Comportement asymptotique de la distribution des pluies extrêmes en France

Aurélie Muller

UNIVERSITE MONTPELLIER II

SCIENCES ET TECHNIQUES DU LANGUEDOC

THÈSE

présentée pour obtenir le titre de

Docteur de l'Université Montpellier II

Statistiques appliquées et Hydrologie

Ecole doctorale : Information, Structures, Systèmes

Comportement asymptotique de la distribution des pluies extrêmes en France

---

Préparée dans l'Unité de Recherche Hydrologie-Hydraulique, ***Cemagref***

par

Aurélie Muller

Soutenue le 24 novembre 2006 devant le jury composé de :

M. Gilles DUCHARME	(Université Montpellier II)	Président du jury
M. Jean-Noël BACRO	(Université Montpellier II)	Directeur de thèse
M. Michel LANG	(*Cemagref*)	Co-directeur
M. Patrick ARNAUD	(*Cemagref*)	Co-directeur
M. Stephane GIRARD	(IMAG)	Rapporteur
M. Jean-Pierre LABORDE	(Université Nice)	Rapporteur
M. Luc NEPPEL	(Hydrosciences Montpellier)	Invité

Remerciements

En premier, je pense naturellement à mes trois directeurs de thèse : Michel Lang, Jean-Noël Bacro et Patrick Arnaud. J'ai beaucoup appris en travaillant avec vous, tant au niveau de la modélisation hydrologique qu'au niveau des statistiques des valeurs extrêmes. Merci pour votre soutien, chacun selon ses compétences, et votre attention au moral de votre étudiante !

Je tiens à remercier Stéphane Girard et Jean-Pierre Laborde, d'avoir accepté d'être les rapporteurs de ce travail, et pour leur lecture attentive.

Je remercie également Luc Neppel et Jacques Lavabre, venus plusieurs fois aux réunions de comité de pilotage. J'en profite pour remercier Aurélien Ribes, pour une discussion qui a beaucoup profité à l'un des chapitre de la thèse.

Cette thèse a été réalisée au sein de l'équipe d'Hydrologie et d'Hydraulique du Cemagref de Lyon, et je voudrais remercier toutes les personnes que j'y ai rencontrées : Jean-Michel Grésillon, notre responsable d'équipe, Hélène Faurant et Anne Escholtz, toujours présentes pour répondre aux questions pratiques, Etienne Leblois, Isabelle Braud, Jean-Baptiste Faure, André Paquier, Bernard Chastan, Gilles Galéa, notre champion Denis Barbet, Alain Recking, Sébastien Proust, Eric Hérouin, Jean-Pierre, Cécile Quignette.

Un merci tout particulier à Benjamin Renard, mon jumeau de thèse, toujours là pour aider, généreusement et dans la bonne humeur. Merci aussi à Mathieu Ribatