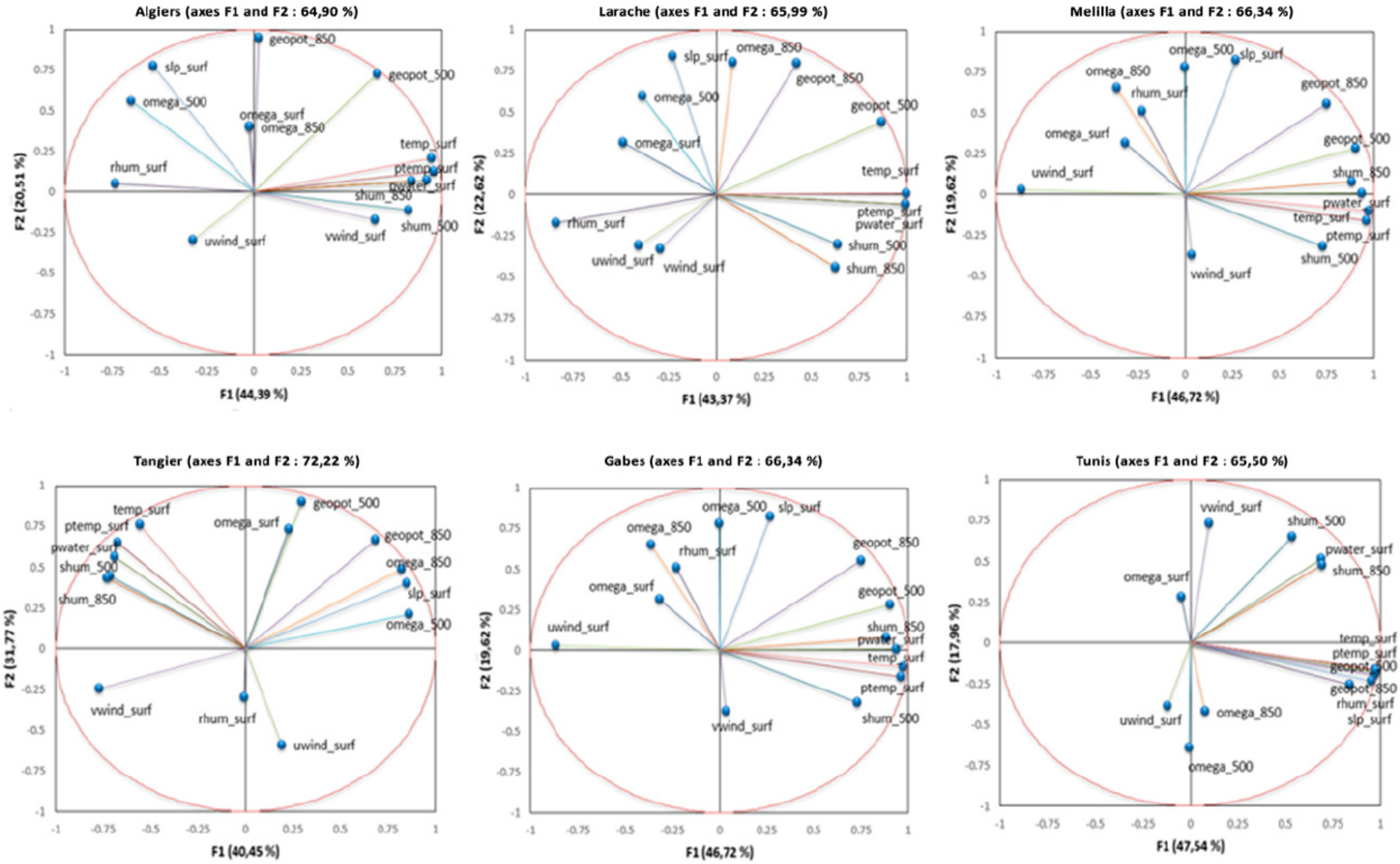

Figure 1.4 – Contributions of the 14 NCEP–NCAR reanalysis covariates on the two principal components (F1 and F2) in selected station (results for case 1). The numbers in the parentheses represent the percentage of explained variance for the represented axes (F1 and F2).

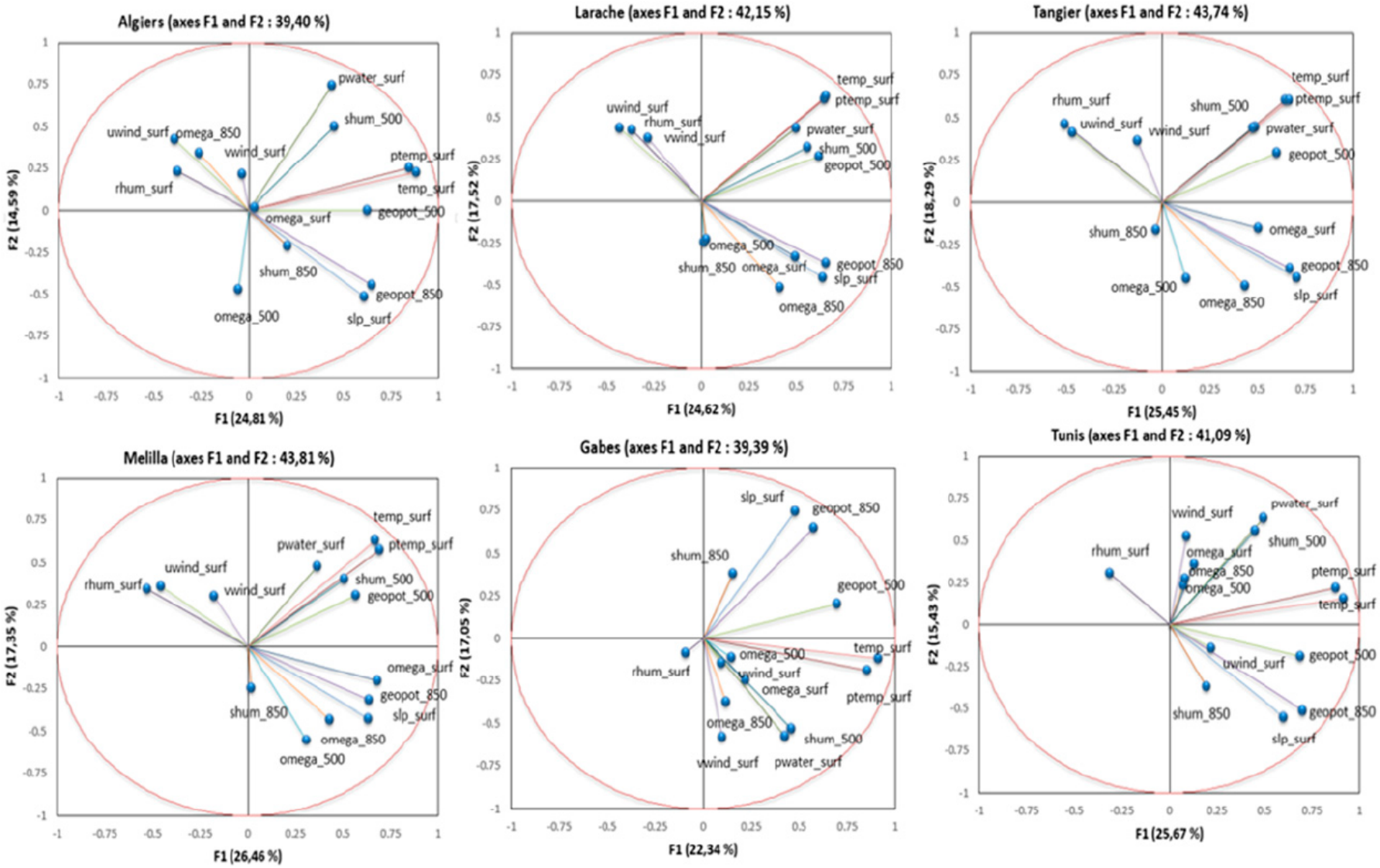

Figure 1.5 – Contributions of the 14 NCEP–NCAR reanalysis covariates on the two principal components (F1 and F2) in selected station (results for case 4). The numbers in the parentheses represent the percentage of explained variance for the represented axes (F1 and F2)

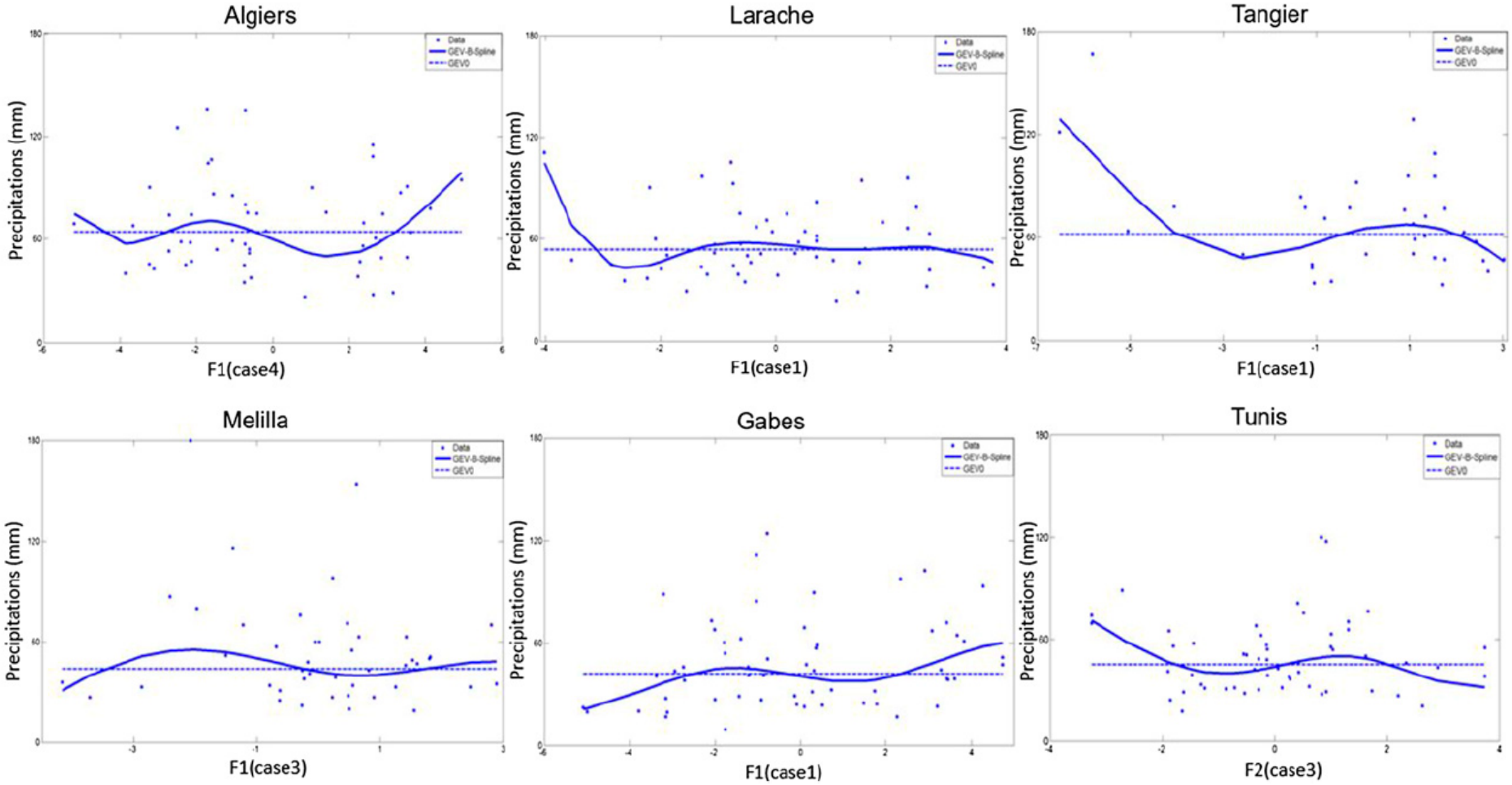

Figure 1.6 – Solid curve and the dotted curve represent an example of nonstationary and stationary median for each station using the first or the second principal component analysis as covariates.

Chapitre 2

Non-Stationary Hydrologic Frequency Analysis using B-Spline Quantile Regression

Titre traduit

Analyse fréquentielle non stationnaire en utilisant la régression des quantiles basée sur les fonctions B-Splines

Auteurs

Bouchra Nasri[1], Taoufik Bouezmarni[2], André St-Hilaire [1] et Taha B.M.J Ouarda [1,3]

[1] Institut national de recherche scientifique, Eau-Terre-Environnement, 490 rue de la couronne, Quebec,G1K 9A9.

[2] Département de Mathématiques, Université de Sherbrooke, 2500 boul. de l'Université, Sherbrooke, J1K 2R1, Canada.

[3] Masdar Institute of Science and Technology, P.O.Box 54224, Abu Dhabi, UAE.

Contribution des auteurs

Bouchra Nasri: proposition du projet, rédaction de la méthodologie, élaboration des codes, proposition des cas d'étude et rédaction de l'article.

Taoufik Bouezmarni: révision de la méthodologie, révision des codes, révision des résultats et correction de l'article.

André St-Hilaire: révision des études de cas, révision des résultats, lecture et correction de l'article.

Taha B.M.J. Ouarda: lecture et révision de l'article.

Remerciements

Je tiens à remercier mes co-auteurs pour la collaboration et la réussite de ce travail.

Publication ciblée

Journal: Journal of Hydrology

Date de soumission: Décembre 2015, resoumis en août 2016

Résumé

L'analyse fréquentielle classique (ou stationnaire) des extrêmes est couramment utilisée par les ingénieurs et les hydrologues afin de fournir des informations de base sur la planification, la conception et la gestion des systèmes de ressources hydriques. Cependant, avec la présence des changements climatiques, il est possible que l'hypothèse de stationnarité ne soit pas vérifiée et par conséquent les résultats de l'analyse fréquentielle stationnaire classique deviendraient discutables. Dans cette étude, nous considérons un cadre d'analyse fréquentielle des extrêmes sur la base de la régression des quantiles avec des fonctions B-Splines. Cette approche permet de fournir un cadre très souple pour évaluer les effets de la non-stationnarité en présence des covariables. Un algorithme de Monte Carlo basé sur les chaines de Markov est utilisé afin d'estimer les quantiles et leurs distributions. Deux critères sont employés, le coefficient de détermination et le critère d'information bayésien, afin de sélectionner le meilleur modèle de B-Splines. Cette méthode est appliquée pour estimer les

risques de crue et d'étiage en Ontario, en fonction de certains indicateurs climatiques. Les résultats montrent une grande différence entre les quantiles de débit stationnaires et leurs équivalents non stationnaires. En effet, cette différence de débit peut dépasser dans certaines stations 92 m^3/s pour des quantiles médians et elle devient encore plus large pour des quantiles supérieurs.

Abstract

Hydrologic frequency analysis is commonly used by engineers and hydrologists to provide the basic information on planning, design and management of hydraulic and water resources systems under the assumption of stationarity. However, with increasing evidence of climate change, it is possible that the assumption of stationarity, which is prerequisite for traditional frequency analysis and hence, the results of conventional analysis would become questionable. In this study, we consider a framework for frequency analysis of extremes based on B-Spline quantile regression which allows to model data in the presence of non-stationarity and/ or dependence on covariates with linear and non-linear dependence. A Markov Chain Monte Carlo (MCMC) algorithm was used to estimate quantiles and their posterior distributions. A coefficient of determination and Bayesian information criterion (BIC) for quantile regression are used in order to select the best model, i.e. for each quantile, we choose the degree and number of knots of the adequate B-spline quantile regression model. The method is applied to annual maximum and minimum streamflow records in Ontario, Canada. Climate indices are considered to describe the non-stationarity in the variable of interest and to estimate the quantiles in this case. The results show large differences between the non-stationary quantiles and their stationary equivalents for an annual maximum and minimum discharge with high annual non-exceedance probabilities. Indeed, this difference can exceed in some stations 92 m^3/s for median quantiles and it becomes larger for higher quantiles.

Keyword

Quantile regression, B-Splines, Bayesian, Streamflow, AMO, PDO.

2.1 Introduction

Understanding the temporal variability of hydrological processes and their associated statistics is essential for better water resource management. Frequency analysis of extreme hydrologic data has been widely used for problems related to engineering design, flood risk management, river navigation planning and water quality management. Generally, current methods of hydrological frequency analysis have been most often based on the assumption of stationarity. Indeed, classical frequency analysis is based on the assumption of underlying independent and identically distributed (i.i.d.) random variables. The last assumption is not valid in non-stationary circumstances. In the context of hydrological processes, non-stationarity is often present because of seasonal effects, perhaps due to different climate patterns in different months, or in the form of trends, possibly due to long-term climate changes [e.g., Stocker *et al.*, 2013; Bates *et al.*, 2008]. Basically, strict-sense stationarity means that the distribution remains constant over time. From a practical point of view, hydrologists assume second-order stationarity, which implies that the first two moments (mean and variance) [1] do not vary over time [Meylan *et al.*, 2012]. Several tests are used to detect non-stationarity in time series including the KPSS test [Kwiatkowski *et al.*, 1992], the Leybourne-McCabe test [Leybourne & McCabe, 1994] and the Mann Kendall test [Mann, 1945]. The last one is the most commonly used in hydro-climatological studies [e.g., Déry & Wood, 2005; Cunderlik & Burn, 2002; Cunderlik & Ouarda, 2009; Nasri *et al.*, 2013; Fiala *et al.*, 2010; Khaliq *et al.*, 2009]. The frequency analysis of a non-stationary series calls for a different understanding than the conventional approach involving stationarity. In fact, in the context of climate change, the distribution parameters and the distribution of hydrological extremes are likely to be modified. As a consequence, the exceedance probability used to estimate the return period also varies over time. Recently, several methods were proposed to take into account, at least partially, non-stationarity in the context of frequency analysis. The

1. the second-order stationarity don't requires a condition on the covariance as in time series data.

most popular approach is the frequency analysis with covariates method. The idea underlying the covariates approach is to incorporate the covariates into the distribution parameters [e.g., Coles, 2001; Olsen *et al.*, 1999; Vrac & Naveau, 2007; Aissaoui-Fqayeh *et al.*, 2009; Cannon, 2010; Ouarda & Adlouni, 2011; El Adlouni & Ouarda, 2009]. Two distributions are generally used in this case: The Generalized Extreme Value (GEV) [e.g., Fisher & Tippett, 1928; Jenkinson, 1955; Hundecha *et al.*, 2008; El Adlouni *et al.*, 2007] and the Generalized Pareto Distribution (GPD) for a Peaks Over a Threshold (POT) approach [Pickands, 1975; Ehsanzadeh *et al.*, 2007]. In the case of stationary data, these distributions are based on limit results of extreme value theory (TVE)[e.g., Fisher & Tippett, 1928; Pickands, 1975]. The use of GEV and GPD for non-stationary data proposed by [Coles, 2001] are not based on the extreme value theory. In fact, it is a "natural" extension of the two models (GEV and GPD) where the stationarity hypothesis is not satisfied. Introducing covariates in one of these distributions can be done through any parameter. The effect of a covariate can be modeled by making one or more parameter linearly [e.g., Coles, 2001; Cannon, 2010] or non-linearly [e.g., Chavez-Demoulin & Davison, 2005; Nasri *et al.*, 2013; Neville *et al.*, 2011] dependent on the covariate. The covariate method is largely developed and has been used in the literature to understand the variation in hydrological time series. Climate indices are commonly used as covariates. This approach works well in the case of linear or quadratic dependence and in the case of one covariate. However, it suffers from several disadvantages in the case of several covariates: (i) the introduction of several covariates in this model increases the number of hyper-parameters, which decreases the model parsimony and potentially increases estimation errors, also can provide, in some cases, some convergence problems (ii) the interactions between the predictors make the model much more complicated because it requires multivariate function modelling. For this reason, some recent studies [e.g., Nasri *et al.*, 2013, 2016] suggest to use the quantile regression method, which was introduced by [Koenker & Bassett, 1987]. Quantile regression provides the conditional

quantiles of the response variable for a fixed value of covariates rather than only the conditional mean. This model can be a good alternative to overcome the problems of convergence raised by the covariate method. Some recent studies in the hydroclimatology context have used linear quantile regression to estimate non-stationary extreme events [e.g., Cannon, 2011; Tareghian & Rasmussen, 2013]. The linear quantile model assumes that the relationship between the variable of interest and covariates is linear. However, in hydroclimatology, the dependence between covariates and variables of interest can take different structures. For this reason, we should investigate the use of a quantile regression model with a more general form of dependency. The nonparametric quantile model allows the assumption of linearity to be relaxed. This model aims to identify the best function according to the data distribution, rather than imposing a restrictive parametric model. Several nonparametric quantile methods have been proposed in the literature [e.g., Koenker *et al.*, 1994; Hendricks & Koenker, 1992]. The most popular is a smoothing regression (or splines regression). Briefly, splines regressions are obtained by joining smoothed polynomial functions separated by a sequence of knots. A larger number of knots leads to a more flexible curve and hence, a better fit. Splines regression for quantile smoothing have been introduced by [Hendricks & Koenker, 1992]. Several variants have been suggested by [Koenker *et al.*, 1994] who proposed to use a natural polynomial splines. This nonparametric quantile regression method is also used in hydrology (see Donner *et al.* [2012], for natural spline quantile regression model). Recently [Nasri *et al.*, 2013] proposed to use B-splines to model nonlinear dependencies. Indeed, B-spline functions are linear combinations of non-negative piecewise polynomials (real functions). This type of functions has some advantages: B-splines do not depend on the response variable, or the variable of interested, but depend only on: (1) the support of the covariates, (2) the number and position of knots and (3) the degree of the B-spline function [De Boor, 2001].

The objective of the present study is to use B-Spline quantile regression for modelling non-stationary

hydrological extreme events (floods and drought) for some rivers located in the province of Ontario (Canada). This province has undergone a number of extreme events over the past two decades. For instance, in 2001: the aggregate level of the Great Lakes plunged to its lowest value in more than 30 years, with lakes Superior and Huron displaying near record lows [Mitchell, 2002; Ashkar & Ouarda, 1996] and in 2005: heavy rainfall and associated flooding resulted in 500 million $ CAD in insured damages [Sandink, 2013].

Hydrologic extreme events are often linked to atmospheric circulation patterns. In fact, several recent studies in North America have modelled the non-stationarity of precipitation and streamflow using climate indices. The most used climate indices in this context are El Nino Southern Oscillation (ENSO) [e.g., Regonda *et al.*, 2005; Cannon, 2010; Nasri *et al.*, 2013], Pacific Decadal Oscillation (PDO) [e.g., Brabets & Walvoord, 2009; Khaliq & Gachon, 2010; Cannon, 2010; Nasri *et al.*, 2013], Atlantic Multi-decadal Oscillation (AMO) [Teegavarapu *et al.*, 1969] and North Atlantic Oscillation (NAO) [Hurrell & Van Loon, 1997]. To our knowledge, no studies have previously studied streamflow extremes using nonparametric quantile regression incorporating B-Spline functions. In the next section, the theoretical background of the nonparametric quantile regression model and its estimation are provided. Data are then presented in Section 3 and results of the B-Splines quantile regression estimation are given in Section 4. Section 5 provides a discussion and a conclusion.

2.2 Theoretical background

2.2.1 Linear quantile regression model

Linear quantile regression is related to linear least-squares regression in that both are used to study the linear relationship between a response variable and one or more independent or expla-

natory variables. However, whereas the least-squares regression is concerned with modelling the conditional mean of the response variable, quantile regression provides a model of the conditional pth quantile of the response variable, for some value of $p \in]0,1[$. For example, the conditional median corresponds to $p = 0.5$. For a vector $y = (y_1, \ldots, y_n)$, the sample mean $\widehat{y}$, solves the least squares problem:

$$\arg\min_{\mu \in \mathbb{R}} \sum_{i=1}^{n} (y_i - \mu)^2 .$$

In many situations, the conditional mean of response variable Y depends on some covariates $\boldsymbol{X} = (X_1, \ldots, X_d)$. For example, the interest variable Y can be the maximum annual precipitation or discharge and $\boldsymbol{X}$ can be a climate index. Based on a sample $(y_i, \boldsymbol{x_i}), i = 1, \ldots, n$, where $\boldsymbol{x_i} = (x_{1i}, \ldots, x_{di})$, for the linear regression model, we suppose that $y_i = \alpha_0 + \boldsymbol{x_i'}\boldsymbol{\alpha} + \varepsilon_i$, where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_d)$ and ε_i are i.i.d with $E(\varepsilon_i|X) = 0$ and $var(\varepsilon_i|X) = c$. An estimate of α_0 and $\boldsymbol{\alpha}$ is obtained by minimizing the following quantity:

$$\sum_{i=1}^{n} (y_i - \boldsymbol{x_i'}\boldsymbol{\alpha} - \alpha_0)^2 .$$

For quantile regression, we suppose that $y_i = \boldsymbol{x_i'}\boldsymbol{\alpha} + \varepsilon_i$, where the pth quantile of $\varepsilon_i \mid \boldsymbol{x_i}$ is defined as: $Q_p(\varepsilon_i \mid \boldsymbol{x_i}) \equiv \inf\{\varepsilon | x_i : F(\varepsilon|x_i) \geq p\} = 0$.

Quantile regression can be derived in a similar manner as mean regression by specifying the pth conditional quantile as $Q_p(Y|X) = \boldsymbol{X}\boldsymbol{\alpha}(\boldsymbol{p}) + \alpha_0(p)$ and estimating $\boldsymbol{\alpha}(\boldsymbol{p})$ and $\alpha_0(p)$ by minimizing

$$\sum_{i=1}^{n} \rho_p (y_i - \boldsymbol{x_i'}\boldsymbol{\alpha} - \alpha_0) \tag{2.1}$$

where $\rho_p(z)$ is a loss function defined as:

$$\rho_p(z) = \begin{cases} z(p-1) \;\; if \;\; z < 0 \\ zp \;\; \text{otherwise.} \end{cases} \tag{2.2}$$

In the case of the linear ordinary mean regression, the loss function can be written as $\rho(z) = z^2$. Figure 2.1 gives a simulated example that shows the difference between the linear mean regression and the linear quantile regression with their loss functions.

2.2.2 Nonparametric quantile regression with B-Spline functions

Nonparametric regression allows the assumption of linearity to be relaxed [Fox, 2000], and it restricts the analysis to smooth and continuous functions. The aim of the nonparametric regression is to identify the best regression function according to the data distribution, rather than estimating the parameters of a specific model. Let us consider the simplest regression case of one explanatory variable:

$$y = f(x) + \varepsilon. \tag{2.3}$$

In nonparametric quantile regression, the function f is not specified and it is commonly assumed that the errors are independent and identically distributed with pth quantile equal to zero. Also, we assume that the errors are independent of the covariate. In the framework of nonparametric quantile regression, several methods are proposed in the literature [Koenker, 2005]. In this work, the B-spline quantile regression is proposed. B-splines, in this case, will be used to approximate the function f. A B-spline is a piecewise polynomial function of degree k and is defined over a domain $x \in [t_0, t_m]$, where m is an integer. The points where $x = t_j$ for $j = 1 \dots m$ are known as knots. A

B-Spline of degree k is a linear combination of basis B-Splines, $B_{i,k}(x)$, of degree k and is given by:

$$f(x) = \sum_{i=1}^{m} \beta_i B_{i,k}(x), \qquad x \in [t_0, t_m]. \tag{2.4}$$

The β_i are called control points and the integer m is the number of knots. Expressions for the polynomial pieces, $B_{i,k}(x)$, can be derived by means of a recursive formula following the definition of the initial polynomial:

$$B_{i,0}(x) = \begin{cases} 1 \ if \ t_i \leq x < t_{i+1} & \forall \quad i = 0 \ldots (m-1) \\ 0 \quad \text{otherwise} \end{cases}$$

and

$$B_{i,k}(x) = \frac{x - t_i}{t_{i+k} - t_i} B_{i,k-1}(x) + \frac{t_{i+k+1} - x}{t_{i+k+1} - t_{i+1}} B_{i+1,k-1}(x) \qquad \forall \quad i = 0 \ldots m-k-1.$$

In this case, B-Spline quantile regression will be described as follows:

$$y = \sum_{i=1}^{m} \beta_i B_{i,k}(x) + \varepsilon. \tag{2.5}$$

2.2.3 Parameter estimation

The classical approach to estimate the parameter vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_m)$ is to use the simplex methods or the interior point methods described in [Koenker, 2005]. In the literature, other methods for estimating the $\boldsymbol{\beta}$ have been developed such as the Bayesian method of [Yu & Moyeed, 2001]. In this work, a Bayesian framework is used to estimate the vector parameters $\boldsymbol{\beta}$ for the B-Spline quantile regression model. The Bayesian approach provides the full distribution for parameters

estimators based on the likelihood function and a prior distribution.

For the prior density of $\boldsymbol{\beta}$, we consider the multivariate Normal distribution (see Green & Silverman [1994], pp. 51-52, for a discussion about the use of multivariate normal density as prior in this context). In fact, in the multivariate case, we do not have a large choice. In the literature, four multivariate parametric distributions are considered: Student, Normal, Gamma and Lognormal distributions. Multivariate Lognormal and Multivariate Gamma distributions are defined for positive random vectors ($\mathbb{R}_+^m$), while the Normal and the Student distributions are defined for random vectors with real support ($\mathbb{R}^m$). Therefore, it seems natural to consider Normal or student distributions. In the present study, we take the multivariate normal distribution. For more flexible distribution the copula functions can provide a good alternative[Nelsen, 2006]. Our prior probability density distribution for $\boldsymbol{\beta}$ in this study is therefore defined by means of the multivariate normal density:

$$\pi_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\boldsymbol{\beta}) = \frac{1}{(2\pi)^{m/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{\beta}-\boldsymbol{\mu})\right\} \tag{2.6}$$

where $\boldsymbol{\mu}$ is the mean of $\boldsymbol{\beta}$, Σ is the variance-covariance matrix of $\boldsymbol{\beta}$ and m is the number of parameters.

The final step in our Bayesian approach is to define the likelihood of (x_i, y_i). The proposed approach is in accordance with [Yu & Moyeed, 2001] and [Thompson *et al.*, 2010]. In these papers, the authors show that the minimization of the loss function is exactly equivalent to the maximization of a likelihood function formed by combining independently distributed asymmetric Laplace densities.

Let us recall the properties of the asymmetric Laplace distribution. A random variable U is said to follow the asymmetric Laplace distribution if its probability density is given by:

$$L_p(u) = p(p-1)\exp\{-\rho_p(u)\}; \quad -\infty < u < +\infty \quad \text{and} \quad p \in]0,1[.$$

Substituting u by $y - \sum_{i=1}^{m} \boldsymbol{\beta}_i B_{ik}(x)$ The resulting likelihood takes the form:

$$L\left(\mathbf{y}|\boldsymbol{\beta}\right) = p^n(1-p)^n \exp\left\{-\sum_{j=1}^{n} \rho_p\left(y_j - \sum_{i=1}^{m} \boldsymbol{\beta}_i B_{ik}(x_j)\right)\right\} \tag{2.7}$$

where ρ_p is the standard quantile regression loss function defined in (2.2). For more information regarding the Laplace distribution, please see 0.3.

Combining $\pi_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\boldsymbol{\beta})$ and $L\left(\mathbf{y}|\boldsymbol{\beta}\right)$, we can write the posterior density function of $\boldsymbol{\beta}$ as:

$$\pi\left(\boldsymbol{\beta}|\mathbf{y}\right) \propto L\left(\mathbf{y}|\boldsymbol{\beta}\right)\pi\left(\boldsymbol{\beta}\right). \tag{2.8}$$

Now, we simulate a realizations of $\boldsymbol{\beta}$ for the posterior density using a Monte Carlo Markov Chain (MCMC) approach implemented through the Metropolis-Hastings (M-H) algorithm. Our inferences are based on these posterior realizations. In particular, we use the posterior mean of $\boldsymbol{\beta}$ to produce our estimated quantile regression. Our algorithm can be summarized as follows:

M-H algorithm

Initialize ($\boldsymbol{\beta^0} \sim \Phi\left(\boldsymbol{\beta}\right)$) ($\Phi$ is called the proposal distribution)

for iteration $i = 1, 2, ...$ **do**

Propose: $\boldsymbol{\beta^*} \sim \Phi\left(\boldsymbol{\beta^i}|\boldsymbol{\beta^{i-1}}\right)$ ($\boldsymbol{\beta^*}$ is called a candidate)

Acceptance probability $\alpha\left(\boldsymbol{\beta^*}|\boldsymbol{\beta^{i-1}}\right) = \min\left\{1, \frac{\Phi\left(\boldsymbol{\beta^{i-1}}|\boldsymbol{\beta^*}\right)\pi\left(\boldsymbol{\beta^*}\right)}{\Phi\left(\boldsymbol{\beta^*}|\boldsymbol{\beta^{i-1}}\right)\pi\left(\boldsymbol{\beta^{i-1}}\right)}\right\}$

$u \sim$ Uniform $(u; 0, 1)$

if $u < \alpha$

Accept the proposal $\boldsymbol{\beta^i} \leftarrow \boldsymbol{\beta^*}$

else

Reject the proposal $\boldsymbol{\beta^i} \leftarrow \boldsymbol{\beta^{i-1}}$

end if

end for

The first step is to initialize the sample value for parameter vector $\boldsymbol{\beta}$ (this value is often sampled

from the parameter's prior distribution). The main loop of the M-H algorithm consists of three components: (1) Generate a proposal sample $\boldsymbol{\beta^*}$ from the proposal distribution $\Phi\left(\boldsymbol{\beta^i}|\boldsymbol{\beta^{i-1}}\right)$; (2) Compute the acceptance probability via the acceptance function based upon the proposal distribution and the full joint density $\pi(.)$; (3) Accept the candidate sample with probability α, the acceptance probability, or reject it with probability $1-\alpha$.

Proposal Distribution: The M-H algorithm starts with simulating a "candidate" sample $\boldsymbol{\beta^*}$ from the proposal distribution $\Phi(.)$. Note that samples from the proposal distribution are not accepted automatically as posterior samples. These candidate samples are accepted probabilistically based on the acceptance probability α. In the literature, the proposal distribution is often the same as the prior distribution [e.g., Gelman *et al.*, 1995; Gilks *et al.*, 1996]. In our study, we choose the multivariate normal distribution as a proposal distribution function. In this case, the acceptance probability $\alpha\left(\boldsymbol{\beta^*}|\boldsymbol{\beta^{i-1}}\right) = \min\left\{1, \frac{\Phi\left(\boldsymbol{\beta^{i-1}}|\boldsymbol{\beta^*}\right)\pi(\boldsymbol{\beta^*})}{\Phi\left(\boldsymbol{\beta^*}|\boldsymbol{\beta^{i-1}}\right)\pi\left(\boldsymbol{\beta^{i-1}}\right)}\right\}$ becomes $\alpha\left(\boldsymbol{\beta^*}|\boldsymbol{\beta^{i-1}}\right) = \min\left\{1, \frac{\pi(\boldsymbol{\beta^*})}{\pi\left(\boldsymbol{\beta^{i-1}}\right)}\right\}$ because of the symmetry of the normal distribution.

2.2.4 Criteria to choose the best model

The B-spline functions depend on two parameters: the number of knots (m) and the degree (k). When $(m,k) = (1,1)$, the B-spline quantile regression model is exactly the linear quantile regression model, and if $(m,k) = (1,2)$, the B-Spline quantile regression model becomes the quadratic quantile regression model. To select the « Best » model, two performance methods are used: (i) the « coefficient of determination » based on the quantile and (ii) the Bayesian information criterion (BIC) for quantile regression.

Coefficient of determination for quantiles

The coefficient of determination for quantiles was proposed by [Bentzien & Friederichs, 2007]. The asymptotic properties and simulation performance for the determination coefficient were studied by [Noh *et al.*, 2012]. This coefficient aims to quantify the goodness of fit measures in the framework of quantile regression. This coefficient aims to compare the mean of residual errors between two models. The coefficient of determination is defined as follows:

$$R(p) = 1 - \frac{E\left(\rho_p\left(Y - \widehat{f}(X)\right)\right)}{E\left(\rho_p\left(Y - \widehat{f}_c(X)\right)\right)} \tag{2.9}$$

where, for fixed $0 < p < 1$, $\widehat{f}$ is the nonparametric conditional pth quantile of Y given X and $\widehat{f}_c$ is the conditional pth quantile of Y given X for the comparative parametric model (e.g. linear or quadratic). We note that, the expectation in the numerator and the denominator are estimated by the sample average. $R(p)$ can be positive, negative or null. When it takes positive values, it means that the nonparametric model is better than the comparative model and when it takes negative values, it means that the comparative model is better. Otherwise the two models are equivalent.

Bayesian information criterion for quantiles

The BIC is a criterion for model selection among a finite set of models; the model with the lowest BIC is preferred. It is based, in part, on the likelihood function. When fitting models, it is possible to increase the likelihood by adding parameters, but leads to overfitting estimation. To overcome this problem, the BIC introduces a penalty term depending on the number of parameters in the model. The BIC was developed, in the first time, by [Schwarz, 1978]. The BIC criterion can

be written as follows:

$$BIC = -2\log\left(\widehat{L(\mathbf{y}|\boldsymbol{\beta})}\right) + p\log(n) \tag{2.10}$$

where $\widehat{L(\mathbf{y}|\boldsymbol{\beta})}$ is the estimator of the likelihood function and it is the same function given in (2.7), p is the number of parameters and n denotes the sample size. In quantile regression, several references have studied this criterion [Nishii, 1984; Wu & Zen, 1999; Zhang *et al.*, 2010]. For a review of literature on the usage of BIC, see [Lee *et al.*, 2014] where they give the BIC form of the linear and nonlinear quantile regression. In hydroclimatology works, we can see the application of this criterion in [Donner *et al.*, 2012] as well as [Koenker & Schorfheide, 1994].

2.3 Data

This study is based on two applications. These applications respectively focus on (a) annual maximum and (b) minimum streamflow records in Ontario, using climate indices to model non-stationarity. For each application, we have selected 5 stations. The data from each station were checked, in particular to look for obvious shifts, outlier values and missing data. The stations that were selected have more than 30 years of complete daily data with identified non-stationarity and they are correlated with at least one climate index.

For this study, we tested the dependence between the variables of interest and the following covariates: Pacific Decadal Oscillation (PDO), Atlantic Multi-decadal Oscillation (AMO), North Atlantic Oscillation (NAO) and El Nino Southern Oscillation (ENSO) indices (see Appendix 0.4 for definitions). Ultimately, only the covariates that have significant dependence with the variables of interest were kept, which are AMO and PDO indices. AMO and PDO were found to have, respectively, significant dependence with maximum and minimum streamflow time series. The dependence in this case is estimated by using the Kendall rank correlation coefficient [Kendall, 1948].

Streamflow data come from the HYDAT database of Environment Canada `ftp://arccf10.tor.ec.gc.ca/wsc/software/HYDAT/`. Climate indices data come from NOAA (National oceanic and atmospheric administration Earth System Research Laboratory) website `http://www.esrl.noaa.gov`. Figure 2.2 illustrates the geographic location of all stations selected for each application. Table 2.1 gives a summary of the description of the selected stations with long precipitation records, the results of Kendall's tau rank correlation coefficients and the stationary quantile estimation for 2 and 10 return period, for applications (a) and (b). Figures 2.3 and 2.4 show the variations of maximum and minimum annual streamflows at each station. Figures 2.5 and 2.6 show, respectively, the variation of maximum and minimum streamflows against AMO and PDO oscillation for each station.

2.4 Results

For model development, the following functions are fitted:

— Maximum annual streamflow$_i = f_{1i}(AMO) + \varepsilon$; $i = 1, ..., 5$

— Minimum annual streamflow$_i = f_{2i}(PDO) + \varepsilon$; $i = 1, .., 5$

The two functions f_1 and f_2 are estimated by the B-spline approach. To select the degree k and the number of knots m, we calculate the determination coefficient and BIC criterion for several values of the couple (k, m). For each application in each station, we choose the couple (k, m) that maximises (resp. minimises) the determination coefficient (resp. the BIC critierion). The following paragraph shows the results of the determination coefficient and the best model of each station. The next one describes the results of the models selected by $R(p)$ and BIC criteria.

For application (a), these criteria are calculated for quantiles 0.5 and 0.9, which correspond, respectively, to return periods of 2 years and 10 years and for application (b), these coefficients are

calculated for quantiles 0.1 and 0.5, which correspond to return periods of 10 years and 2 years. Tables 2.2 and 2.3 show the determination coefficient corresponding to different values of the couple (m, k), for application (a) and application (b) respectively. Tables 2.4 and 2.5 show the BIC criterion corresponding to different values of (m, k), for application (a) and application (b) respectively. From Table 2.2, it can be noticed that the best model chosen using the determination coefficient is the model with 3 knots and 3 degrees for all stations. From Tables 2.3, we can notice that the best model is model with 2 knots and 2 degrees for all stations except for station 02HC029, where the model with 3 knots and 3 degrees is selected. Different results are provided using BIC criterion. In fact, for all the studied cases and for both application, we can see that the models selected using the BIC criterion have less degrees and number of knots than those chosen using the coefficient of determination. For both applications, the values of the degrees and knots are less or equal to (2,3). This result was expected, as the BIC criterion penalizes for the number of parameters to estimate and favours parsimony. Whereas, the coefficient of determination only considers estimation errors. An "adjusted" determination coefficient which takes into account the number of parameters to estimate is an interesting problem to study in future works. For this reason, quantiles estimation will be done by using the models selected by the BIC criterion.

Figures 2.7 and 2.8 show the estimated 2 and 10-year return period maximum and minimum streamflow quantiles, respectively as function of the covariates AMO and PDO. It can be seen that generally, quantiles take different values as the covariate values change. Non-stationary quantiles can take much larger values than stationary quantiles. For example, the results for station 02AC001 show that the stationary median is equal to 48 (m^3/s). However, the non-stationary median can reach 140 (m^3/s). This difference between stationary and nonstationary quantiles become very large for high quantile levels. In fact, for station 02AC001, the 0.9 stationary quantile is equal to 96.9 (m^3/s), but for the nonstationary quantile it can reach 250 (m^3/s). The stationary quantile results

are given in Table 2.1. We can see similar results for all stations in both applications which confirms the importance of considering the nonstationary quantiles for better water resource management practices.

2.5 Discussion and conclusion

The two last decades have witnessed the development of a large number of statistical modelling approaches for extreme value variables in the presence of non-stationarity or dependence on covariates. In this study, we present the B-spline quantile model, a nonparametric approach which model linear and nonlinear conditional quantiles or quantile with covariates and offers great flexibility for smoothing the quantile regression. Estimating the parameters of the proposed model is carried out using the Bayesian approach. It combines observed and prior information, estimates the entire posterior distribution of the parameters and quantiles and allows to give better or similar estimation results than the frequentist approach (similar results, in the case of non-informative prior and better if we have a prior information concerning the parameters).

Despite the advantages of the nonparametric model, this kind of model is often criticized in the literature for the possibility of over-parameterization, leading to a parsimony problem. Some studies have suggested to use a classical model such as a linear or quadratic quantile models to avoid this problem. In this study, we propose to adapt a performance criterion for quantile regression that allows the comparison of B-Spline approach with classical model, by using a coefficient of determination and BIC criterion for quantile regressions. However, only the BIC criterion allows to select the best performing model with less parameters.

In this work, two case studies are proposed to show the performance of the proposed method. The first one is focused on maximum annual streamflow quantiles for five stations in Ontario using the

AMO oscillation index, and the second one estimates the quantiles for minimum annual streamflow at five other stations in Ontario, using the PDO oscillation index. Our results show that:

1. Different covariates can influence different metrics of one variable of interest in the same study area. In this case, we have the PDO index that influences the minimum annual streamflow while it does not affect the annual maximum streamflow and conversely for the AMO index. Looking at the time series of PDO and AMO oscillations, we notice that the relationship between the AMO and PDO are negative during the period of 1942-1965 and 1968-1998 [Rowan & Daniel, 2005]; which can explain the influence of AMO for maximum annual discharge and PDO for minimum annual flow values.

2. Moreover, it can be noticed that, although the shape of the relationship between AMO and floods is similar for all five studied stations, it is not the case for low flows and PDO. Two of the five stations, located further north show a negative relationship between low flows and PDO, while for the three stations located in southern Ontario, this relationship is positive. Looking at the daily datasets of flows in these stations, we noticed that minimum discharge values in the northern stations are most often observed late in the winter, generally between March and April. However, the minimum flow values at the other stations are often observed during the summer or autumn period, especially between July and November. This can explain the difference in signs (+ or -) of Kendall's tau.

3. The quantile regression model was used in a framework that includes nonparametric smoothing B-spline functions. These functions can capture linear and nonlinear dependence between covariates (e.g climate indices) and the variables of interest (annual minimum and maximum streamflows). For both case studies, conditional quantiles are calculated for two return periods T = 2 years and 10 years. Quantiles for higher return periods (e.g. T = 50, T = 100 years) could be calculated if we disposed of longer data sets. The proposed model allows to modulate

conditional quantile estimation as a function of low frequency atmospheric patterns and in some cases, this can lead to quantile estimations that are much higher than those obtained in a stationary framework. For nonstationary quantiles, several values are possible for a given return period, depending on the value of the covariate. Also, we see that nonlinear quantile values are greater than stationary quantile values for all stations.

4. The proposed model shows several advantages and some drawbacks. Indeed, according to the results described above, we can easily conclude that it is a relatively complex model for describing the linear and nonlinear quantiles in the presence of covariates. However, it is a flexible model that allows to reach more extreme values than classical models like linear and quadratic quantile regression models. The proposed model also has some disadvantages. Indeed, the optimal number of knots and degrees of smoothing for the B-spline functions are always based on the calculation of a specific criterion (coefficient of determination and BIC criterion in this case), which can take much computing time and some programming skill.

5. In this work, a Bayesian framework is used to estimate $\boldsymbol{\beta}$ parameters for the B-Spline quantile regression model. The Bayesian approach provides the full distribution for estimators of the parameters. Posterior distributions are calculated for each β, in each station, and each probability level. In this work, we have excluded more details about the posterior distribution in order to reduce the size of paper. We give an example of the MCMC results for one station 04JF001 and for one quantile level = 0.5. Please see Figure 2.9.

6. Two criteria are used to choose the best model with more parsimony. The first criterion is the determination coefficient for quantile regression and the second one is the BIC criterion. Both criteria are applied for several combinations of number of knots and degree of smoothing. These criteria gave different results. In fact, the models selected by the BIC criterion contained less coefficients than the models selected by the determination coefficient for quantiles. The

results of the BIC show the advantage of using non-linear functions in this context. Indeed, error estimation decreases with the use of degree greater than 2. This indicates the usefulness of the proposed model.

7. All the BIC results were confirmed by the MCMC confidence interval. In fact, we can notice from Figure 2.9, that if all the β are different than 0, then the MCMC confidence interval does not contain 0. And we can notice from the results of BIC criterion that, the best selected model for 04JF001 station is the model with degree equal to 2 and the number of knots equal to 1. The same results are shown for all the other stations.

8. In this study, we propose an alternative way to estimate conditional quantile (or nonstationnary quantile). This method is based on the estimation of B-Spline of regression model which is easier than the B-Spline model based on the GEV (GPD). The results of this work give the return streamflow quantile for each value of covariate (here, we have a range of values of AMO and PDO). In the future, if new values of AMO and PDO are observed and are inside the interval values of the covariates as in the present study, we can easily predict the streamflow quantiles. However, this type of model allows the description of the impact of covariates on the variable of interest and cannot be used for predictions outside the interval values of covariate. Indeed, outside this interval, we can never guess how the variable of interest varies depending on the covariate.

Here, we have used a single covariate to explain the temporal variations of the maximum and minimum flows. This unique covariate partially explained these variations. The introduction of more covariates may allow for better quantile estimation. Hence, future researches may deal with the introduction of additional covariates. In the context of climate change studies, additional covariates could include GCM/RCM outputs or NCEP / NCAR reanalysis predictors to better explain the temporal variation of the flow in relation to climate.

Tables

Table 2.1 – Description of the selected stations with length of discharge records for application (a) and (b). The five first stations are station chosen for application (a) and the five last stations are the stations chosen for application (b). $Q_{T=2}$ and $Q_{T=10}$ correspond to the stationary quantiles, repectively, for 2 and 10 years return period estimated by using the inverse of cumulative distribution function.

Station	Length of data	Latitude	Longitude	Kendall's tau	$Q_{T=2}$	$Q_{T=10}$
02AC001	1971-2010	48.821	-88.534	AMO (-0.25)	48.9	96.6
02HB012	1965-2010	43.301	-79.869	AMO (-0.24)	11.5	18.1
02HD012	1975-2010	43.991	-78.3282	AMO (-0.26)	28.4	52.5
02LA007	1969-2010	45.249	-75.7906	AMO (-0.27)	79.2	133.0
04LM001	1972-2010	50.585	-82.091	AMO (-0.29)	1880.0	2796.0
02FB007	1959-2010	44.522	-80.930	PDO (0.4)	0.4	0.6
02HC009	1959-2010	43.790	-79.584	PDO (0.39)	0.1	0.2
02HC029	1964-1996	43.757	-79.345	PDO (0.42)	0.4	0.5
04FA001	1970-2010	51.823	-89.602	PDO (-0.35)	15.3	20.8
04JF001	1980-2010	50.658	-86.532	PDO (-0.24)	13.3	16.5

Table 2.2 – Coefficient of Determination for B-spline quantile regression model vs linear quantile model (l) and quadratic quantile model (q) for application (a).

(Degree, Knots)	02AC001		02HB012		02HD012		02LA007		04LM001	
	$p = 0.5$	$p = 0.9$	$p = 0.5$	$p = 0.9$	$p = 0.5$	$p = 0.9$	$p = 0.5$	$p = 0.9$	$p = 0.5$	$p = 0.9$
$(1,1)^l$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$(1,1)^q$	0.04	-0.41	0.28	0.05	0.10	-0.13	-0.22	0.03	-0.37	-0.17
$(1,2)^l$	-0.09	0.17	-0.63	-0.37	0.61	-0.56	0.09	-0.02	0.12	0.07
$(1,2)^q$	-0.09	0.28	-0.63	-0.37	0.61	-0.56	0.26	-0.02	0.34	0.07
$(1,3)^l$	-0.09	-0.06	-0.63	-0.13	0.60	0.02	0.12	-0.02	0.04	0.28
$(1,3)^q$	-0.09	0.27	-0.63	-0.13	0.60	0.02	0.26	-0.02	0.38	0.28
$(2,1)^l$	-0.04	0.29	-0.39	-0.06	-0.11	0.12	0.18	-0.03	0.27	0.15
$(2,1)^q$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$(2,2)^l$	0.20	0.16	-0.46	-0.01	-0.05	-0.31	0.69	0.01	0.38	0.20
$(2,2)^q$	0.20	0.42	-0.46	-0.01	-0.05	-0.31	0.20	0.01	-0.22	0.20
$(2,3)^l$	0.26	0.11	0.10	-0.25	0.60	0.09	0.32	0.06	0.37	0.35
$(2,3)^q$	0.26	0.43	0.10	0.13	0.60	0.19	0.33	0.06	0.40	0.35
$(3,1)^l$	0.09	0.04	-0.62	-0.15	0.59	0.01	0.10	-0.07	0.04	0.28
$(3,1)^q$	0.09	0.33	-0.62	-0.15	0.59	0.01	0.23	-0.07	0.38	0.28
$(3,2)^l$	0.29	0.12	-0.50	0.13	0.46	0.01	0.02	0.12	0.22	0.28
$(3,2)^q$	0.29	0.50	-0.50	0.13	0.46	0.01	0.20	0.12	0.43	0.28
$(3,3)^l$	0.34	0.31	0.17	0.13	0.95	0.20	0.48	0.20	0.38	0.45
$(3,3)^q$	0.34	0.54	0.17	0.13	0.95	0.20	0.50	0.20	0.49	0.45
$(3,4)^l$	0.15	0.13	0.02	0.03	0.60	0.17	0.30	0.02	1.29	0.29
$(3,4)^q$	0.14	0.39	0.00	0.04	0.33	0.13	0.28	0.04	0.27	0.22

Table 2.3 – Coefficient of Determination for B-spline quantile regression model vs linear quantile model (l) and quadratic quantile model (q) for application (b).

(Degree, Knots)	02FB007		02HC009		02HC029		04FA001		04JF001	
	$p=0.5$	$p=0.9$	$p=0.5$	$p=0.9$	$p=0.5$	$p=0.9$	$p=0.5$	$p=0.9$	$p=0.5$	$p=0.9$
$(1,1)^l$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$(1,1)^q$	0.17	0.00	-0.11	-0.07	0.29	-0.47	0.15	0.16	-0.32	0.08
$(1,2)^l$	-0.74	0.06	0.01	0.11	-0.05	-0.08	0.87	-0.11	0.33	0.12
$(1,2)^q$	-0.45	0.06	-0.09	0.05	0.25	-0.59	0.89	0.06	0.11	0.19
$(1,3)^l$	-1.92	0.06	-0.15	0.13	0.63	0.12	0.03	1.12	0.98	1.12
$(1,3)^q$	-1.43	0.06	-0.27	0.07	0.74	-0.30	0.18	1.10	0.97	1.11
$(2,1)^l$	-0.20	0.00	0.10	0.06	-0.41	0.32	-0.18	-0.18	0.24	-0.09
$(2,2)^l$	0.30	0.08	0.27	0.14	0.60	0.12	0.69	0.29	0.79	0.16
$(2,2)^q$	0.42	0.08	0.19	0.08	0.71	0.30	0.74	0.40	0.72	0.23
$(2,3)^l$	0.69	0.14	-0.16	0.01	0.19	0.10	0.62	0.11	0.39	0.11
$(2,3)^q$	0.75	0.14	-0.28	-0.05	0.43	-0.32	0.68	0.25	0.19	0.18
$(3,1)^l$	-0.61	-0.01	0.07	0.07	-0.36	0.32	0.54	-0.26	-0.51	0.02
$(3,1)^q$	-0.34	-0.01	-0.03	0.01	0.04	-0.01	0.61	-0.06	-1.00	0.10
$(3,2)^l$	-1.45	0.17	0.14	0.12	0.54	0.11	0.34	0.11	-0.93	0.11
$(3,3)^l$	0.48	0.20	-0.32	0.11	0.18	0.08	0.32	0.09	0.24	0.08
$(3,3)^q$	0.56	0.20	-0.46	0.05	0.42	-0.35	0.42	0.23	0.00	0.15

Table 2.4 – BIC results for different couple of degree and knots in the B-Spline quantile model for the application (a). The results in the table are $*100$

(Degree, Knots)	02AC001		02HB012		02HD012		02LA007		04LM001	
	$p=0.5$	$p=0.1$	$p=0.5$	$p=0.1$	$p=0.5$	$p=0.1$	$p=0.5$	$p=0.1$	$p=0.5$	$p=0.1$
(1,1)	252.60	223.40	789.60	346.90	195.36	158.74	273.82	164.18	554.46	488.73
(1,2)	253.60	223.50	793.10	358.10	207.53	157.83	279.10	171.40	559.12	483.90
(1,3)	253.70	223.90	914.20	356.30	209.74	157.91	277.40	169.10	557.48	479.52
(2,1)	252.20	223.20	755.60	353.30	197.70	156.29	273.90	167.06	550.63	473.10
(2,2)	240.60	210.20	853.00	277.20	204.69	162.77	280.41	175.13	550.91	476.80
(2,3)	251.60	222.11	733.20	275.80	194.96	148.89	272.20	163.45	555.00	479.84
(3,1)	252.60	223.11	876.50	310.70	205.72	161.44	280.82	173.58	558.98	483.31
(3,2)	258.56	226.85	917.10	317.70	209.62	163.72	285.01	177.77	562.83	487.26
(3,3)	262.68	226.75	957.00	356.10	211.14	161.65	284.19	182.12	565.67	481.60

Table 2.5 – BIC results for different couple of degree and knots in the B-Spline quantile model for the application (b). The results in the table are $*100$

(Degree, Knots)	**02FB007**		**02HC009**		**02HC029**		**04FA001**		**04JF001**	
	$p = 0.5$	$p = 0.1$	$p = 0.5$	$p = 0.1$	$p = 0.5$	$p = 0.1$	$p = 0.5$	$p = 0.1$	$p = 0.5$	$p = 0.1$
((1,1)	-280.53	-386.01	-359.32	-464.85	-326.43	-397.23	-241.20	-541.80	-184.10	-478.50
(1,2)	-268.86	-366.45	-359.06	-452.84	-328.02	-405.39	-502.65	-632.83	-446.78	-555.52
(1,3)	-271.28	-373.30	-349.47	-443.26	-313.80	-383.12	-263.94	-549.12	-278.00	-543.01
(2,1)	-277.18	-377.78	-360.17	-456.85	-323.51	-394.97	-518.20	-652.40	-460.60	-572.70
(2,2)	-291.25	-396.22	-370.28	-466.97	-338.17	-417.93	-272.10	-566.10	-286.60	-559.80
(2,3)	-279.67	-384.85	-347.09	-462.71	-314.13	-387.95	-227.40	-579.10	-337.10	-547.20
(3,1)	-277.18	-374.36	-347.10	-462.91	-310.05	-382.84	-424.10	-516.40	-395.00	-520.90
(3,2)	-269.70	-374.62	-358.66	-462.33	-311.31	-378.48	-423.70	-621.70	-366.60	-489.30
(3,3)	-269.87	-378.03	-354.92	-466.34	-306.19	-374.87	-449.30	-615.30	-365.70	-449.40

Figures

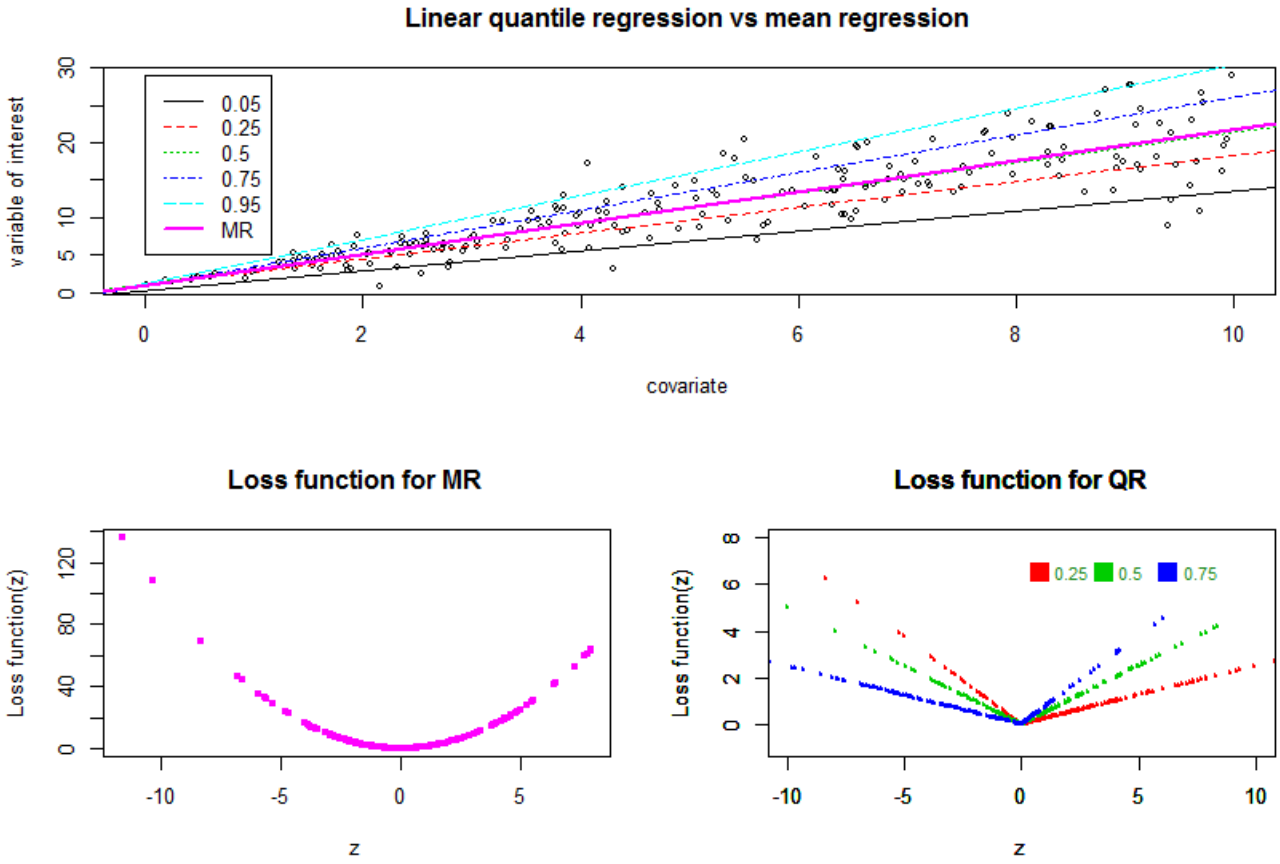

Figure 2.1 – Example of linear mean regression (MR) and linear quantile regression (QR) with their corresponding Loss function. Note that, we use here the Bayesian method for quantile and mean estimation.

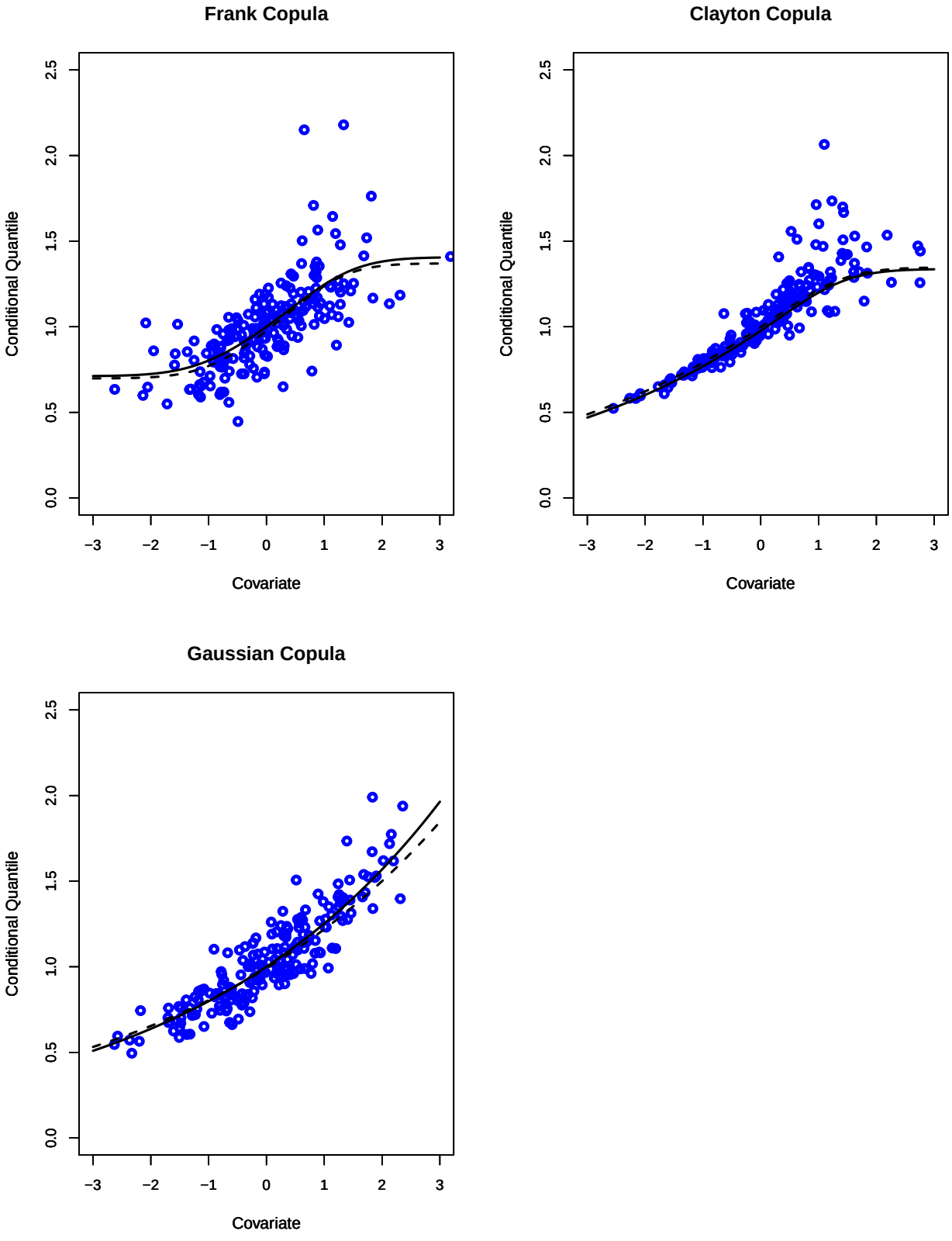

Figure 2.2 – Geographic location of all stations for application (a) and application (b)

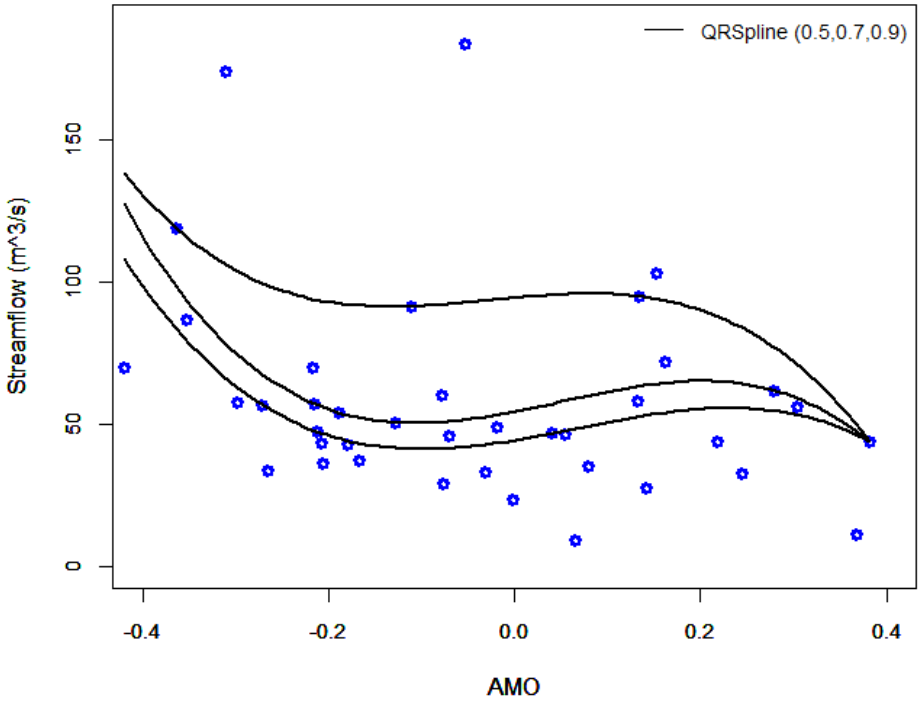

Figure 2.4 – Variation of minimum annual streamflows for each station- Application (b)

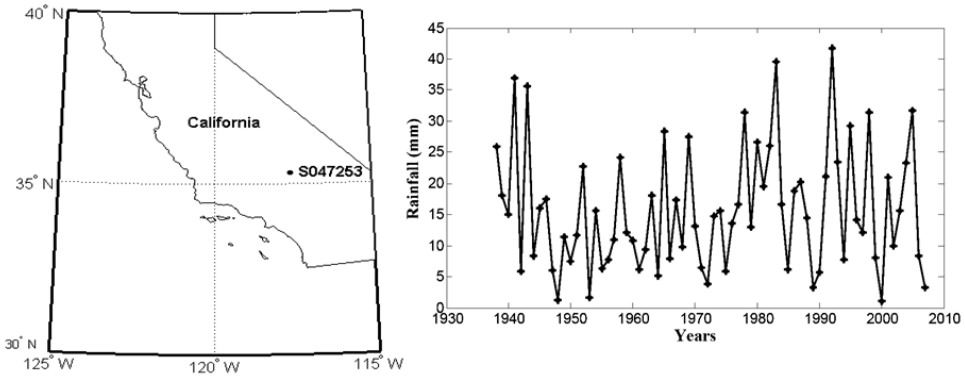

Figure 2.5 – Annual maximum streamflows vs AMO oscillation

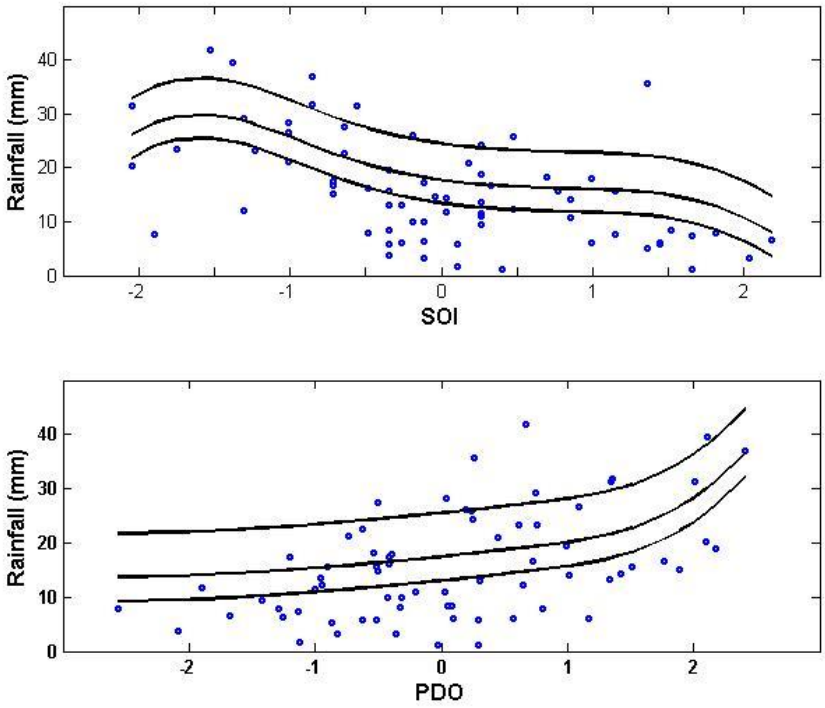

Figure 2.6 – Annual minimum streamflows vs PDO oscillation

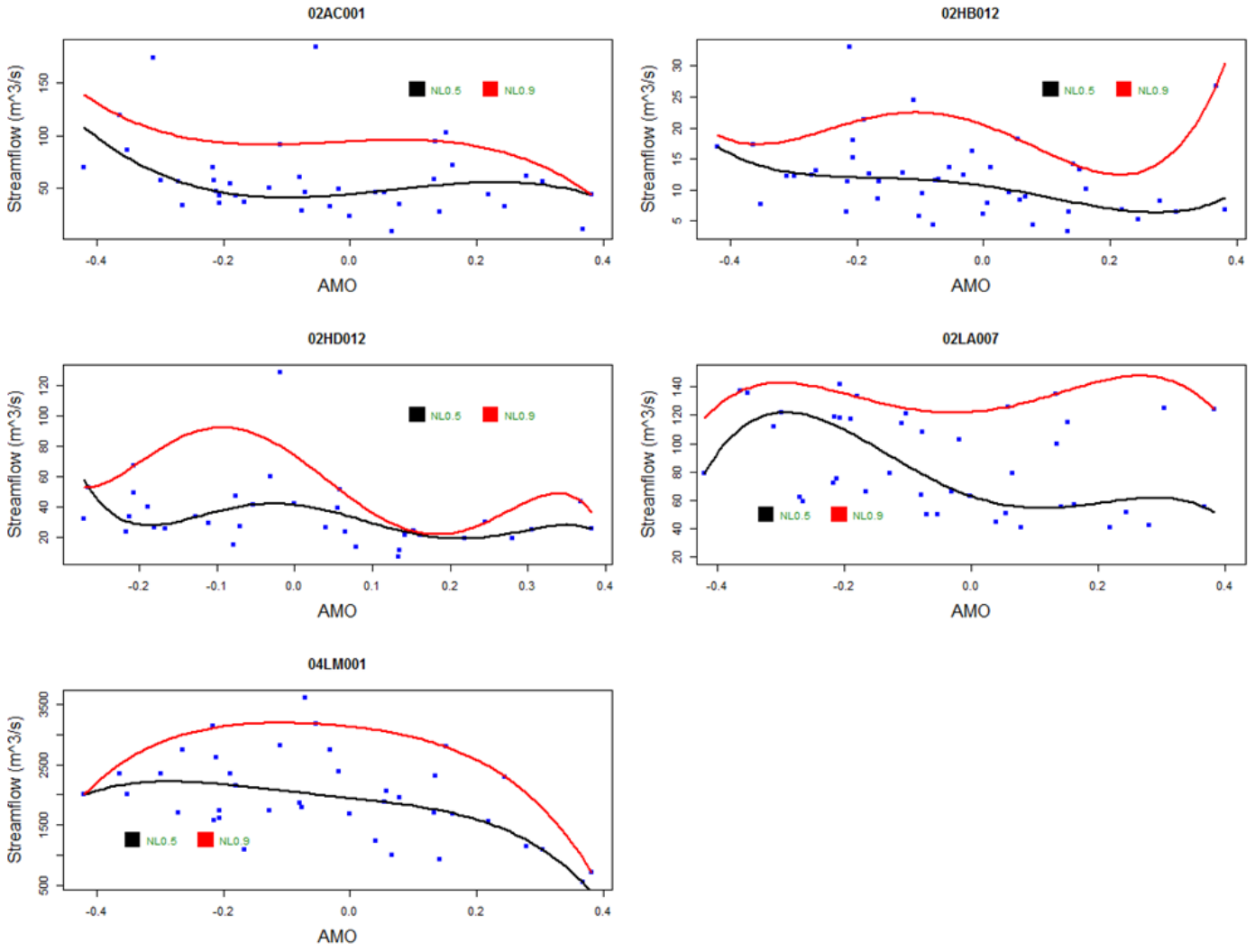

Figure 2.7 – 0.5 and 0.9 quantile results estimated by using the B-spline quantile regression model -Application (a)

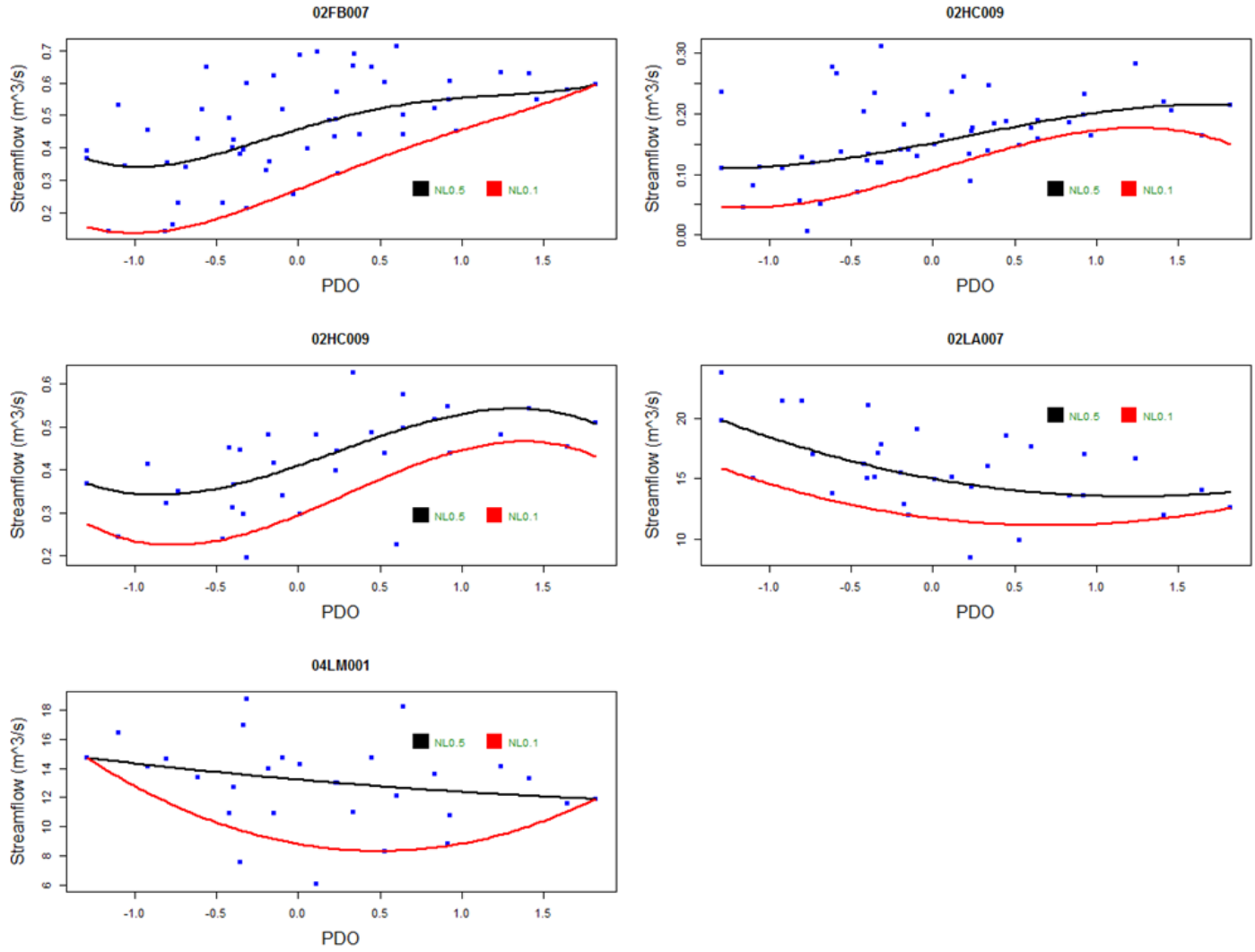

Figure 2.8 – 0.1 and 0.5 quantile results estimated by using the B-spline quantile regression model-Application (b)

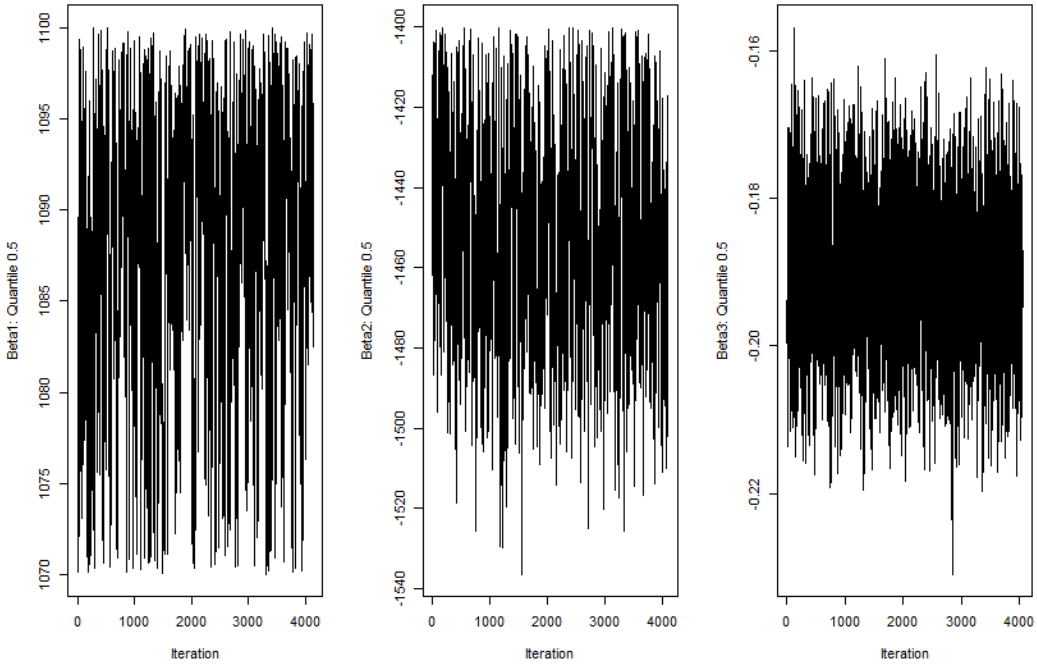

Figure 2.9 – MCMC results for 04LM001 station for $p = 0.5$

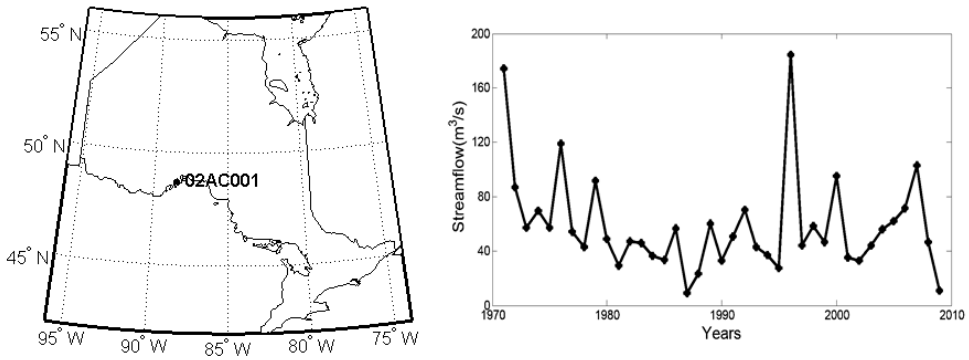

Figure 2.3 – Variation of maximum annual streamflows for each station- Application (a)

Chapitre 3

Copula-Based Conditional Quantile and Inference

Titre traduit

Quantile conditionnel basé sur les copules, méthode et inférence.

Auteurs

Bouchra Nasri[1], Taoufik Bouezmarni[2]

[1] Institut national de recherche scientifique, Eau-Terre-Environnement, 490 rue de la couronne, Quebec,G1K 9A9.

[2] Département de Mathématiques, Université de Sherbrooke, 2500 boul. de l'Université, Sherbrooke, J1K 2R1, Canada.

Contribution des auteurs

Bouchra Nasri: proposition du projet, rédaction de la méthodologie, élaboration des codes, proposition des cas d'étude et co-rédaction de l'article.

Taoufik Bouezmarni: révision de la méthodologie, révision des codes, révision des résultats, co-

rédaction de l'article.

Remerciements

Je tiens à remercier mon co-auteur pour la collaboration et la réussite de ce travail. Je remercie également André St-Hilaire pour la correction de l'article et pour ses commentaires. Aussi, je remercie Taha B.M.J Ouarda pour son grand soutien. Mes remerciements vont finalement à M. Bruno Rémillard pour ses conseils concernant les méthodes d'adéquation pour le choix de copule.

Publication ciblée

Journal: Environmetrics

Date de soumission: Novembre 2016

Résumé

Dans cet article, nous proposons deux nouvelles approches pour estimer les quantiles conditionnels en incorporant une fonction copule qui lie les covariables et la variable d'intérêt. L'idée principale de nos méthodes est d'explorer le lien entre la copule et les fonctions de distribution marginale afin de construire la fonction des quantiles conditionnels. Pour ce faire, deux estimateurs ont été proposés, un paramétrique et l'autre semi-paramétrique. La normalité a été validée/vérifiée, ce qui permet de construire des intervalles de confiance selon une approche asymptotique. En comparant l'erreur quadratique moyenne intégrée de nos estimateurs et celle proposée dans la littérature, nous avons montré la performance de nos approches proposées. En fait, inclure l'information de dépendance 'copule' pour l'estimation des quantiles conditionnels a permis de fournir une large amélioration de l'estimation et de réduire le biais. Enfin, ces deux estimateurs ont été utilisés pour estimer les quantiles conditionnels pour les deux cas d'étude suivants: (i) l'estimation des quantiles du débit maximum annuel pour une station en Ontario en présence de l'indice d'oscillation multidécennale de l'Atlantique (AMO), (ii) l'estimation des quantiles conditionnels des précipitations annuelles maximales pour une station en Californie en présence de l'oscillation décennale du Pacifique (PDO)

et de l'indice de l'oscillation australe (SOI).

Abstract

In this paper, we propose two new approaches to estimate conditional quantiles by incorporating the copula functions of the covariates and the response variable. The main idea behind our methods is to explore the link between copula functions and marginal distribution functions in order to construct the conditional quantile function. Parametric and semiparametric estimators are proposed for the conditional quantile functions. The asymptotic normality is established and a finite-sample of these estimators is investigated. By comparing the integrated mean squared error of our estimators and those proposed in the literature, we show the performance of our proposed approaches. In fact, including the dependence information in the estimation provides a huge estimation improvement. Finally, two applications of hydro-climatic data were used to illustrate the usefulness of the proposed estimators. The first application studies the maximum annual streamflow from one station in Ontario (Canada) and investigates the case of one covariate (AMO). The second application explores the maximum annual precipitations at the Randsburg Meteorological Station in California (USA) by using two covariates (PDO and SOI).

Keyword

Conditional quantiles, copula, covariates, parametric estimation, semiparametric estimation, hydrology, climatology.

3.1 Introduction

Understanding the temporal variability of hydrological processes and their associated statistics is essential for better water resource management. Frequency analysis of hydrologic data has been widely used for problems related to engineering design, flood risk management, river navigation

planning and water quality management [e.g., Vogel *et al.*, 1993; Yu *et al.*, 2015; Hirabayashi *et al.*, 2013]. Generally, current methods of hydrological frequency analysis have been most often based on the estimation of unconditional quantile, also called in hydrology stationary quantiles [1]. In fact, unconditional quantile functions can be modelled by estimating the cumulative distribution function (c.d.f). Therefore, the quantile function is equal to the generalized inverse of the c.d.f, given some probability level. Parametric and nonparametric methods were used, in general, to estimate the c.d.f function. In the parametric estimation, the c.d.f is assumed to belong to a parent distribution with known parameters (i.e, Gamma, Lognormal, Generalized Extreme Value, or GEV, etc.). In the nonparametric estimation, there are several methods of estimating the c.d.f, for example, the empirical distribution function [e.g., Van der Vaart, 1998; Coles, 2001] and kernel distribution functions [Yamato, 1973] among others. The estimation of unconditional quantile functions has been the subject of several studies in hydrology in the 80s and 90s [e.g., Buishand, 1984, 1989, 1991; Carter & Challenor, 1981; Cunnane, 1989; Grehys, 1996; Madsen *et al.*, 1997; Lang *et al.*, 1999]. The quantile function estimation assumes that the random variables are independent and identically distributed (i.i.d.). But, in the context of climate change, the distributions of hydrological series are likely to be modified. As a consequence, the quantile function also varies over time and/or covariates. Recently, some methods have been proposed to model the conditional quantile, also called non-stationary quantile in hydrology, by taking into account the covariates and their variability. In fact, two approaches are mostly used in the literature. The first approach incorporates covariates through the distribution parameters. In fact, the parameters will vary according to the covariates. The link between these parameters and the covariates can be linear [e.g., Coles, 2001; Cannon, 2010] or nonlinear [e.g., Chavez-Demoulin & Davison, 2005; Neville *et al.*, 2011; Nasri *et al.*, 2013, 2016].

1. Basically, strict-sense stationarity means that the distribution remains constant over time. From a practical point of view, hydrologists assume second-order stationarity, which implies that the first two moments (mean and variance) are constant. In time series analysis, the second-order stationarity requires that the covariance depends only on the horizon and does not change over time. Here, this condition is not required [Meylan *et al.*, 2012].

The second approach is based on the quantile regression models. Parametric models, including linear and non linear models, are investigated by [Koenker & Bassett, 1987; Koenker, 2005] and nonparametric quantile regression models are studied by [Hendricks & Koenker, 1992; Koenker *et al.*, 1994]. However, all these previous studies estimate the conditional quantile based only on the information on the marginal distribution function or on the regression between the covariates and the response variable. Recently, [Noh *et al.*, 2013] have proposed a new semi-parametric approach based on the quantile regression estimator weighted by a copula function. In fact, the quantile regression model is rewritten in terms of copula density function and the marginal distributions.

In this paper, we present two new approaches for estimating the conditional quantile function by incorporating the copula function. In fact, our proposed estimators are based on the conditional quantile functions. These functions are reformulated in terms of copula functions and their marginal distributions. The first approach proposes a parametric estimator for the conditional quantile function by assuming that the copula function and the marginal distributions follow a parametric model with unknown parameters. The second approach proposes a semi-parametric estimator for the conditional quantile function by supposing a parametric model for a copula function with unknown parameters and by estimating the marginal function non-parametrically. The asymptotic normality of the proposed estimators is established. Also, comparison between our estimators and those proposed in the literature are studied. Finally, the new estimators are applied for two case studies based on hydro-climatic data. The first application studies the maximum annual streamflow from one station in Ontario (Canada) and investigates the case of one covariate. The second application explores the maximum annual precipitations at the Randsburg Meteorological Station in California (USA) by using two covariates.

The paper is organized as follows: Section 2 presents an overview of the conditional quantile estimators developed in the literature and our two proposed estimators. In Section 3, we establish the

asymptotic i.i.d. representation of the new estimators. In Section 4, we investigate the finite-sample performance of our approaches, compared to the existing estimators, by running simulations based on various models. We analyze, in Section 5, real data from two case studies by applying our proposed estimators. Section 6 concludes.

3.2 Conditional quantile estimators

3.2.1 Existing estimators of the conditional quantile

In the literature, two conditional quantile estimators are frequently used. The first approach incorporates the covariates in the distribution parameters and the second method is related to quantile regression models.

— *The first approach: covariates in the parameters of the distribution*

Let $\boldsymbol{X} = (X_1, \ldots, X_d)$ be a covariate vector of dimension d ($d \geq 1$), Y be a response variable with a continuous c.d.f, F_Y, with parameter vector $\boldsymbol{\alpha}$ and τ is a fixed probability ($\tau \in [0, 1]$). The conditional quantile function using the conditional c.d.f function is given by:

$$\begin{aligned} \widehat{Q}_{CDF} &\equiv \widehat{Q}_Y\left(\tau | \boldsymbol{X} = \boldsymbol{x}\right) \\ &= F_Y^{-1}\left(\widehat{\boldsymbol{\alpha}}_{\boldsymbol{X}}, \tau\right), \end{aligned} \tag{3.1}$$

where $\widehat{\boldsymbol{\alpha}}_{\boldsymbol{X}}$ can be a linear [e.g., Coles, 2001; Cannon, 2010] or a nonlinear [e.g., Chavez-Demoulin & Davison, 2005; Neville *et al.*, 2011; Nasri *et al.*, 2013, 2016] function of the covariables $\boldsymbol{X}$.

— *The second approach: quantile regression models*

The conditional quantile regression model is defined by:

$$\begin{aligned} \widehat{Q}_{Qr} &\equiv Q_Y\left(\tau|\boldsymbol{X}=\boldsymbol{x}\right) \\ &= \arg\min_{\beta} \mathrm{E}\left[\rho_\tau\left(Y-s\left(\boldsymbol{X},\boldsymbol{\beta}\right)\right)\right], \end{aligned} \tag{3.2}$$

where ρ_τ is the loss function [1], s is a predetermined link function between Y and $\boldsymbol{X}$ and $\boldsymbol{\beta}$ is a set of unknown coefficients to be estimated. For example, in the case of the linear quantile regression model, the link function is given by $s(\boldsymbol{X},\boldsymbol{\beta}) = \boldsymbol{X'\beta}$; for more details see [Koenker & Bassett, 1987]. Several parametric and nonparametric methods for estimating the quantile regression function are studied in Koenker [2005].

However, in the two aforementioned approaches, the relation between $\boldsymbol{X}$ and Y is contained either in the parameters vector $\widehat{\boldsymbol{\alpha}}$ or in the link function s, both of which are too restrictive to fully describe the dependence. One way to model this dependence is to use the copula function [Nelson, 1999]. Recently [Noh *et al.*, 2015] have investigated copula for estimating the conditional quantile regression function. In their approach, they studied a semiparametric estimator for a conditional quantile function. This estimator is based on a regression model, see Equation (3.2), weighted by a parametric copula density function of Y and $\boldsymbol{X}$. In fact, the right term of Equation (3.2) is replaced by a copula density function. Let us denote by F_Y, $\boldsymbol{F}$ the c.d.f function of Y and $\boldsymbol{X}$ respectively and $C(.,.,\boldsymbol{\theta})$ (resp. $c(.,.,\boldsymbol{\theta})$) the copula function (resp. the copula density function) with $\boldsymbol{\theta}$ the vector of copula parameters to be estimated. Please see the supplementary material for the definition of the copula function, Sklar's theorem and some copula families. [Noh *et al.*, 2015] have proposed the following conditional quantile

1. $\rho_\tau = u\left(\tau - \mathbb{I}\left(u<0\right)\right)$, where $\mathbb{I}$ is an indicator function. In the case of median quantile regression $\tau = .5$, then the check function takes $-\frac{u}{2}$ when $u<0$ and $\frac{u}{2}$ when $u \geq 0$.

estimator:

$$\widehat{Q}_{nLr} \equiv Q_Y(\tau|\boldsymbol{X}=\boldsymbol{x}) \tag{3.3}$$

$$= \arg\min_a \mathrm{E}\left[\rho_\tau(Y-a)\,\hat{c}\left(\widehat{F_Y}(y), \widehat{\boldsymbol{F}}(\boldsymbol{x}), \widehat{\boldsymbol{\theta}}\right)\right] \tag{3.4}$$

where $\widehat{\boldsymbol{\theta}}$ is a consistent estimator of the copula function, $\widehat{F_Y}$ and $\widehat{\boldsymbol{F}}$ are, respectively, the empirical c.d.f functions of Y and $\boldsymbol{X}$.

3.2.2 Copula-based conditional quantile estimators

Relationship between copula and the conditional quantile function

The expression of conditional distribution function in terms of copula function and the marginal distributions is established in [Bouyé & Salmon, 2002] and given by:

$$F_{Y|\boldsymbol{X}}(y|\boldsymbol{x}) = \tilde{C}(F_0(y), \boldsymbol{F}(\boldsymbol{x})) \tag{3.5}$$

where $F_{Y|\boldsymbol{X}}$ is the conditional distribution of Y given $\boldsymbol{X}$ and

$$\tilde{C}(u_0, u_1, \ldots, u_d) = \frac{f_1(x_1)\ldots f_d(x_d)}{f(x_1,\ldots,x_d)} \frac{\partial^d C(u_0, u_1, \ldots, u_d)}{\partial u_1 \ldots \partial u_d},$$

where f_j (resp. F_j) the density (resp. distribution) function of X_j, $U_j = F_j(X_j)$, $j = 1, \ldots, d$ and $U_0 = F_Y(Y)$.

Note that, if $d = 1$ or $X_1, \ldots, X_d$ are independent, then $\tilde{C}(u_0, u_1, \ldots, u_d)$ becomes:

$$\tilde{C}(u_0, u_1, \ldots, u_d) = \frac{\partial^d C(u_0, u_1, \ldots, u_d)}{\partial u_1 \ldots \partial u_d}.$$

The conditional quantile function can be defined as the inverse of the conditional distribution. Fixing the conditional distribution given by (3.5) at some τ, the conditional quantile function is giving by:

$$Q_Y\left(\tau|\boldsymbol{X}=\boldsymbol{x}\right)=F_Y^{-1}\left[\Gamma\left(\boldsymbol{F}\left(\boldsymbol{x}\right),\tau\right)\right], \tag{3.6}$$

where Γ is the partial inverse, with respect to the second argument of $\tilde{C}$, and F_Y^{-1} is the generalized inverse function of F_Y. Below, we give some examples for conditional quantile functions based on copula in the case of one covariate, for more details see [Xie, 2015; Koenker & Bassett, 1978; Bouyé & Salmon, 2002].

— **Examples of Archimedean copula-based conditional quantile**

The τth quantile function with one covariate for the Archimedean copula with generator $\varphi(.)$ is given by

$$Q_Y\left(\tau|X_1=x_1\right)=F_Y^{-1}\left(\varphi^{[-1]}\left[\varphi\left(\varphi'^{[-1]}\left(\frac{1}{\tau}\varphi'\left(F_1\left(x_1\right)\right)\right)\right)-\varphi\left(F_1\left(x_1\right)\right)\right]\right),$$

where $\varphi^{[-1]}(.)$ denotes the pseudo-inverse of $\varphi(.)$ and $\varphi'(.)$ is the first order derivative of $\varphi(.)$.

(a) Clayton copula: the generator function of the Clayton copula is given by: $\varphi\left(t\right)=\theta^{-1}\left(t^{\theta}-1\right)$. Hence, $\varphi'\left(t\right)=t^{\theta-1}$, $\varphi'^{[-1]}=t^{\frac{1}{\theta-1}}$ and $\varphi^{[-1]}\left(t\right)=\left(\theta t+1\right)^{\frac{1}{\theta}}$. Therefore, the τth quantile function of Clayton copula is given by

$$Q_Y\left(\tau\mid X_1=x_1\right)=F_Y^{-1}\left[\left(\left(\tau^{\frac{-\theta}{1+\theta}}-1\right)F_1^{-\theta}\left(x\right)+1\right)^{-\frac{1}{\theta}}\right].$$

(b) Gumbel copula: the generator function of the Gumbel copula is given by: $\varphi=\left(-\log\left(t\right)\right)^{\theta}$. In this case, an explicit formula for the pseudo-inverse function of $\varphi'\left(t\right)$ does not exist.

Therefore, the τth quantile function of Gumbel copula does not have an explicit expression but can be found numerically.

(c) Frank copula: the generator function of the Frank copula is given by: $\varphi(t) = -\log\left(\frac{\exp(-\theta t)-1}{\exp(-\theta)-1}\right)$. Hence, $\varphi'(t) = \frac{\theta\exp(-\theta t)}{\exp(-\theta t)-1}$; $\varphi^{[-1]}(t) = -\frac{\log(t\exp(-\theta)-t+1)}{\theta}$ and $\varphi'^{[-1]}(t) = \frac{-1}{\theta}\log\left(\frac{t}{t-\theta}\right)$. Therefore, the τth quantile function of the Frank copula is given by:

$$Q_Y(\tau \mid X_1 = x_1) = F_Y^{-1}\left[-\frac{1}{\theta}\log\left(1-(1-\exp(-\theta))\left[1+\exp(-\theta F_1(x_1))\left(\tau^{-1}-1\right)\right]^{-1}\right)\right].$$

Remark 1 *For the three copulas given in (a,b and c), we see that when θ tends towards zero (i.e. X_1 and Y tends towards independency), the quantile $Q_Y(\tau \mid X_1 = x_1)$ tends towards $F_Y^{-1}(\tau)$ which corresponds to the unconditional quantile.*

— **Examples of Elliptical copula-based conditional quantile**

(d) Gaussian copula: the τth quantile function with one covariate for the Gaussian copula is given by:

$$Q_Y(\tau|X_1 = x_1) = F_Y^{-1}\left[\Phi\left(\tilde{\rho}\Phi^{-1}(F_1(x_1)) + \Phi^{-1}(\tau)\sqrt{1-\tilde{\rho}^2}\right)\right],$$

where $\Phi(.)$ denotes the standard normal distribution function and $\tilde{\rho}$ is the Pearson correlation coefficient between X_1 and Y.

(e) Student t-copula: the τth quantile function with one covariate for the Student t-copula is given by:

$$Q_Y(\tau|X_1 = x_1) = F_Y^{-1}\left[t_d\left(\tilde{\rho}\, t_d^{-1}(F_1(x_1)) + t_{d+1}^{-1}(\tau)\sqrt{\frac{d+\left[t_d^{-1}(F_1(x_1))\right]}{d+1}(1-\tilde{\rho}^2)}\right)\right],$$

where $t_d(.)$ is the Student t-distribution function with d degrees of freedom and $\tilde{\rho}$ is the Pearson correlation coefficient between X_1 and Y.

Remark 2 *For the two elliptical copulas, presented in (d and e), we see that when $\tilde{\rho}$ tends towards zero, i.e. the independent case, the quantile $Q_Y(\tau \mid X_1 = x_1)$ tends towards $F_Y^{-1}(\tau)$.*

Proposed estimators

The expression given in Equation (3.6) allows more flexibility for estimating the conditional quantile. Indeed, parametric, semi-parametric or nonparametric approaches, depending on the estimation method of the marginal distributions and the copula function, could be used in order to estimate the conditional quantile function. For example, we can estimate the marginal distributions and the copula function non-parametrically. However, this method suffers from the curse of dimensionality. In this paper, we propose both parametric and semi-parametric approaches.

Parametric estimator: The parametric approach assumes a parametric model for the copula function C, denoted by $C(.,.,\boldsymbol{\theta})$, and parametric models for the marginal distributions F_Y and $\boldsymbol{F}$, denoted by $F_Y(.,\boldsymbol{\alpha})$ and $\boldsymbol{F}(.,\boldsymbol{\beta})$ respectively. Then, the proposed estimator is given by:

$$\widehat{Q}_Y^p(\tau|\boldsymbol{X}=\boldsymbol{x}) = F_Y^{-1}\left[\Gamma\left(\boldsymbol{F}\left(\boldsymbol{x},\widehat{\boldsymbol{\beta}}\right),\widehat{\boldsymbol{\theta}},\tau\right),\widehat{\boldsymbol{\alpha}}\right] \tag{3.7}$$

$$\equiv H_{\boldsymbol{x}}^p\left(\widehat{\boldsymbol{\alpha}},\widehat{\boldsymbol{\beta}};\widehat{\boldsymbol{\theta}},\tau\right), \tag{3.8}$$

where $\widehat{\boldsymbol{\alpha}}$, $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\theta}}$ are the estimators of the parameters $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$. Two methods for estimating the parameters are developed in the literature. First, we can estimate simultaneously $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ using the complete maximum likelihood, see [Shih & Louis, 1995] and [Joe, 1997]. However, this

method requires intensive calculations and sometimes the optimization problem is difficult to solve. Second, [Joe & Xu, 1996] have proposed a two-step process, called inference function for margins (IFM), in order to estimate the marginal functions and copula parameters. This method estimates, in the first step, the marginal function parameters and the copula parameters in the second step. The IFM method is frequently used in the literature because it is efficient and easy to implement. For more details, see Oakes [1982], Romano [2002] and Joe [2005]. Here, we use the IFM method for estimating the parameters of the model.

To illustrate the parametric estimator for the conditional quantile, we consider the Frank, Clayton and Gaussian copulas and the standard normal distribution as margin for covariate and lognormal distribution with parameters $(0, 0.25)$ as margin for the variable of interest. For the two Archimedean copulas, observations are generated from a bivariate copula with $\theta = 8$ which corresponds to Kendall's tau $\tau_c = 0.6$ for Frank and $\tau_c = 0.8$ for the Clayton model. For the Gaussian model, observations are generated from a bivariate copula with correlation coefficient $\rho = 0.9$. In Figure 3.1, we provide the true conditional median and its parametric estimator curves for one realization of sample size $n = 200$.

Remark 3 *In practice, the copula model and the margins are unknown and they have to be selected. To choose the best copula model for fitted data, several goodness-of-fit procedures have recently been proposed to this end. These goodness-of-fit tests can be divided into three broad classes: (1) tests based on the probability integral transformation of Rosenblatt's*[1] *[Breymann* et al.*, 2003; Dobric & Schmid, 2005; Genest* et al.*, 2009], (2) tests that involve kernel smoothing [Fermanian, 2005; Scaillet, 2007] and (3) omnibus tests derived from continuous functionals of the empirical copula process and Kendall's process [Genest* et al.*, 2006]. Genest* et al. *[2009] have provided a comparison of all copulas goodness-of-fit criteria given in classes (1) and (3) and they recommended the use of*

1. [Rosenblatt, 1952], for more information about Resenblatt's transformation

their two proposed goodness-of-fit tests based on the Rosenblatt's transform. These tests are noted $S_n(C)$ and $S_n(B)$. In the case studies of our paper, only these tests will be used. Note that several packages in R are developed for the copula goodness-of fit tests. In this work, we will use "copula" package developed by [Hofert et al.*, 2016]. To choose the marginal distributions for variables of interest and covariates, the Bayesian information criterion (BIC) [Schwarz, 1978] will be used.*

Semiparametric estimator: The semiparametric approach assumes a parametric model for the copula, denoted by $C(.,.,\boldsymbol{\theta})$, and estimates nonparametrically the marginal distributions. To estimate the copula parameters, we use the Canonical Maximum Likelihood (CML) proposed by [Genest *et al.*, 1995]. In fact, the CML approach does not requires assumptions on the marginal distributions. The proposed estimator for the semiparametric approach is defined as follows:

$$\widehat{Q}_Y^{sp}(\tau|\boldsymbol{X}=\boldsymbol{x}) = F_Y^{-1}\left[\Gamma\left(\widehat{\boldsymbol{F}}(\boldsymbol{x});\widehat{\boldsymbol{\theta}},\tau\right)\right] \tag{3.9}$$

$$\equiv H_{\boldsymbol{x}}^{sp}\left(\widehat{F_Y},\widehat{\boldsymbol{F}},\widehat{\boldsymbol{\theta}},\tau\right) \tag{3.10}$$

where $\widehat{\boldsymbol{\theta}}$ is the CML estimator of the copula parameter of $\boldsymbol{\theta}$ and $\widehat{F_0}$, $\widehat{\boldsymbol{F}}$ are the rescaled empirical distribution function of F_Y and $\boldsymbol{F}$ defined by:

$$\widehat{F_Y}(y) = (n+1)^{-1}\sum_{i=1}^{n} I(Y_i \leq y) \tag{3.11}$$

$$\widehat{\boldsymbol{F}}(\boldsymbol{x}) = (n+1)^{-1}\sum_{i=1}^{n} I(\boldsymbol{X}_i \leq \boldsymbol{x}) \tag{3.12}$$

where $(Y_i, \boldsymbol{X}_i)$, $i = 1 \ldots n$, is an independent and identically distributed (i.i.d) sample of n observations generated from the distribution of $(Y, \boldsymbol{X})$. To illustrate the performance of the proposed

semi-parametric model, we consider the same three models presented before in Figure 3.1. Figure 3.2 presents the true conditional median and its semi-parametric estimator.

Remark 4 *In this case, we will also need to select the copula model. Note that, the same tests and package described in Remark 3 will be used.*

3.3 Theoretical Results

In this section, we establish the asymptotic i.i.d representation for the two proposed estimators. This representation implies that the estimators follow asymptotically a normal distribution with finite variance. The results, in this section, show that the performance of the proposed estimators increases with the sample size, i.e., the estimators are close to the true conditional quantile function when the number of observations is large. Also, the asymptotic normality of our estimators allows the construction of the confidence interval. For the rest of the paper, we assume that the regularity conditions given in [Joe & Xu, 1996] and [Noh *et al.*, 2013] are satisfied. Supplementary material contains the proof of the two theorems in this section.

3.3.1 Convergence of the parametric estimator

In this section, we give the asymptotic i.i.d of $\sqrt{n}\left(H_{\boldsymbol{x}}^{p}\left(\widehat{\boldsymbol{\alpha}},\widehat{\boldsymbol{\beta}};\widehat{\boldsymbol{\theta}},\tau\right)-H_{\boldsymbol{x}}^{p}\left(\boldsymbol{\alpha},\boldsymbol{\beta};\boldsymbol{\theta},\tau\right)\right)$ and we derive the asymptotic distribution of the conditional quantile parametric estimator. The following notations will be needed:

Notations

— $H^p_{\boldsymbol{x},1}(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\theta},.)$, $H^p_{\boldsymbol{x},2}(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\theta},.)$ and $H^p_{\boldsymbol{x},3}(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\theta},.)$ denote the partial derivative of $H^p_{\boldsymbol{x}}(.)$ with respect to $\boldsymbol{\alpha},\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ respectively.

— $\boldsymbol{\eta}=(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\theta})$ is the vector of the parametric model and $\widehat{\boldsymbol{\eta}}=\left(\widehat{\boldsymbol{\alpha}},\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\theta}}\right)$ is the vector of the IFM estimators.

The next theorem provides the asymptotic i.i.d representation and the asymptotic normality for the parametric estimator of the conditional quantile.

Théorème 1 *Under regular conditions in Joe & Xu [1996] and if $H^p_{\boldsymbol{x}}$ admits a continuous first derivative, we have*

$$\sqrt{n}\left(H^p_{\boldsymbol{x}}\left(\widehat{\boldsymbol{\alpha}},\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\theta}},\tau\right)-H^p_{\boldsymbol{x}}(\boldsymbol{\alpha},\boldsymbol{\beta};\boldsymbol{\theta},\tau)\right)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\gamma_i+o_p(1)$$

where γ_i are i.i.d random vectors, depending on $(Y_i,\boldsymbol{X_i},\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\theta},\tau,\boldsymbol{x})$, with zero mean and finite variance given by

$$E\left(\gamma_i^2\right)=\nabla H^p_{\boldsymbol{x}}(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\theta},\tau)\left(\mathcal{G}(\boldsymbol{\eta})\right)\nabla H^p_{\boldsymbol{x}}(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\theta},\tau)^T$$

with $\nabla H^p_{\boldsymbol{x}}(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\theta},\tau)=\left(H^p_{\boldsymbol{x},1}(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\theta},.),H^p_{\boldsymbol{x},2}(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\theta},.),H^p_{\boldsymbol{x},3}(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\theta},.)\right)$ and $(\mathcal{G}(\boldsymbol{\eta}))$ is the Godambe information matrix. Therefore, $\sqrt{n}\left(H^p_{\boldsymbol{x}}\left(\widehat{\boldsymbol{\alpha}},\widehat{\boldsymbol{\beta}},\boldsymbol{\theta},\tau\right)-H^p_{\boldsymbol{x}}(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\theta},\tau)\right)$ is asymptotically normally distributed, i.e.,

$$\sqrt{n}\left(H^p_{\boldsymbol{x}}\left(\widehat{\boldsymbol{\alpha}},\widehat{\boldsymbol{\beta}},\boldsymbol{\theta},\tau\right)-H^p_{\boldsymbol{x}}(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\theta},\tau)\right)\xrightarrow{d}N\left(\mathbf{0},\nabla H^p_{\boldsymbol{x}}(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\theta},\tau)\left(\mathcal{G}(\boldsymbol{\eta})\right)\nabla H^p_{\boldsymbol{x}}(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\theta},\tau)^T\right).$$

■

Remark 5 *Theorem 1 shows the asymptotic normality of our parametric estimator with unknown variance. In order to construct the confidence interval, we suggest to use resampling methods, for example bootstrap or Jackknife, to estimate variance.*

3.3.2 Convergence of the Semiparametric Estimator

In this section, we give the asymptotic i.i.d of $\sqrt{n}\left(H_{\boldsymbol{x}}^{sp}\left(\widehat{F_0},\widehat{\boldsymbol{F}},\widehat{\boldsymbol{\theta}},\tau\right)-H_{\boldsymbol{x}}^{sp}\left(F_0,\boldsymbol{F},\boldsymbol{\theta},\tau\right)\right)$ and we deduce the asymptotic distribution of the conditional quantile semiparametric estimator. The following notations will be needed:

Notations

— $H_{\boldsymbol{x},1}^{sp}\left(F_Y,\boldsymbol{F},\boldsymbol{\theta},\tau\right)$, $H_{\boldsymbol{x},2}^{sp}\left(F_Y,\boldsymbol{F},\boldsymbol{\theta},\tau\right)$ and $H_{\boldsymbol{x},3}^{sp}\left(F_0,\boldsymbol{F},\boldsymbol{\theta},\tau\right)$ denotes the partial derivative of $H^{sp}\left(.\right)$ with respect to the first three components respectively.

— $\boldsymbol{\kappa}=\left(F_Y,\boldsymbol{F},\boldsymbol{\theta},\tau\right)$ is the vector of the semiparametric model and $\widehat{\boldsymbol{\kappa}}=\left(\widehat{F_0},\widehat{\boldsymbol{F}},\widehat{\boldsymbol{\theta}},\tau\right)$ is the vector of the CML estimators.

Théorème 2 *Under regular conditions in Noh* et al. *[2013] and if H^{sp} admits a continuous first derivative, we have*

$$\sqrt{n}\left(H_{\boldsymbol{x}}^{sp}\left(\left(\widehat{F_Y},\widehat{\boldsymbol{F}};\widehat{\boldsymbol{\theta}},\tau\right)-H_{\boldsymbol{x}}^{sp}\left(F_Y,\boldsymbol{F},\boldsymbol{\theta},\tau\right)\right)\right)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\xi_i+o_p\left(1\right)$$

where ξ_i are i.i.d random vectors, depending on $(Y_i,\boldsymbol{X_i},\boldsymbol{\theta},\tau,\boldsymbol{x})$, with zero mean and finite implicit variance Σ.

Therefore, $\sqrt{n}\left(H_{\boldsymbol{x}}^{sp}\left(\widehat{F_Y},\widehat{\boldsymbol{F}},\widehat{\boldsymbol{\theta}},\tau\right)-H_{\boldsymbol{x}}^{sp}\left(F_Y,\boldsymbol{F},\boldsymbol{\theta},\tau\right)\right)$ is asymptotically normally distributed, i.e.,

$$\sqrt{n}\left(H_{\boldsymbol{x}}^{sp}\left(\widehat{F_Y},\widehat{\boldsymbol{F}};\widehat{\boldsymbol{\theta}},\tau\right)-H_{\boldsymbol{x}}^{sp}\left(F_Y,\boldsymbol{F},\boldsymbol{\theta},\tau\right)\right)\xrightarrow{d}N\left(\boldsymbol{0},\Sigma\right).$$

■

Remark 6 *Theorem 2 shows the asymptotic normality of our parametric estimator with unknown and implicit variance. In order to construct the confidence interval, we suggest to use the same resampling methods listed earlier to estimate variance.*

3.4 Simulation

The objective of this section is to compare our parametric and semi-parametric estimators with the following estimators:

— Estimator based on the inverse of c.d.f with linear conditional parameter $\boldsymbol{\alpha} = \boldsymbol{X'\beta}$, where $\boldsymbol{\beta}$ are unknown parameters to be estimated (see Equation (3.1)).

— Estimator based on linear quantile regression model (see Equation (3.2)).

— Estimator based on copula-weighted quantile regression (see Equation (3.3)).

To this end, we consider the following data generating procedures (DGP):

— **DGP a** $(F_Y(Y), F_1(X_1)) \sim$ bivariate Frank copula with parameter $\theta = -3$; $Y \sim \mathcal{W}(\lambda = 0.5, \kappa = 2)$ (2-parameters Weibull distribution, λ is the scale parameter and κ is the shape parameter) and $X_1 \sim \mathcal{N}(\mu = 0, \sigma = 1)$.

— **DGP b** $(F_Y(Y), F_1(X_1)) \sim$ bivariate Frank copula with parameter $\theta = -3$; $Y \sim \ln\mathcal{N}(\mu = 0.5, \sigma = 2)$ (log-normal distribution) and $X_1 \sim \mathcal{N}(\mu = 0, \sigma = 1)$.

— **DGP c** $(F_Y(Y), F_1(X_1)) \sim$ bivariate Gumbel copula with parameter $\theta = 3$; $Y \sim \ln\mathcal{N}(\mu = 0.5, \sigma = 2)$ and $X_1 \sim \mathcal{N}(\mu = 0, \sigma = 1)$.

— **DGP d** $(F_Y(Y), F_1(X_1)) \sim$ bivariate Clayton copula with parameter $\theta = 3$; $Y \sim \ln\mathcal{N}(\mu = 0.5, \sigma = 2)$ and $X_1 \sim \mathcal{N}(\mu = 0, \sigma = 1)$.

— **DGP e** $(F_Y(Y), F_1(X_1), F_2(X_2)) \sim$ trivariate Clayton copula with parameter $\theta = 8$; $Y \sim \ln\mathcal{N}(\mu = 0.5, \sigma = 2)$, $X_1 \sim \mathcal{N}(\mu = 0, \sigma = 1)$ and $X_2 \sim \mathcal{N}(\mu = 0, \sigma = 1)$. We suppose that X_1 and X_2 are independent.

The resulting quantile function of **DGP a**, **b** and **d** are given in Example (a) and (c) for Archimedean copulas (Paragraph 3.2.2). The corresponding quantile function of **DGP c** is calculated numerically; for more information please see Example (b) for Archimedean copulas (Paragraph 3.2.2). For **DGP e**, the quantile function is defined as follows:

$$Q_Y(\tau \mid X_1, X_2 = x_1, x_2) = F_Y^{-1}\left[\left(\left(\frac{\tau F_1(X_1)^{1+\theta} F_2(X_2)^{1+\theta}}{1+\theta}\right)^{\frac{-\theta}{2\theta+1}} - F_1(X_1)^{-\theta} - F_2(X_2)^{-\theta} + 1\right)^{-1/\theta}\right].$$

For each DGP model, data are generated for different sample sizes $n = 50$, $n = 100$ and $n = 200$. As a comparison criterion, for fixed τ, we calculate the empirical Integrated Mean Squared Error ($IMSE$) which is defined by:

$$IMSE = \frac{1}{N}\sum_{j=1}^{N} ISE\left(\widehat{Q}_Y^j(\tau|\boldsymbol{X} = \boldsymbol{x})\right) := \frac{1}{N}\sum_{j=1}^{N}\left[\frac{1}{I}\sum_{l=1}^{I}\left(\widehat{Q}_Y^l(\tau|\boldsymbol{X} = \boldsymbol{x_i}) - Q_Y(\tau|\boldsymbol{X} = \boldsymbol{x_i})\right)^2\right] \tag{3.13}$$

where $\boldsymbol{x_i}, i = 1, \ldots, I$, is a random variable drawn from distribution of $\boldsymbol{X}$ and $\widehat{Q}^l$ is the estimated quantile function from the l-th data sample. Here, we consider $I = 500$ and the iteration number is $N = 1000$.

The $IMSE$ is calculated for the proposed estimators and the compared estimators, for three probabilities values of $\tau = 0.2$, $\tau = 0.5$ and $\tau = 0.8$.

Table 3.1 gives the simulation results. From this table, we can see that our proposed semi-parametric estimator performs better that the estimator based on the inverse c.d.f function and the linear quantile regression estimator. This means that including the dependance information in the estimation

provides a huge estimation improvement. In fact, if we take the case of **DGP a** and for median, we can see that the IMSE is reduced from 0.42 for quantile regression estimator and c.d.f estimator to 0.07 for the semi-parametric estimator. Similar remarks can be seen for other quantile levels and all DGP models.

Also, we remark that, our semiparametric estimator slightly outperforms the semiparametric estimator based on copula weighted quantile regression.

Finally, as expected, the proposed parametric estimator performed better than all the others estimators. In fact, including information concerning the distribution of marginal and copulas improves estimation results.

3.5 Real Data Analysis

3.5.1 Dataset

The two proposed approaches are considered to model the data of the following two case studies. The first case study aims to model the maximum annual streamflow (MAS) at Wolf River station (Environment Canada station # 02AC001) in the province of Ontario, Canada for the period of 1971-2010. This station is located in the northwest portion of this province (48.821°, 88.534°). Figure 3.3 illustrates the geographic location of this station and shows the MAS time series. We consider the 40-year annual multidecadal Oscillation (AMO) index as a covariate, which shows the fluctuation in the sea surface temperature in the North Atlantic Ocean [Teegavarapu *et al.*, 2013]. By using AMO as a covariate in estimating the MAS-quantile function, we consider the effect of climate on hydrological events. The same data have been used for previous work to estimate the conditional quantiles using B-Spline quantile regression model (see article 2). Figure 3.4 shows the results of

conditional quantiles using the B-Spline quantile regression model. The second case study has for goal to model the maximum daily annual rainfall (MAR) at Randsburg station (NOAA station # 047253) in California (USA) for the period of 1938-2007. The Randsburg station is located in the south east of the state of California (35.37°, 117.65°). Figure 3.5 illustrates the geographic location of the Randsburg station and shows the 70-year variation of MAR at Randsburg Station. We consider the 70-year Southern Oscillation Index (SOI) and Pacific Decadal Oscillation (PDO) time series as covariates. The SOI and PDO describe the pressure and temperature anomalies over the Pacific Ocean and have a clear impact on water systems in North America [Bjerknes, 1969; Nathan & Hare, 2002]. By using SOI and PDO as covariates in estimating the conditional quantile functions, we can take into account the effect of multiannual climate fluctuations on rainfall events at this station. These data were used for another study estimating the conditional quantiles using the GEV-B-Spline model [1] [Nasri *et al.*, 2013]. Figure 3.6 shows the results of conditional quantile using the GEV-Spline model.

3.5.2 Choice of copula and margins

Before estimating the conditional quantile functions, for each case study, we have to choose firstly, the best copula function which links the variable of interest (i.e. MAS or MAR) and the covariates (i.e. AMO or (SOI, PDO)) and secondly, the marginal distributions for these variables. Several copula models are compared, including Frank, Clayton, Gumbel, Normal and Student, in order to select the best copula model. This comparison is done by calculating the p-value of $S_n(B)$ and $S_n(C)$ goodness-of fit tests. For the marginal distributions, several cumulative functions are compared, including Normal, 2-parameter Weibull, GEV, lognormal and Gamma. This comparison

1. This model aims to estimate conditional quantile by using Equation (3.1), with F_Y following a GEV distribution and $\boldsymbol{\alpha X'}$ estimated by using a B-Spline function.

is done by using BIC criterion. Table 3.2 gives the results for the best selected copula for the first and the second case studies. The copula models chosen for the first and the second case studies are the Frank and the Normal copulas, respectively. These results mean that in both case studies, we have a symmetric dependence between the variable of interest and the covariates. Table 3.3 shows the results for the marginal distributions and their parameters for the first and the second case studies. For both case studies, the Normal distributions are selected for covariates, while the lognormal and the GEV distributions are chosen for the variable of interest, respectively.

3.5.3 Conditional quantiles results

Finally, having all the needed information about copula functions and margins, we can estimate conditional quantile function using our proposed approaches. Note that the parametric margins will be used only for the parametric approach. Figures 3.7 and 3.8 show the results of the $\tau = 0.5, 0.7$ and 0.9 conditional quantile functions for the first and the second case studies, respectively.

The results show that the conditional quantile curves are very similar for the two approaches and for both case studies. The most obvious remark is that the conditional quantile curves given by semiparametric approach are less smooth than the conditional quantile curve given by the parametric approach. This difference is due to the use of the empirical cumulative function to estimate margins in the semiparametric approach.

By comparing Figure 3.7 to Figure 3.4, we can conclude, in general, that negative (resp. positive) values of the AMO covariate correspond to the high (resp. low) values of MAS-quantiles. Also, we can notice a considerable difference between the maximum median value between the two figures (please see the value of quartile which corresponds to AMO=-0.4) and that is due to the use of nonparametric smoothing functions (e.g., Spline functions, local polynomial functions, etc.). In fact, these smoothing functions divide an interval of data to piecewise intervals and in each piecewise

interval, a polynomial interpolation function is fitted. Thus, in the first part of data which corresponds to negative values of AMO, we can see that the polynomial fitted function follows the two extreme MAS values. This is not the case for copula functions which take into account the structure of the dependence of all observation at the same time. Therefore, the use of copula, in this context, can be useful to overcome this kind of problem. The same problem can be observed if we compare Figure 3.8 to Figure 3.6. In fact, in these figures, it can be seen that generally, the rainfall has a negative dependence with SOI, while it is positively dependant with PDO. The negative values of SOI and positive values of PDO coincide with the relatively high MAR observations. MAR-quantiles increase slowly with increasing PDO values and then increase exponentially for PDO values grather than 1. From Figure 3.6, different inflexion points in the relationship between SOI and MAR are observed (for example at SOI= 1.5, SOI= 0 and SOI= 1.5), that is also due to the use of Spline function. These inflexion points are missed in Figure 1.6.

3.6 Conclusions and Recommandations

Statistical risk assessment is of great importance in hydrology and many other fields of applied statistics. The last two decades have witnessed the development of a number of statistical modeling approaches for extreme quantile functions in the presence of non-stationarity or dependence on covariates.

In this study, we present two new approaches to estimate conditional quantile functions based on copula models. The first approach proposes a parametric estimator for the conditional quantile function and assumes a parametric model for the copula function and for the marginal distributions. The second approach proposes a semiparametric estimator for the conditional quantile function by supposing a parametric model for the copula function and nonparametric model for the marginal

distributions. Under some regularity conditions, the asymptotic normality of the proposed estimators is obtained.

Simulation results show the efficiency of our approaches compared to the approaches proposed in the literature. From the results of both case studies, we concluded the usefulness of copula to formally encapsulate the dependence and the nonlinearity without overestimating results. In fact, the use of copulas can be an alternative to overcome some problems observed in nonparametric or semiparametric conditional quantile models such as Spline quantile regression model or GEV-Spline model. Despite the advantages of the use of our proposed estimators, this kind of model is dependent on the inversion of the derivative of copula function, which can be a limitation in some copula families. In fact, in the case of more than two covariates, some copula families such as Archimedean copulas do not admit an explicit conditional quantile function. Then, the conditional quantiles must be calculated numerically which can be difficult in the presence of several covariates. One way to use our approaches in high dimensional data, is by decomposing the multivariate copula into a bivariate copulas and then use the Vine copula structures. Therefore, future work can focus on the development of some conditional quantile estimators based on Vine copula models.

Acknowledgments

Special thanks to Professor Taha Ouarda and André St-Hilaire for their availability, suggestions and comments. The second author gratefully acknowledges the research support of the Natural Sciences and Engineering Research Council of Canada.

Figures

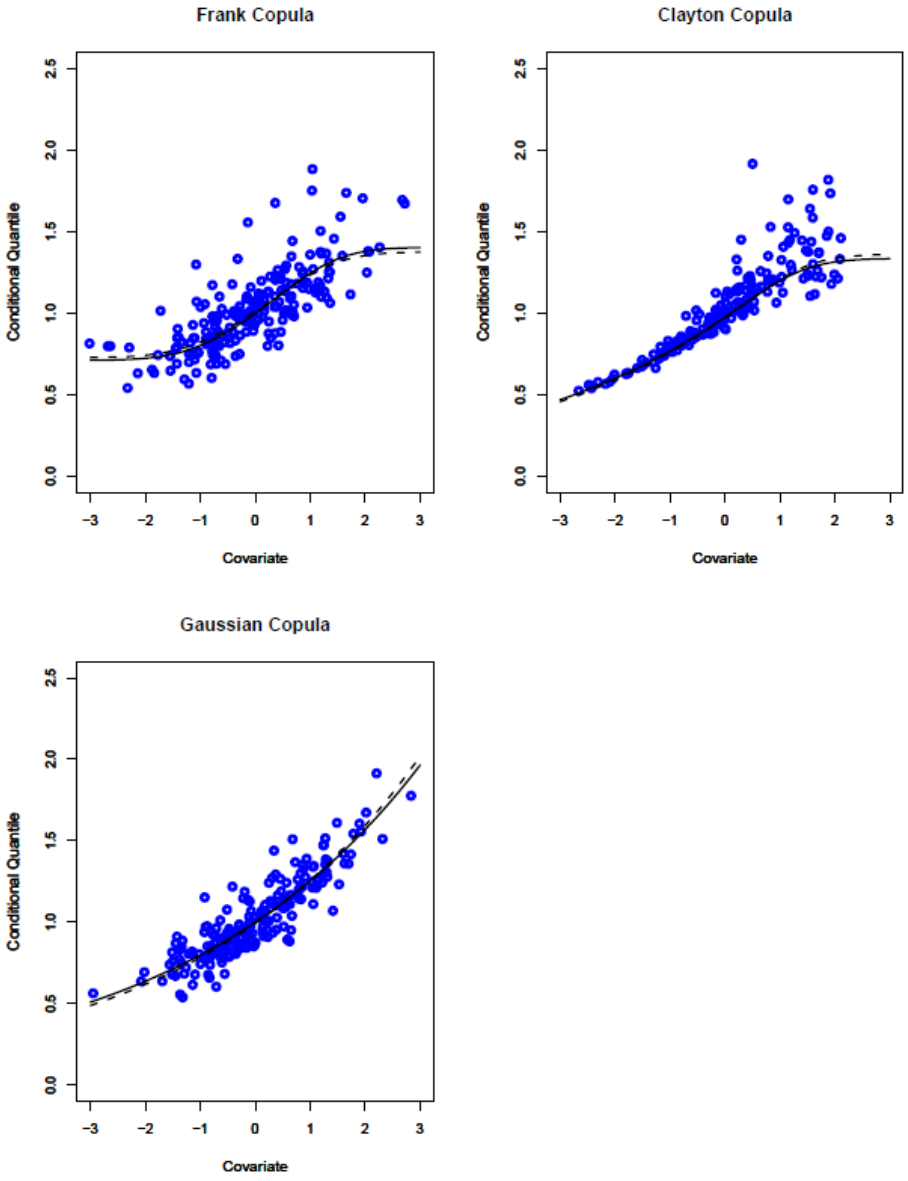

Figure 3.1 – Frank, Clayton and Gaussian Copula models. The figure illustrates 200 observations from bivariate version of Frank, Clayton and Gaussian, respectively. Median true parametric conditional quantile appears in solid curve and the dotted curve represents the estimate.

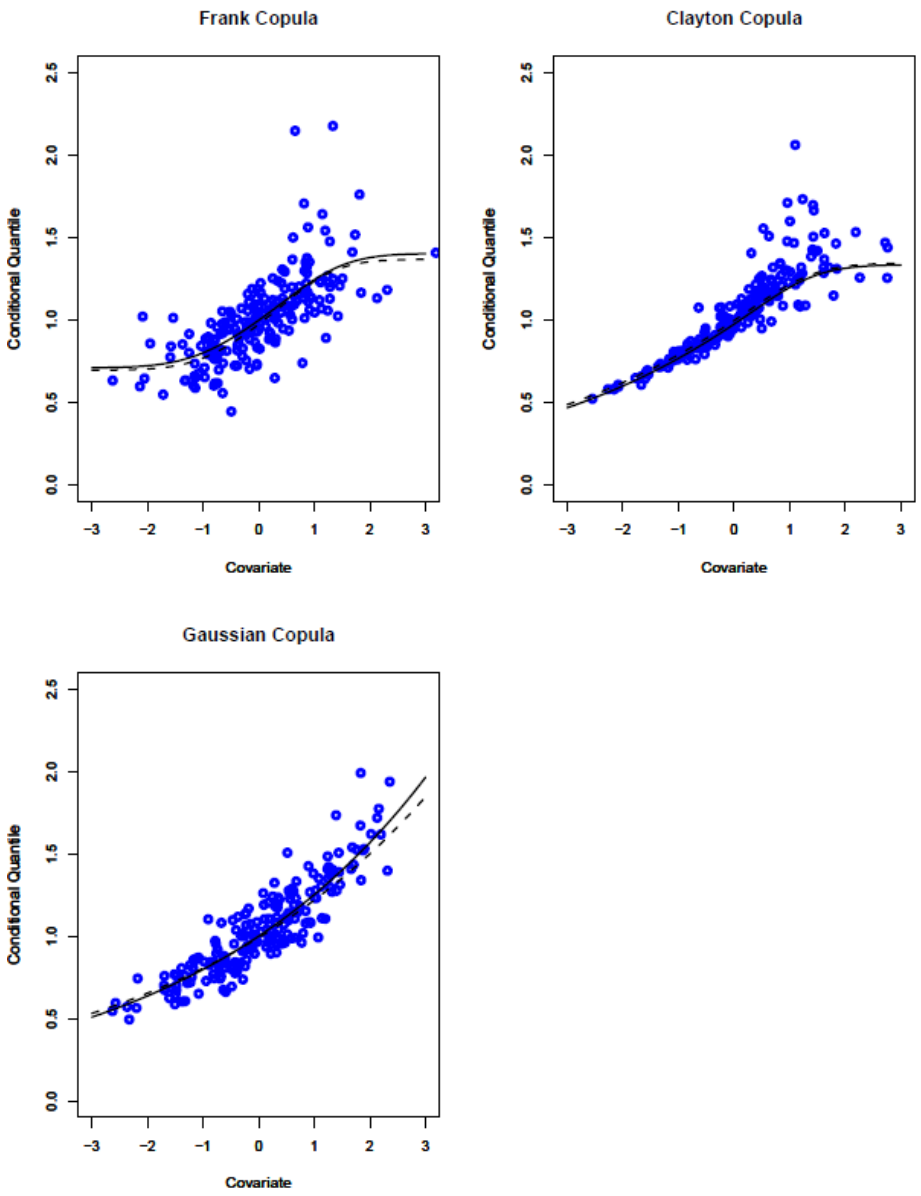

Figure 3.2 – Frank, Clayton and Gaussian Copula models. The figure illustrates 200 observations from bivariate version of Frank, Clayton and Gaussian, respectively. Median true semiparametric conditional quantile appears in solid curve and the dotted curve represents the estimate.

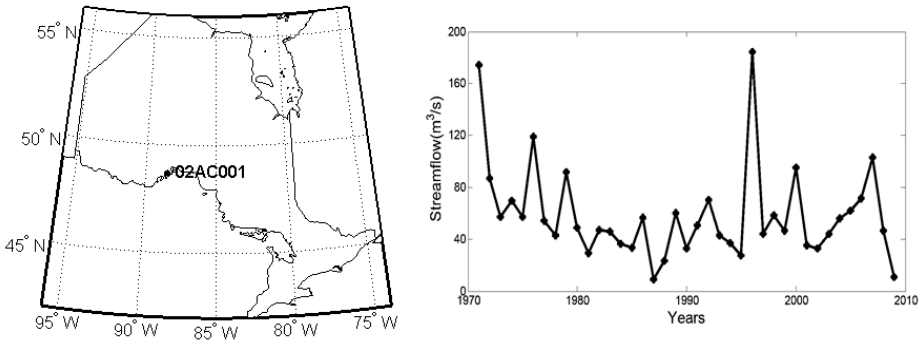

Figure 3.3 – In the left of the figure we see the geographic location of Ontario station and in the right, the time series of annual maximum streamflows.

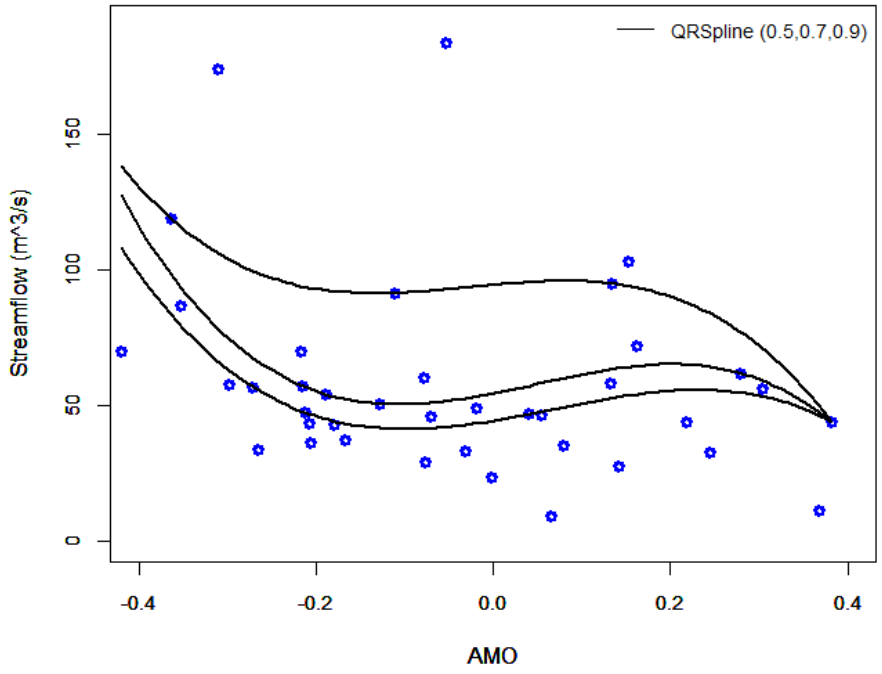

Figure 3.4 – $\tau = 0.5$, 0.7 **and** 0.9 **Conditional quantiles using B-Spline quantile regression model-case study 1**

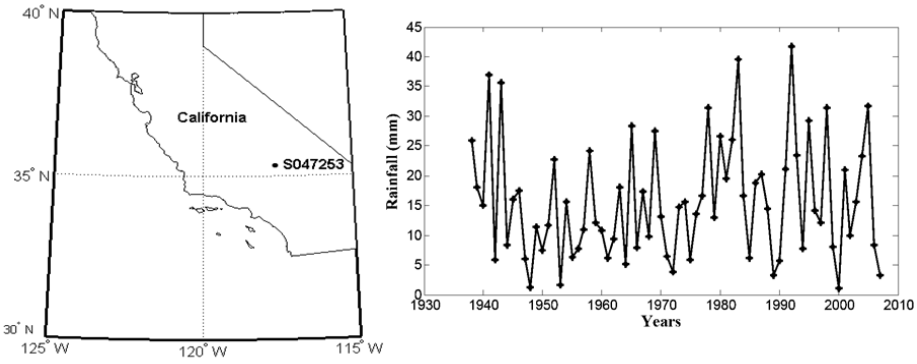

Figure 3.5 – In the left of the figure we see the geographic location of California station and in the right, the time series of annual maximum rainfall.

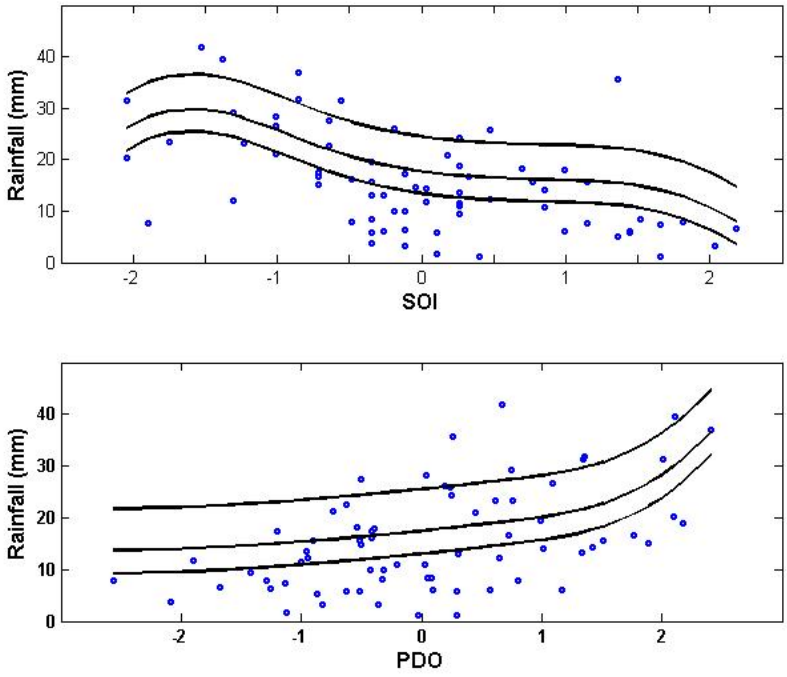

Figure 3.6 – $\tau = 0.5$, 0.7 **and** 0.9 **Conditional quantiles using GEV-B-Spline model-case study 2**

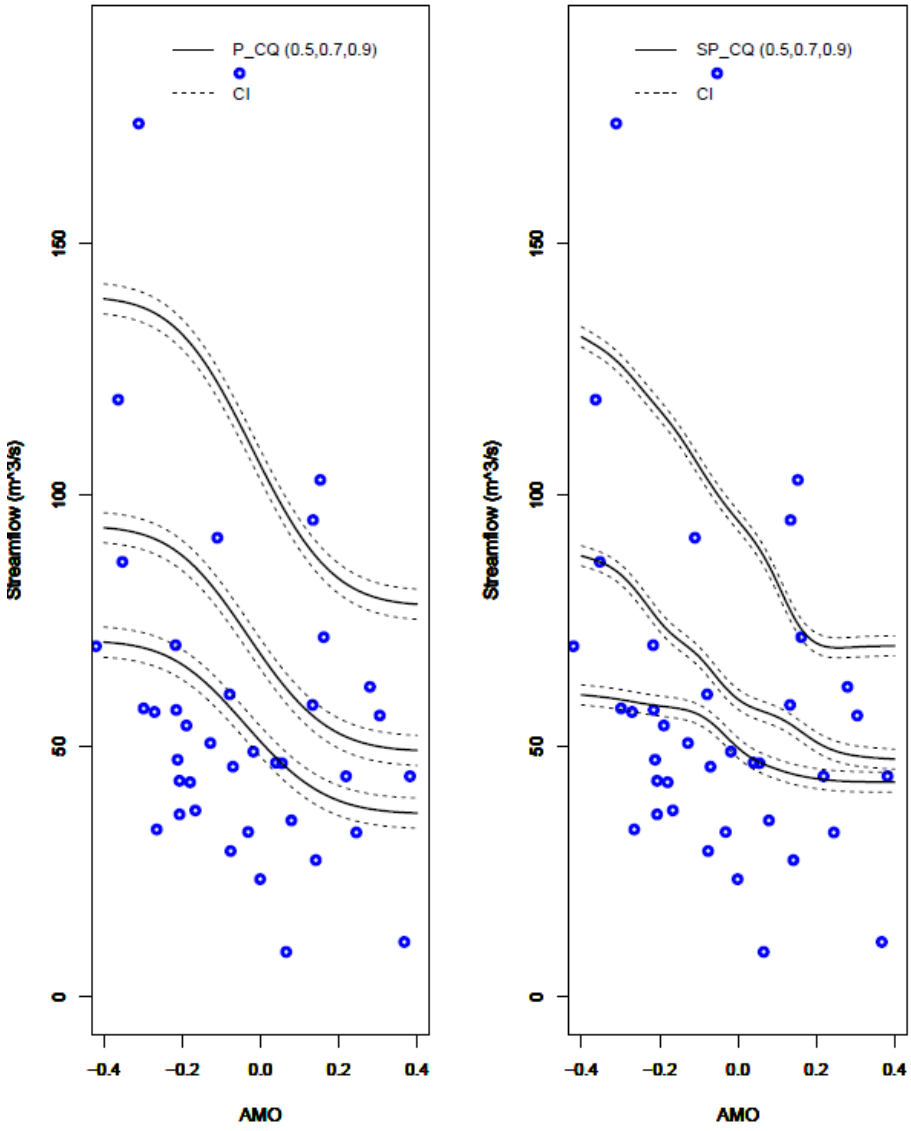

Figure 3.7 – $\tau = 0.5, 0.7$ and 0.9 **Conditional quantiles (P__CQ for the parametric estimator and SP__CQ for the semiparametric estimator) and their respective confidence intervals (CI)-case study 1**

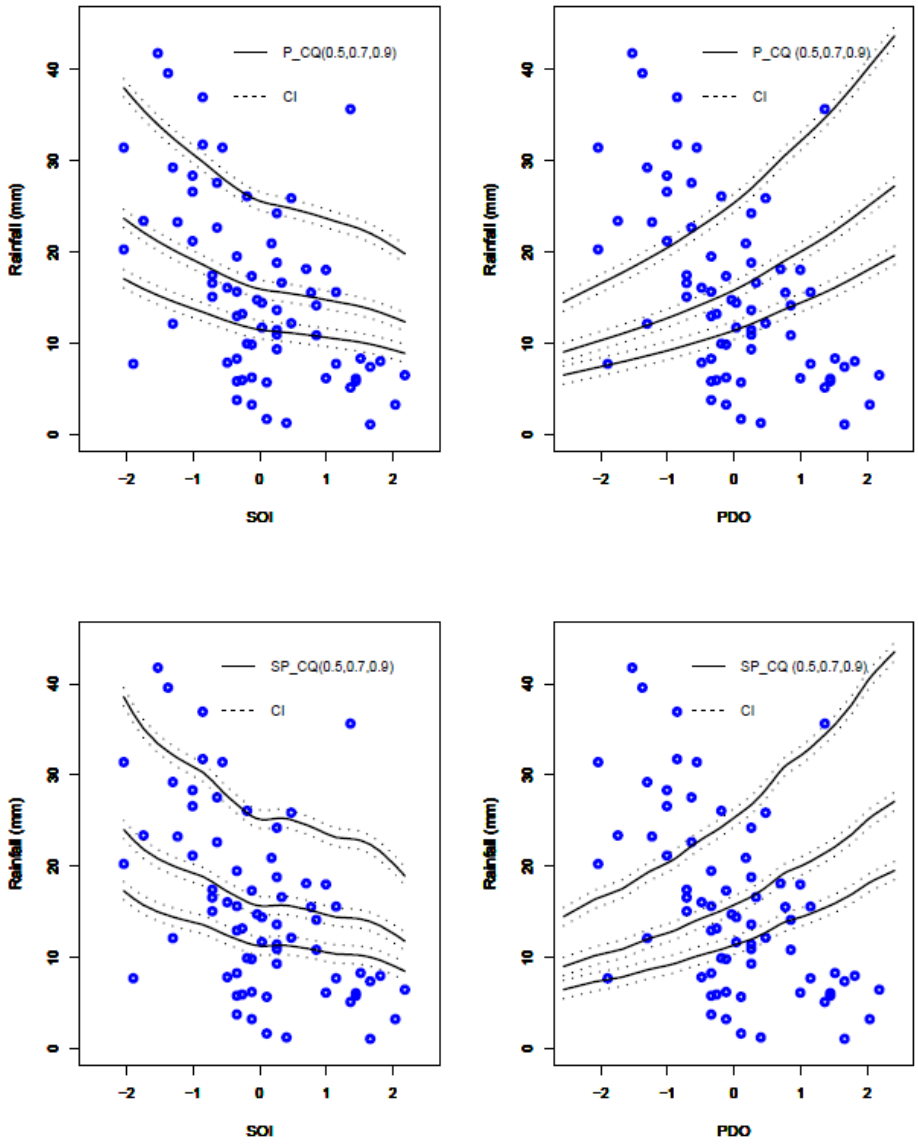

Figure 3.8 – $\tau = 0.5, 0.7$ **and** 0.9 **Conditional quantiles (P__CQ for the parametric estimator and SP__CQ for the semiparametric estimator) and their respective confidence intervals (CI)-case study 2**

Tables

Table 3.1 – 100*$IMSE$ results calculated for the proposed estimators and the competitors estimators.

		DGP s.a					**DGP s.b**					**DGP s.c**					**DGP s.d**					**DGP s.e**				
n	τ	$\widehat{Q}_{CDF}$	$\widehat{Q}_{Lr}$	$\widehat{Q}_{nLr}$	$\widehat{Q}_p$	$\widehat{Q}_{sp}$	$\widehat{Q}_{CDF}$	$\widehat{Q}_{Lr}$	$\widehat{Q}_{nLr}$	$\widehat{Q}_p$	$\widehat{Q}_{sp}$	$\widehat{Q}_{CDF}$	$\widehat{Q}_{Lr}$	$\widehat{Q}_{nLr}$	$\widehat{Q}_p$	$\widehat{Q}_{sp}$	$\widehat{Q}_{CDF}$	$\widehat{Q}_{Lr}$	$\widehat{Q}_{nLr}$	$\widehat{Q}_p$	$\widehat{Q}_{sp}$	$\widehat{Q}_{CDF}$	$\widehat{Q}_{Lr}$	$\widehat{Q}_{nLr}$	$\widehat{Q}_p$	$\widehat{Q}_{sp}$
50	0.20	0.02	0.02	0.01	**0.00**	**0.01**	1.01	1.14	0.92	**0.31**	**0.71**	0.014	0.016	0.013	**0.009**	**0.009**	0.010	0.011	0.009	**0.006**	**0.003**	0.018	0.021	0.017	**0.011**	**0.005**
	0.50	0.42	0.46	0.06	**0.06**	**0.06**	1.61	1.92	1.45	**0.35**	**0.85**	1.298	1.473	0.932	**0.113**	**0.521**	0.813	0.982	0.622	**0.075**	**0.347**	1.721	1.866	1.181	**0.143**	**0.660**
	0.80	0.59	0.67	0.53	**0.11**	**0.13**	2.85	3.29	1.72	**0.42**	**1.38**	2.013	2.154	1.963	**0.018**	**1.275**	1.321	1.436	1.309	**0.012**	**0.850**	2.612	2.728	2.487	**0.023**	**1.615**
100	0.20	0.02	0.02	0.01	**0.00**	**0.03**	0.35	0.41	0.33	**0.11**	**0.25**	0.012	0.014	0.012	**0.003**	**0.004**	0.007	0.008	0.006	**0.004**	**0.002**	0.014	0.016	0.013	**0.009**	**0.004**
	0.50	0.81	0.73	0.09	**0.07**	**0.07**	0.60	0.69	0.52	**0.13**	**0.31**	1.161	1.277	0.808	**0.098**	**0.451**	0.521	0.752	0.476	**0.058**	**0.266**	1.231	1.473	0.932	**0.113**	**0.521**
	0.80	1.51	1.34	1.07	**0.19**	**0.21**	0.85	1.18	0.62	**0.15**	**0.49**	1.521	1.867	1.702	**0.016**	**1.105**	1.021	1.100	1.002	**0.009**	**0.65**	2.031	2.154	1.963	**0.018**	**1.275**
200	0.2	0.02	0.02	0.01	**0.01**	**0.03**	0.29	0.19	0.15	**0.05**	**0.12**	0.011	0.013	0.011	**0.003**	**0.003**	0.005	0.006	0.005	**0.003**	**0.002**	0.01	0.02	0.01	**0.01**	**0.01**
	0.5	0.35	0.30	0.10	**0.02**	**0.09**	0.34	0.31	0.24	**0.06**	**0.14**	1.087	1.179	0.746	**0.090**	**0.417**	0.521	0.611	0.387	**0.047**	**0.216**	1.61	1.83	0.87	**0.11**	**0.49**
	0.8	2.50	2.08	1.19	**0.15**	**0.25**	0.35	0.54	0.28	**0.07**	**0.23**	1.621	1.723	1.571	**0.014**	**1.020**	0.721	0.893	0.814	**0.007**	**0.529**	1.91	2.01	1.83	**0.01**	**1.19**

Table 3.2 – P-value results of Goodness-of-fit tests $S_n(B)$ and $S_n(C)$ based on Rosemblatt's transformation. Here, the selected copula model is the copula with the higher p-value greater than 0.05 and is indicated in bold with its corresponding parameter.

case study 1					
Tests	Normal	Student	Frank	Gumbel	Clayton
$S_n(B)$	0.05 -	0.06 -	**0.82** **(-1.92)**	0.38 -	0.44 -
$S_n(C)$	0.05 -	0.07 -	**0.83** **(-1.92)**	0.39 -	0.43 -
case study 2					
Tests	Normal	Student	Frank	Gumbel	Clayton
$S_n(B)$	**0.94** $\mathbf{R} = \begin{pmatrix} \mathbf{1} & \mathbf{-0.52} & \mathbf{-0.47} \\ \mathbf{-0.52} & \mathbf{1} & \mathbf{-0.44} \\ \mathbf{-0.47} & \mathbf{-0.44} & \mathbf{1} \end{pmatrix}$	0.61 -	0.21 -	0.06 -	0.03 -
$S_n(C)$	**0.71** $\mathbf{R} = \begin{pmatrix} \mathbf{1} & \mathbf{-0.52} & \mathbf{-0.47} \\ \mathbf{-0.52} & \mathbf{1} & \mathbf{-0.44} \\ \mathbf{-0.47} & \mathbf{-0.44} & \mathbf{1} \end{pmatrix}$	0.51 -	0.13 -	0.05 -	0.01 -

Table 3.3 – Results of Goodness-of-fit based on BIC criterion. "NA" indicates that the probability distribution cannot be fitted for the data. Here, the best selected probability distribution is the distribution with the lowest value of BIC and is indicated in bold with its corresponding parameter.

case study 1					
Variables	Weibull	Gamma	Normal	GEV	Lognormal
MAS	627.30	629.44	628.01	627.12	**626.96**
	-	-	-	-	**(3.91,0.58)**
AMO	NA	NA	**-4.76**	-3.16	NA
	-	-	**(-0.04,0.21)**	-	-
case study 2					
Variables	Weibull	Gamma	Normal	GEV	Lognormal
MAR	628.21	629.37	628.12	**626.52**	628.96
	-	-	-	**(10.9, 7.48,0.03)**	-
SOI	NA	NA	**-5.15**	-4.28	NA
	-	-	**(-1.67,11.04)**	-	-
PDO	NA	NA	**-3.12**	-2.95	NA
	-	-	**(0.07,1.09)**	-	-

Troisième partie

Annexe

Annexe

0.1 Phénomène de Runge

Le phénomène de Runge se manifeste dans le contexte de l'interpolation polynomiale. Avec certaines fonctions (même infiniment dérivables), l'augmentation du nombre de points d'interpolation ne constitue pas nécessairement une bonne stratégie d'approximation.

Prenons l'exemple de la fonction suivante:

$$f(x) = \frac{1}{1+25x^2}, \qquad x\,[-1,1]\,.$$

Runge en 1901 a montré que si cette fonction est interpolée aux points équidistants, x_k entre -1 et 1: $x_k = -1 + (k-1)\,\frac{2}{n}, \quad k = 0 \ldots n$, par une polynôme P_n de degré $\leq n$ alors:

$$\lim_{n\to\infty} \left(\max_{-1\leq x\leq 1} |f(x) - P_n(x)| \right) = \infty.$$

Lorsqu'on augmente le nombre de points, on constate que le polynôme se met à osciller fortement entre les points x_k avec une amplitude de plus en plus grande. La figure 1 montre une illustration graphique du phénomène de Runge. Nous remarquons que l'approximation est de plus en plus mauvaise quand n est plus grand.

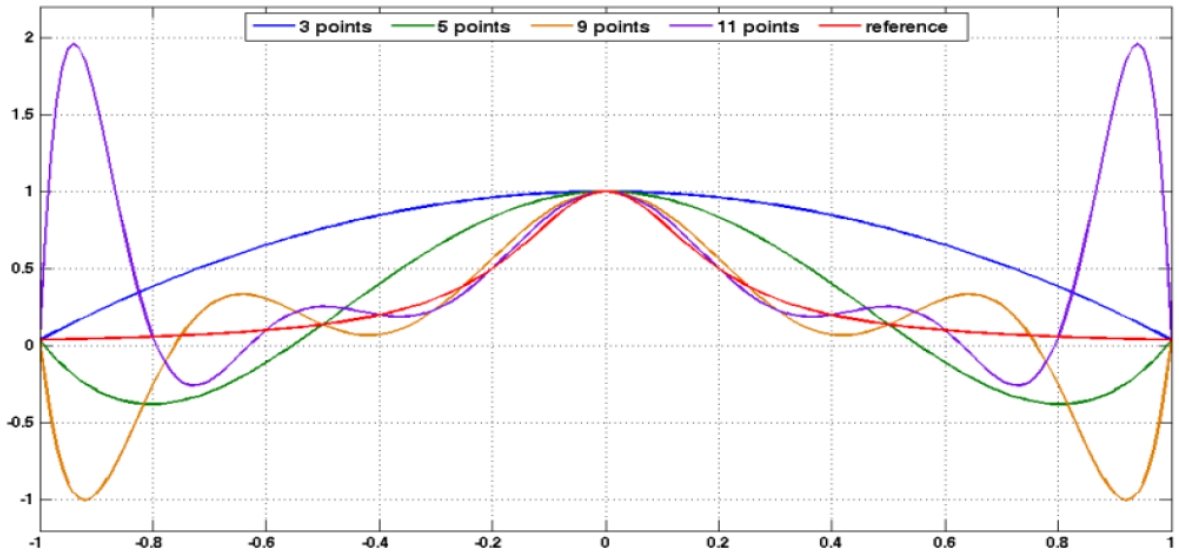

Figure 1 – Illustration graphique du phénomène de Runge. La courbe rouge représente la fonction $\frac{1}{1+25x^2}, x\,[-1,1]$ et les autres courbes sont les courbes d'interpolation

0.2 MCMC algorithm for GEV B-Splines model

The basic idea of the MCMC method is, for each parameter, to construct a Markov chain with the posterior distribution being a stationary and ergodic distribution. After running the Markov chain, of size N , for a given burn-in period N_0 , one obtains a sample from the posterior distribution $f(\boldsymbol{\theta}|y)$. One popular method for constructing a Markov chain is via the Metropolis-Hastings (M-H) algorithm [Metropolis *et al.*, 1953; Hastings, 1970]. We simulated the realizations from the posterior distribution by way of a single-component M-H algorithm [Grehys, 1996]. Each parameter was updated using a random-walk Metropolis algorithm with a Gaussian proposal density centered at the current state of the chain. Some methods to assess the convergence of the MCMC methods make it possible to determine the length of the chain and the burn-in time such as the Raftery and Lewis diagnostic [Raftery & Lewis, 1992, 1995] and subsampling methods [El Adlouni *et al.*, 2006]. In all cases, the convergence methods indicated that the Markov chains converged within a few iterations.

In this study, we considered chains of size $N = 15000$ and a burn-in period of $N_0 = 8000$ runs. In every case, a sample of $N - N_0 = 7000$ values is collected from the posterior of each of the elements of $\boldsymbol{\theta}$. The principal step of the M-H algorithm can be summarized as follows:

— Initialization: Assign initial value $\boldsymbol{\theta}^0$ and choose an arbitrary proposal probability density $Q(\theta^*|\theta)$. In this case we propose a multivariate normal distribution.

— For each iteration t: generate θ^* a candidate for the next sample by picking from the distribution $Q(\theta^*|\theta_t)$.

— Calculate the acceptance ratio, given by $\alpha(\theta^*, \theta_t) = \frac{\pi(\theta^*|y)}{\pi(\theta_t|y)}$.

— If $\alpha \geq 1$, then the candidate is more likely than θ_t; automatically accept the candidate by setting $\theta_{t+1} = \theta^*$. Otherwise, accept the candidate with probability α; if the candidate is rejected, set $\theta_{t+1} = \theta_t$ instead.

0.3 Loss function and Laplace distribution

The pth linear regression quantile $(0 < p < 1)$ is defined as any solution, $\widehat{\boldsymbol{\alpha}(p)}, \widehat{\alpha_0(p)}$, to the quantile regression minimization problem

$$\arg \min_{\boldsymbol{\alpha}, \alpha_0 \in \mathbb{R}} \sum_{i=1}^{n} \rho_p \left(y_i - \boldsymbol{x_i'}\boldsymbol{\alpha} - \alpha_0 \right)$$

where $\rho_p(z)$ is a loss function defined as:

$$\rho_p(z) = \frac{|z| + (2p-1)z}{2}.$$

Minimization of the loss function $\rho_p(z)$ is equivalent to the maximization of a likelihood function formed by combining independently distributed asymmetric Laplace densities

$$f(y, x_{,\alpha}, \alpha_0, p) = p(p-1)\exp\{-\rho_p(y - \boldsymbol{x'}\boldsymbol{\alpha} - \alpha_0)\}$$

Indeed,

$$\arg\min_{\alpha,\alpha_0\in\mathbb{R}} \sum_{i=1}^{n} \rho_p(y_i - \boldsymbol{x_i'}\boldsymbol{\alpha} - \alpha_0) \text{ is equivalent to } \quad p(p-1)\arg\max_{\alpha,\alpha_0\in\mathbb{R}} \exp\{-\rho_p(y - \boldsymbol{x'}\boldsymbol{\alpha} - \alpha_0)\}$$

This equivalence is due simply to the fact that the exponential function is strictly increasing.

0.4 Climate Indices

North Atlantic Oscillation (NAO)

NAO is an irregular fluctuation of atmospheric pressure over the North Atlantic Ocean that has a strong effect on winter weather in Europe, northeastern North America, North Africa, and northern Asia [Hurrell & Van Loon, 1997].

El Nino Southern Oscillation (ENSO)

ENSO is a naturally occurring phenomenon that involves fluctuating ocean temperatures in the equatorial Pacific. For North America and much of the globe, the phenomenon is known as a dominant force causing variations in regional climate patterns [Bjerknes, 1969].

Pacific Decadal Oscillation (PDO)

PDO is a pattern of Pacific climate variability similar to ENSO in character, but which varies over a much longer time scale. The PDO can remain in the same phase for 20 to 30 years, while ENSO cycles typically only last 6 to 18 months [Nathan & Hare, 2002].

Atlantic Multi-decadal Oscillation (AMO)

AMO is a fluctuation in the sea surface temperature in the North Atlantic Ocean. It seems to occur with a period of roughly 70 years [Teegavarapu *et al.*, 2013].

0.5 Copula: definition, properties and Sklar's theorem

Definition

Copula is a function which joins the marginal distribution functions to form a multivariate joined distribution function. [Meylan *et al.*, 2012] gave the definition of the multivariate copula function. A function $C : [0,1]^d \to [0,1]$ is a d-dimensional Copula, when the following conditions are satisfied.

(i) $C(u_1, \ldots, u_{i-1}, 0, u_{i+1}, \ldots, u_d) = 0$; the copula is zero if one of the argument is zero.

(ii) $C(1, \ldots, 1, u, 1, \ldots, 1) = u$; the copula is equal to u if one argument is u and all others equal to 1.

(iii) C is d-increasing; i.e.,

$$\frac{\partial^d C}{\partial u_1 \ldots \partial u_d} \geq 0.$$

C is, in fact, a multivariate distribution with a uniform marginal distributions.

For instance, in the bivariate case (i.e. $d = 2$), $C : [0,1] \to [0,1] \times [0,1]$ is a bivariate copula if:

(i) $C(0, u_2) = C(u_1, 0) = 0$.

(ii) $C(1, u_2) = u_2$ and $C(u_1, 1) = u_1$.

(iii) $C(u_2, v_2) - C(u_2, v_1) - C(u_1, v_2) + C(u_1, v_1) \geq 0$; $\forall\, 0 \leq u_1 \leq u_2 \leq 1$ and $0 \leq v_1 \leq v_2 \leq 1$.

Sklar's theorem

For a given $\mathbf{x} = (x_1, \ldots, x_d)^\top$, from the seminal work of [Nelsen, 2006], the c.d.f of $(Y, \boldsymbol{X})$ evaluated at $(y, \boldsymbol{x})$ can be expressed by $C(F_0(y), \boldsymbol{F(x)})$, where $\boldsymbol{F}(\boldsymbol{x}) = (F_1(x_1), \ldots, F_d(x_d))$ and C is the copula function of $(Y, \boldsymbol{X})$, defined by $C(u_0, u_1, \ldots, u_d) = P(U_0 \leq u_0, U_1 \leq u_1, \ldots, U_d \leq u_d)$.

Examples of copula families

— **Archimedean copulas**: Archimedean copulas are an associative class of copulas. Most common Archimedean copulas admit an explicit formula. In practice, Archimedean copulas are popular because they allow modeling the dependence in arbitrarily high dimensions with only one parameter, governing the strength of dependence. A copula C is called Archimedean if it admits the representation

$$C(u_0, u_1, \ldots, u_d; \theta) = \varphi^{[-1]}(\varphi(u_0; \theta) + \varphi(u_1; \theta) + \ldots + \varphi(u_d; \theta); \theta),$$

where $\varphi \colon [0,1] \times \Theta \to [0, \infty)$ is a continuous, strictly decreasing and convex function such that $\varphi(1; \theta) = 0$. θ is a parameter within some parameter space Θ. φ is the so-called generator function and φ^{-1} is its pseudo-inverse. Here, we give the generator function for some important Archimedean copula:

(a) Clayton: $\varphi(t) = \theta^{-1}\left(t^\theta - 1\right), \theta \in [-1, +\infty[\setminus 0$.

(b) Gumbel: $\varphi = (-\log(t))^{\theta}, \theta \in [1, +\infty[$.

(c) Frank: $\varphi(t) = -\log\left(\frac{\exp(-\theta t)-1}{\exp(-\theta)-1}\right), \theta \in]-\infty, +\infty[\setminus 0$.

The Clayton copula is mostly used to study correlated risks because of its ability to capture lower tail dependence. The Gumbel copula is used to model asymmetric dependence in the data and it is famous for its ability to capture strong upper tail dependence and weak lower tail dependence. Unlike the Clayton and the Gumbel copula, the Frank copula allows the maximum range of dependence which means that the dependence parameter of the Frank copula permits modeling positive as negative dependence.

— **Elliptical copulas**: The elliptical copulas differ from the Archimedean classes of copulas in the approach that only implicit analytical expression is available. These copulas are derived from the related elliptical distribution (e.g. normal distribution, Student t-distribution).

(a) Gaussian: the Gaussian copula is a distribution over the unit cube $[0,1]^d$. It is constructed from a multivariate normal distribution over $\mathbb{R}^d$ by using the probability integral transform. For a given correlation matrix $R \in [-1,1]^{d \times d}$, the Gaussian copula with parameter matrix R can be written as:

$$C(u_0, u_1, \ldots, u_d) = \Phi_R\left(\Phi^{-1}(u_0), \Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_d)\right),$$

where Φ^{-1} is the inverse cumulative distribution function of a standard normal and Φ_R is the joint cumulative distribution function of a multivariate normal distribution with mean vector zero and covariance matrix equal to the correlation matrix R.

(b) Student: the Student copula is a distribution over the unit cube $[0,1]^d$. It is constructed from a multivariate Student-t distribution over $\mathbb{R}^d$. For a given correlation matrix $R \in [-1,1]^{d \times d}$, and d degrees of freedom. The Student copula with parameter matrix R and

d can be written as:

$$C\left(u_0, u_1, \ldots, u_d\right) = \boldsymbol{t}_{d,R}\left(t_d^{-1}\left(u_0\right), \ldots, t_d^{-1}\left(u_d\right)\right),$$

where d is the degree of freedom parameter, t_d^{-1} is the inverse of the univariate standard Student-t distribution function, and $\boldsymbol{t}_{d,R}$ is the multivariate standard Student-t distribution parameterized by the correlation matrix R and d is degrees of freedom.

0.6 Proofs-Article 3

0.6.1 Proof of Theorem 1-Article 3

Proof 1 *To provide the asymptotic i.i.d representation of the conditional quantile parametric estimator, we apply 1-time differentiable Taylor's theorem to* $H_{\boldsymbol{x}}^p\left(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}; \widehat{\boldsymbol{\theta}}, \tau\right)$ *around* $H_{\boldsymbol{x}}^p\left(\boldsymbol{\alpha}, \boldsymbol{\beta}; \boldsymbol{\theta}, \tau\right)$ *:*

$$\begin{aligned} H_{\boldsymbol{x}}^p\left(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\theta}}, \tau\right) &\approx H_{\boldsymbol{x}}^p\left(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\theta}, \tau\right) + H_{\boldsymbol{x},1}^p\left(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\theta}, \tau\right)\left(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}\right) + H_{\boldsymbol{x},2}^p\left(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\theta}, \tau\right)\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \\ &\quad + H_{\boldsymbol{x},3}^p\left(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\theta}, \tau\right)\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) \\ &= H_{\boldsymbol{x}}^p\left(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\theta}, \tau\right) + \left(\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}\right)\nabla H_{\boldsymbol{x}}^p\left(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\theta}, \tau\right). \end{aligned}$$

Now, Bentzien & Friederichs [2007] have showed that under general regularity conditions, the asymptotic i.i.d representation for the IMF estimator can be written as:

$$\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\gamma_i} + o_p\left(n^{-1/2}\right),$$

where $\boldsymbol{\gamma}_i$ are i.i.d random vectors, depending on $(Y_i, \boldsymbol{X_i}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\theta}, \tau, \boldsymbol{x})$, with zero mean 0 and finite variance given by

$$E\left(\gamma_i^2\right) = \mathcal{G}(\boldsymbol{\eta}),$$

where $\mathcal{G}(\boldsymbol{\eta})$ is the information matrix of Godambe [Godambe, 1960]. If we define $g(\boldsymbol{\eta}) = \left(\frac{\partial l_1}{\partial \boldsymbol{\alpha}}, \frac{\partial l_2}{\partial \boldsymbol{\beta}}\right)$, where l_1 and l_2 are the likelihood function for Y and $\boldsymbol{X}$. Then, the Godombe matrix can be written as:

$$\mathcal{G}(\boldsymbol{\eta}) = \left(D^{-1}\right) M \left(D^{-1}\right)^T,$$

where $D = E\left[\frac{\partial}{\partial \boldsymbol{\eta}}\left(g(\boldsymbol{\eta})^T\right)\right]$ and $M = \left[g(\boldsymbol{\eta})^T g(\boldsymbol{\eta})\right]$. Which concludes the proof.

■

0.6.2 Proof of of Theorem 2- Article 3

Proof 2 *we will follow similar steps as in Section 0.6.1 to show the asymptotic i.i.d. representation of the conditional quantile semiparametric estimator.*

In fact, we apply the Taylor's theorem to $H_{\boldsymbol{x}}^{sp}\left(\widehat{F_0}, \widehat{\boldsymbol{F}}; \widehat{\boldsymbol{\theta}}, \tau\right)$ around $H_{\boldsymbol{x}}^{sp}(F_0, \boldsymbol{F}, \boldsymbol{\theta}, \tau)$, which gives the following results:

$$\begin{aligned} H_{\boldsymbol{x}}^{sp}\left(\widehat{F_0}, \widehat{\boldsymbol{F}}, \widehat{\boldsymbol{\theta}}, \tau\right) - H_{\boldsymbol{x}}^{sp}(F_0, \boldsymbol{F}, \boldsymbol{\theta}, \tau) &\approx H_{\boldsymbol{x},1}^{sp}(F_0, \boldsymbol{F}, \boldsymbol{\theta}, \tau)\left(\widehat{F_0} - F_0\right) + H_{\boldsymbol{x},2}^{sp}(F_0, \boldsymbol{F}, \boldsymbol{\theta}, \tau)\left(\widehat{\boldsymbol{F}} - \boldsymbol{F}\right) \\ &\quad + H_{\boldsymbol{x},3}^{sp}(F_0, \boldsymbol{F}, \boldsymbol{\theta}, \tau))\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) \\ &= (\widehat{\boldsymbol{\kappa}} - \boldsymbol{\kappa}) \nabla H_{\boldsymbol{x}}^{sp}(F_0, \boldsymbol{F}, \boldsymbol{\theta}, \tau). \end{aligned}$$

According to Noh et al. *[2013], we have:* $\widehat{\boldsymbol{\kappa}} - \boldsymbol{\kappa} \approx \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\gamma}_i$*, where* $\boldsymbol{\gamma}_i$ *is a i.i.d random vector such that* $E(\boldsymbol{\gamma}) = \mathbf{0}$ *and* $E(\boldsymbol{\gamma}^2) < \infty$*.*

Therefore,

$$H_{\boldsymbol{x}}^{sp}\left(\widehat{F_0}, \widehat{\boldsymbol{F}}, \widehat{\boldsymbol{\theta}}, \tau\right) - H_{\boldsymbol{x}}^{sp}(F_0, \boldsymbol{F}, \boldsymbol{\theta}, \tau) \approx \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\gamma}_i \nabla H_{\boldsymbol{x}}^{sp}(F_0, \boldsymbol{F}, \boldsymbol{\theta}, \tau).$$

Finally, using limit central theorem we can deduce the asymptotic normality of the proposed estimator.

Bibliographie

Abi-Zeid I & Bobée B (1999). La modélisation stochastique des étiages: une revue bibliographique. *Revue des sciences de l'eau*, 12(3):459–484.

Aissaoui-Fqayeh I, El-Adlouni S, Ouarda TBMJ & St-Hilaire A (2009). Non-stationary lognormal model development and comparison with the non-stationary GEV model. *Hydrological Sciences Journal*, 54(6):1141–1156.

Anctil F, Rousselle J & Lauzon N (2012). *Hydrologie. Cheminements de l'eau.* Montreal, Presses internationales polytechniques.

Argence S, Lambert D, Richard E, Chaboureau JP & Söhne N (2008). Impact of initial condition uncertainties on the predictability of heavy rainfall in the Mediterranean: a case study. *Quarterly Journal of the Royal Meteorological Society*, 134(636):1775–1788.

Ashkar F & Ouarda T (1996). On some methods of fitting the generalized pareto distribution. *Journal of Hydrology*, 177:117–141.

Bates B, Kundzewicz Z, Wu S & Palutikof J (2008). Climate change and water. *Technical Paper of the Intergovernmental Panel on Climate Change.Geneva: IPCC Secretariat.*

Beguería S, Angulo-Martínez M, Vicente-Serrano SM, López-Moreno JI & El-Kenawy A (2011). Assessing trends in extreme precipitation events intensity and magnitude using non-stationary peaks-over-threshold analysis: a case study in northeast spain from 1930 to 2006. *International Journal of Climatology*, 31(14):2102–2114.

Bentzien S & Friederichs P (2007). Statistical downscaling of extreme precipitation events using censored quantile regression. *Monthly weather review*, 135:2365–2378.

Bjerknes J (1969). Atmospheric teleconnections from the equatorial pacific. *Monthly weather review*, 97:163–172.

Bliefernicht J & Bárdossy A (2007). Probabilistic forecast of daily areal precipitation focusing on extreme events. *Natural Hazards and Earth System Sciences*, 7(2):263–269. DOI:10.5194/nhess-7-263-2007.

Bouyé E & Salmon M (2002). Dynamic copula quantile regression and tail area dynamic dependence in forex markets. *Manuscript, Financial Econometrics Research Centre, Warwick Business School, UK*.

Brabets TP & Walvoord MA (2009). Trends in streamflow in the Yukon River basin from 1944 to 2005 and the influence of the pacific decadal oscillation. *Journal of Hydrology*, 371:108–119.

Breymann W, Dias A & Embrechts P (2003). Dependence structures for multivariate high-frequency data in finance. *Quantitative Finance*, 3:1–14.

Buchinsky M (1998). The dynamics of changes in the female wage distribution in the USA: a quantile regression approach. *Journal of Applied Econometrics*, 13(1):1–30. DOI:10.1002/(SICI)1099-1255(199801/02)13:1<1::AID-JAE474>3.0.CO;2-A.

Buishand TA (1984). Bivariate extreme value data and the stationyear method. *Journal of Hydrology*, 69:77–95.

Buishand TA (1989). Statistics of extremes in climatology. *Statistica Neerlandica*, 36:1–30.

Buishand TA (1991). Extreme rainfall estimation by combining data from several sites. *Hydrological Science Journal*, 36:345–365.

Cannon A (2010). A flexible nonlinear modelling framework for nonstationary generalized extreme value analysis in hydroclimatology. *Hydrological Processes*, 24(6):673–685.

Cannon A (2011). Quantile regression neural networks:implementation in R and application to precipitation downscaling. *Computers and Geosciences*, 37:1277–1284.

Carter DJT & Challenor PG (1981). Estimating return values of environmental parameters. *Quarterly Journal of the Royal Meteorological Society*, 101:259–266.

Cavazos T & Hewitson BC (2005). Performance of ncep–ncar reanalysis variables in statistical downscaling of daily precipitation. *Climate Ressources*, 28:95–107. DOI:doi:10.3354/cr028095.

Chandran A, Basha G & Ouarda TBMJ (2016). Influence of climate oscillations on temperature and precipitation over the United Arab Emirates. *International Journal of Climatology*, 36(1):225–235. DOI:10.1002/joc.4339.

Chavez-Demoulin V & Davison A (2005). Generalized additive modeling of sample extremes. *Applied Statistics*, 54:207–222.

Christopeit N (1994). Estimating parameters of an extreme value distribution by the method of moments. *Journal of Statistical Planning and Inference*, 42:173–186.

Clayton DG (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika*, 65:141–151.

Coles S (2001). *An Introduction to Statistical Modeling of Extreme Values*. Springer Series in Statistics.

Conte M, Giuffrida A & Tedesco S (1989). The Mediterranean oscillation: Impact on precipitation and hydrology in Italy. *Proc. Conf. on Climate, Water, Helsinki, Finland, Academy of Finland*, 121–137 pages.

Cunderlik JM & Burn DH (2002). Local and regional trends in monthly maximum flows in southern British Columbia. *Canadian Water Resources Journal*, 27:191–212.

Cunderlik JM & Ouarda TBMJ (2009). Trends in the timing and magnitude of floods in Canada. *Journal of hydrology*, 375:471–480.

Cunnane C (1989). Statistical distributions for flood frequency analysis. *WMO No. 718, WMP, Geneva*.

De Boor C (2001). *A practical guide to spline*. Springer Series in Statistics.

Di Baldassarre G, Montanari A, Lins H, Koutsoyiannis D, Brandimarte L & Blöschl G (2010). Flood fatalities in Africa: From diagnosis to mitigation. *Geophysical Research Letters*, 37(22):n/a–n/a. DOI:10.1029/2010GL045467. L22402.

Dobric J & Schmid F (2005). Testing goodness of fit for parametric families of copulas: Application to financial data. *Communications in Statistics.Simulation and Computation*, 34:1053–1068.

Donat MG, Peterson TC, Brunet M, King AD, Almazroui M, Kolli RK, Boucherf D, Al-Mulla AY, Nour AY, Aly AA, Nada TAA, Semawi MM, Al Dashti HA, Salhab TG, El Fadli KI, Muftah MK, Dah Eida S, Badi W, Driouech F, El Rhaz K, Abubaker MJY, Ghulam AS, Erayah AS, Mansour MB, Alabdouli WO, Al Dhanhani JS & Al Shekaili MN (2014). Changes in extreme temperature and precipitation in the arab region: long-term trends and variability related to ENSO and NAO. *International Journal of Climatology*, 34(3):581–592. DOI:10.1002/joc.3707.

Donner RV, Ehrcke R, Barbosa SM, Wagner J, Donges JF & Kurths J (2012). Spatial patterns of linear and nonparametric longterm trends in baltic sealevel variability. *Nonlinear Processes in Geophysics*, 19:9511.

Déry SJ & Wood EF (2005). Decreasing river in northern Canada. *Geophysical Research Letters*, 32.

Ehsanzadeh E & Adamowski K (2007). Detection of trends in low flows across Canada. *Canadian Water Resources Journal*, 32(4):251–264.

Ehsanzadeh E, Saley H, Ouarda TB, Burn D, Pietroniro A, Seidou O, Charron C & Lee D (2007). Analysis of changes in the great lakes hydro-climatic variables. *Journal of Great Lakes Research*, 39(3):383–394.

El Adlouni S, Favre A & Bobée B (2006). Comparison of methodologies to assess the convergence of Markov Chain Monte Carlo methods. *Computational Statistics and Data Analysis*, 50(10): 2685–2701.

El Adlouni S, Ouarda T, Zhang X, Roy R & Bobée B (2007). Generalized maximum likelihood estimators for the nonstationary generalized extreme value model. *Water Resources Research*, 43(W03410).

El Adlouni S & Ouarda TB (2009). Joint bayesian model selection and parameter estimation of the generalized extreme value model with covariates using birth-death Markov Chain Monte Carlo. *Water Resources Research*, 45(W06403).

Fermanian JD (2005). Goodness-of-fit tests for copulas. *Journal of Multivariate Analysis*, 95:119–152.

Fiala T, Ouarda T & Hladny J (2010). Evolution of low flows in the Czech Republic. *Journal of Hydrology*, 393:206–218.

Fisher R & Tippett L (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample. *Proceedings of the Cambridge Philosophical Society*, 24:180–190.

Fowler HJ, Blenkinsop S & Tebaldi C (2007). Linking climate change modelling to impacts studies: recent advances in downscaling techniques for hydrological modelling. *International Journal of Climatology*, 27(12):1547–1578. DOI:10.1002/joc.1556.

Fox J (2000). Multiple and generalized nonparametric regression. series: Quantitative applications in the social sciences. *SAGE Publications Inc*, 131.

Frank MJ (1960). Distributions des valeurs extrêmes en plusieurs dimensions. *Publications de l'Institut de Statistique de l'Université de Paris*, 9:171–173.

Frank MJ (1979). On the simultaneous associativity of $f(x, y)$ and $x + y - f(x, y)$. *Aequationes Mathematicae*, 19:194–226.

Friederichs P (2010). Statistical downscaling of extreme precipitation events using extreme value theory. *Extremes*, 13(2):109–132.

Friederichs P & Hense A (2007). Statistical downscaling of extreme precipitation events using censored quantile regression. *Monthly Weather Review*, 135:2365–2378.

Friederichs P & Hense A (2008). A probabilistic forecast approach for daily precipitation totals. *Weather and Forecasting*, 23(4):659–673.

Gelman A, Carlin, B. J, S. SH & Rubin DB (1995). Bayesian data analysis. *London: Chapman and Hall*.

Genest C, Ghoudi K & Rivest L (1995). A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. *Biometrika*, 82:543–52.

Genest C, Quessy JF & Remillard B (2006). Goodness-of-fit procedures for copula models based on the integral probability transformation. *Scandinavian Journal of Statistics*, 33:337–366.

Genest C, Remillard B & Beaudoin D (2009). Goodness-of-fit tests for copulas: A review and a power study. *Insurance: Mathematics and Economics*, 44:199–213.

GIEC (2007). Climate Change 2007. The Physical Science Basis. Summary for Policymakers, Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change. *Report of the Intergovernmental Panel on Climate Change*.

Gilks WR, Richardson S & Spiegelhalter DJ (1996). *Markov Chain Monte Carlo in Practice*. London: Chapman and Hall.

Godambe V (1960). An optimum property of regular maximum likelihood estimation. *The Annals of Mathematical Statistics*, 31:1208–1211.

Green P & Silverman B (1994). *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*. Chapman AND Hall.

Grehys (1996). Presentation and review of some methods for regional flood frequency analysis. *Journal of Hydrology*, 186:63–84.

Hastings WK (1970). Monte Carlo sampling methods using Markov Chains and their applications. *Biometrika*, 57(1):97–109.

Hendricks W & Koenker R (1992). Hierarchical spline models for conditional quantiles and the demand for electricity. *Journal of the American Statistical Association*, 87(417):58–68.

Hertig E, Paxian A, Vogt G, Seubert S, Paeth H & Jacobeit J (2012, pages =). Statistical and dynamical downscaling assessments of precipitation extremes in the mediterranean area. *Meteorologische Zeitschrift*, 21(1).

Hertig E, Seubert S, Paxian A, Vogt G, Paeth H & Jacobeit J (2013). Changes of total versus extreme precipitation and dry periods until the end of the twenty-first century: statistical assessments for the mediterranean area. *Theoretical and Applied Climatology*, 111(1):1–20. DOI:10.1007/s00704-012-0639-5.

Hill B (1975). A simple general approach to inference about the tail of a distribution. *The Annals of Statistics*, 3:1163–1173.

Hirabayashi Y, Mahendran R, Koirala S, Konoshima L, Yamazaki D, Watanabe S, Kim H & Kanae S (2013). Global flood risk under climate change. *Nature Climate Change*, 3:816–821.

Hofer M, Marzeion B & Mölg T (2012). Comparing the skill of different reanalyses and their ensembles as predictors for daily air temperature on a glaciated mountain (Peru). *Climate Dynamics*, 39(7):1969–1980. DOI:10.1007/s00382-012-1501-2.

Hofert M, Kojadinovic I, Maechler M & Yan J (2016). Multivariate dependence with copulas: Copulas package. *cran.r-project.org*.

Hundecha Y, St-Hilaire A, Ouarda T, El Adlouni S & Gachon P (2008). A non-stationary extreme value analysis for the assessment of changes in extreme annual wind speed over the gulf of st.lawrence, Canada. *Journal of Applied Meteorology and Climatology*, 47:2745–2759.

Hurrell JW & Van Loon H (1997). Decadal variations in climate associated with the north atlantic oscillation. *Clim. Change*, 36:301–326.

Jagger TH & Elsner JB (2006). Climatology models for extreme hurricane winds near the united states. *Journal of Climate*, 19(13):3220–3236. DOI:10.1175/JCLI3913.1.

Jenkinson AF (1955). The frequency distribution of the annual maximum (or minimum) of meteorological elements. *Q. J. R. Meteorol. Soc*, 81:158–171.

Joe H (1997). *Multivariate models and dependence concepts*. Chapman & Hall, London.

Joe H (2005). Asymptotic efficiency of the two-stage estimation method for copula-based models. *Journal of Multivariate Analysis*, 94:401–419.

Joe H & Xu J (1996). The estimation method of inference functions for margins for multivariate models. Department of Statistics, University of British Columbia.

Kaiser HF (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement*, 20(1):141–151. DOI:10.1177/001316446002000116.

Kallache M, Vrac M, Naveau P & Michelangeli PA (2011). Nonstationary probabilistic downscaling of extreme precipitation. *Journal of Geophysical Research: Atmospheres*, 116(D5):n/a–n/a. DOI:10.1029/2010JD014892. D05113.

Kalnay E, Kanamitsu M, Kistler R, Collins W, Deaven D, Gandin L, Iredell M, Saha S, White G, Woollen J, Zhu Y, Leetmaa A, Reynolds R, Chelliah M, Ebisuzaki W, Higgins W, Janowiak J, Mo KC, Ropelewski C, Wang J, Jenne R & Joseph D (1996). The ncep/ncar 40-year reanalysis project. *Bulletin of the American Meteorological Society*, 77(3):437–471. DOI:10.1175/1520-0477(1996)077<0437:TNYRP>2.0.CO;2.

Kendall M (1948). Rank correlation methods. *Charles Griffin and Company Limited*.

Kendall MG (1975). Rank correlation methods london. *Charles Griffin*.

Khaliq M, Ouarda T & Gachon P (2009). Identification of temporal trends in annual and seasonal low flows occurring in Canadian rivers: The effect of short-and long-term persistence. *Journal of Hydrology*, 368:183–197.

Khaliq M, Ouarda T, Ondo JC, Gachon P & Bobée B (2006). Frequency analysis of a sequence of dependent and/or non-stationary hydro-meteorological observations: A review. *Journal of Hydrology*, 329(3–4):534 – 552. DOI:http://dx.doi.org/10.1016/j.jhydrol.2006.03.004.

Khaliq MN & Gachon P (2010). Pacific decadal oscillation climate variability and temporal pattern of winter flows in northwestern north america. *Journal of Hydrometeorology*, 11:917–933.

Khaliq MN & Ouarda TBMJ (2007). On the critical values of the standard normal homogeneity test (snht). *International Journal of Climatology*, 27(5):681–687. DOI:10.1002/joc.1438.

Kistler R, Collins W, Saha S, White G, Woollen J, Kalnay E, Chelliah M, Ebisuzaki W, Kanamitsu M, Kousky V, van den Dool H, Jenne R & Fiorino M (2001). The ncep–ncar 50–year reanalysis: Monthly means cd–rom and documentation. *Bulletin of the American Meteorological Society*, 82(2):247–267. DOI:10.1175/1520-0477(2001)082<0247:TNNYRM>2.3.CO;2.

Knippertz P, Christoph M & Speth P (2003). Long-term precipitation variability in morocco and the link to the large-scale circulation in recent and future climates. *Meteorology and Atmospheric Physics*, 83(1):67–88. DOI:10.1007/s00703-002-0561-y.

Koenker R (2005). *Quantile Regression*. Cambridge University Press.

Koenker R & Bassett G (1978). Regression quantiles. *Econometrica*, pages 33–50.

Koenker R & Bassett G (1987). Regression quantiles. *Econometrica*, 46(1):33–50.

Koenker R, Ng P & Portnoy S (1994). Quantile smoothing splines. *Biometrika*, 4(81):673–680.

Koenker R & Schorfheide F (1994). Quantile spline models for global temperature. *Climatic change*, 28.

Kwiatkowski D, Phillips PCB, Schmidt P & Shin Y (1992). Testing the null hypothesis of stationarity against the alternative of a unit root. *Journal of Econometrics*, 54(1–3):159–178.

Lang M, Ouarda T & Bobee B (1999). Towards operational guidelines for over-threshold modeling. *Journal of Hydrology*, 225:103–117.

Lee E, Noh H & Park B (2014). Model selection via bayesian information criterion for quantile regression models. *J. Am. Statist. Ass*, 109.

Leybourne SJ & McCabe BPM (1994). A consistent test for a unit root. *Journal of Business and Economic Statistics*, 12:157–166.

Madsen H, Pearson C & Rosbjerg D (1997). Comparison of annual maximum series and partial duration series methods for modeling extreme hydrologic events.2.regional modeling. *Water Resources Research*, 33:759–769.

Mann H (1945). Nonparametric tests against trend. *Econometrica*, 3:245–259.

Maraun D, Wetterhall F, Ireson AM, Chandler RE, Kendon EJ, Widmann M, Brienen S, Rust HW, Sauter T, Themeßl M, Venema VKC, Chun KP, Goodess CM, Jones RG, Onof C, Vrac M & Thiele-Eich I (2010). Precipitation downscaling under climate change: Recent developments to bridge the gap between dynamical models and the end user. *Reviews of Geophysics*, 48(3).

Martin-Vide J & Lopez-Bustins JA (2006). The western mediterranean oscillation and rainfall in the iberian peninsula. *International Journal of Climatology*, 26(11):1455–1475. DOI:10.1002/joc.1388.

Martins E & Stedinger J (2000). Generalized maximum likelihood gev quantile estimators for hydrologic data. *Water Resour. Res.*, 36:737–744.

Metropolis N, Rosenbluth A, Rosenbluth M, Teller A & Teller E (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092.

Meylan P, Favre A & Musy A (2012). *Predictive Hydrology: A Frequency Analysis Approach*. CRC Press, Taylor and Francis Group. Science Publishers, Enfield, NH, USA.

Mitchell J (Septembre,2002). Down the drain? the incredible shrinking great lakes. *National Geographic*.

Nasri B, El-Adlouni S & Ouarda T (2013). Bayesian estimation for gev-b-spline model. *Open Journal of Statistics*, 3:118–128.

Nasri B, Tramblay Y, El-Adlouni S, Hertig E & Ouarda BMJ (2016). Atmospheric predictors for annual maximum precipitation in north africa. *Journal of Applied Meteorology and Climatology*.

Nathan MJ & Hare SR (2002). The pacific decadal oscillation. *Journal of Oceanography*, 58:35–44.

Nelsen R (2006). *An Introduction to Copulas*. Springer-Verlag New York.

Nelson RB (1999). *An Introduction to Copulas*. Springer New York.

Neville S, Palmer M & Wand M (2011). Generalized extreme value additive model analysis via mean field variational bayes. *Australian AND New Zealand Journal of Statistics*, 53(3):305 –330.

Nishii R (1984). Asymptotic properties of criteria for selection of variables in multiple regression. *Annals of Statistics*, 12.

Noh H, El Ghouch A & Keilegom I (2012). Quality of fit measures in the framework of quantile regression. *Scandinavian journal of statistics: theory and application*, 40:105–118.

Noh H, El Gouch A & Bouezmarni T (2013). Copula-based regression estimation and inference. *Journal of the American Statistical Association*, 108:676–688.

Noh H, Ghouch AE & Keilegom IV (2015). Semiparametric conditional quantile estimation through copula-based multivariate models. *Journal of Business & Economic Statistics*, 33(2):167–178. DOI:10.1080/07350015.2014.926171.

Oakes D (1982). A coefficient of concordance for censored data. *Biometrics*, 38:451–455.

Olsen J, Stedinger J, Matalas N & Stakhiv E (1999). Climate variability and flood frequency estimation for the upper mississippi and lower missouri rivers. *Journal of the American Water Resources Association*, 35(6):1509–1523.

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

Ouarda TBMJ & Adlouni SE (2011). Bayesian nonstationary frequency analysis of hydrological variables. *Journal of the American Water Resources Association*, 47(3):496–505.

Pettitt AN (1988). A non-parametric approach to the change-point problem. *J. Roy. Stat. Soc.*, 28C:126–135.

Pickands J (1975). Statistical inference using extreme order statistics. *Annals of Statistics*, 3:119–131.

Preisendorfer RW (1988a). La "catastrophe" climatique de l'automne 1969 en tunisie. *Annales de Géographie*, 79:581–595.

Preisendorfer RW (1988b). Principal component analysis in meteorology and oceanography. *Developments in Atmospheric Sciences*, 17.

Raftery AE & Lewis SM (1992). [practical Markov Chain Monte Carlo]: Comment: One long run with diagnostics: Implementation strategies for Markov chain monte carlo. *Statist. Sci.*, 7(4):493–497. DOI:10.1214/ss/1177011143.

Raftery AE & Lewis SM (1995). *The number of iterations, convergence diagnostics and generic Metropolis algorithms. Practical Markov Chain Monte Carlo.* Chapman and Hall.

Regonda SK, Rajagopalan B, Clark M & Pitlick J (2005). Seasonal cycle shifts in hydroclimatology over the western united states. *Journal of climate*, 18:372–384.

Roche PA, Miquel J & Gaume E (2012). *Hydrologie quantitative.* Springer.

Romano C (2002). Calibrating and simulating copula functions: An application to the italian stock market. *Working Paper, Università di Roma*, 12.

Rosenblatt M (1952). Remarks on a multivariate transformation. *The Annals of Mathematical Statistics*, 23:470–472.

Rowan TS & Daniel LRH (2005). Atlantic ocean forcing of north american and european summer climate. *Science*, 309.

Saidi MEM, Daoudi L, Aresmouk MEH & Blali A (2003). Rôle du milieu physique dans l'amplification des crues en milieu montagne montagnard: Exemple de la crue du 17 août 1995 dans la vallée de l'ourika (haut-atlas, maroc). *Sécheresse*, pages 107–114.

Sandink D (2013). Urban ooding in Canada: Lot-side risk reduction through voluntary retrofit programs, code interpretation and by-laws. *Toronto, Ontario: The Institute for Catastrophic Loss Reduction (ICLR):Building resilient communities*, 52:105–118.

Scaillet O (2007). Kernel based goodness-of-fit tests for copulas with fixed smoothing parameters. *Journal of Multivariate Analysis*, 98:533–543.

Scheipl F (2011). spikeslabgam: Bayesian variable selection, model choice and regularization for generalized additive mixed models in r. *Journal of Statistical Software*, 43(14):1–24.

Schwarz GE (1978). Estimating the dimension of a model. *Annals of Statistics*, 6.

Schweizer B & Wolff EF (1981). On nonparametric measures of dependence for random variables. *The Annals of Statistics*, 9:879–885.

Shih J & Louis TA (1995). Inference on association parameter in copula models for bivariate survival data. *Biometrics*, 26:183–214.

Sillmann J, Kharin VV, Zhang X, Zwiers FW & Bronaugh D (2013). Climate extremes indices in the cmip5 multimodel ensemble: Part 1. model evaluation in the present climate. *Journal of Geophysical Research: Atmospheres*, 118(4):1716–1733. DOI:10.1002/jgrd.50203.

Smakhtin VU (2001). Low flow hydrology: a review. *Journal of hydrology*, 240(3-4):147–186.

Smith RL (1985). Maximum likelihood estimation in a class of non-regular cases. *Biometrika*, 72:67–92.

Stocker T, Qin D, Plattner GK, Tignor M, Allen S, Boschung J, Nauels A, Xia Y, Bex V & Midgley P (2013). *Climate Change: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, chapitre SPM, 1–30 pages. Cambridge University Press.

Tareghian R & Rasmussen P (2013). Analysis of arctic and antarctic sea ice extent using quantile regression. *International Journal of Climatology*, 33(5):1079–1086.

Teegavarapu RSV, Goly A & Obeysekera J (1969). Atmospheric teleconnections from the equatorial pacific. *Monthly weather review*, 97:163–172.

Teegavarapu RSV, Goly A & Obeysekera J (2013). Influences of atlantic multi-decadal oscillation on regional precipitation extremes. *Journal of Hydrology*, 495:74–93.

Thompson P, Cai Y, Moyeed R, Reeve D & Stander J (2010). Bayesian nonparametric quantile regression using splines. *Computational Statistics and Data Analysis*, 54:1138–1150.

Tibshirani R (1996). regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288.

Toreti A, Xoplaki E, Maraun D, Kuglitsch FG, Wanner H & Luterbacher J (2010). Characterisation of extreme winter precipitation in mediterranean coastal sites and associated anomalous atmospheric circulation patterns. *Natural Hazards and Earth System Sciences*, 10(5):1037–1050. DOI:10.5194/nhess-10-1037-2010.

Tramblay Y, Badi W, Driouech F, Adlouni SE, Neppel L & Servat E (2012). Climate change impacts on extreme precipitation in morocco. *Global and Planetary Change*, 82–83:104 – 114.

Tramblay Y, El Adlouni S & Servat E (2013). Trends and variability in extreme precipitation indices over maghreb countries. *Natural Hazards and Earth System Sciences*, 13(12):3235–3248. DOI:10.5194/nhess-13-3235-2013.

Tramblay Y, Neppel L & Carreau J (2011). Brief communication "climatic covariates for the frequency analysis of heavy rainfall in the mediterranean region". *Natural Hazards and Earth System Sciences*, 11(9):2463–2468. DOI:10.5194/nhess-11-2463-2011.

Van der Vaart A (1998). *Asymptotic statistics*. Cambridge University Press.

Vicente-Serrano SM, Beguería S, López-Moreno JI, El Kenawy AM & Angulo-Martínez M (2009). Daily atmospheric circulation events and extreme precipitation risk in northeast spain: Role of the north atlantic oscillation, the western mediterranean oscillation, and the mediterranean oscillation. *Journal of Geophysical Research: Atmospheres*, 114(D8):n/a–n/a. DOI:10.1029/2008JD011492. D08106.

Vogel R, Thomas WJ & McMahon T (1993). Flood flow frequency model selection in southwestern united states. *J. Water Resour. Plann. Manage.*, 353:353–366.

Vrac M & Naveau P (2007). Stochastic downscaling of precipitation: From dry events to heavy rainfalls. *Water Resources Research Journal*, 43(7).

Wald A & Wolfowitz J (1940). On a test whether two samples are from the same population. *Ann. Math. Statist.*, 11(2):147–162. DOI:10.1214/aoms/1177731909.

Wanner H, Brönnimann S, Casty C, Gyalistras D, Luterbacher J, Schmutz C, Stephenson DB & Xoplaki E (2001). North atlantic oscillation – concepts and studies. *Surveys in Geophysics*, 22(4):321–381. DOI:10.1023/A:1014217317898.

Wetterhall F, Halldin S & yu Xu C (2005). Statistical precipitation downscaling in central sweden with the analogue method. *Journal of Hydrology*, 306(1–4):174 – 190. DOI:http://dx.doi.org/10.1016/j.jhydrol.2004.09.008.

Wilcoxon F (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83.

Wu Y & Zen MM (1999). A strongly consistent information criterion for linear model selection based on m-estimation. *Probability Theory and Related Fields*, 625.

Xie Q (2015). Computation and application of copula-based weighted average quantile regression. *Journal of Computational and Applied Mathematics*, 281:182–195.

Xoplaki E, González-Rouco JF, Luterbacher J & Wanner H (2004). Wet season mediterranean precipitation variability: influence of large-scale dynamics and trends. *Climate Dynamics*, 23(1): 63–78. DOI:10.1007/s00382-004-0422-0.

Yamato H (1973). Uniform convergence of an estimator of a distribution function. *Bull.Math.Statist*, 15:69–78.

Yu K & Moyeed R (2001). Bayesian quantile regression. *Statistics and Probability Letters*, 54:437–447.

Yu X, Cohn T & Stedinger J (2015). Flood frequency analysis in the context of climate change. *World Environmental and Water Resources Congress 2015*, pages 2376–2385.

Zhang Y, Li R & Tsai CL (2010). Regularization parameter selections via generalized information criterion. *Journal of the American Statistical Association*, 105.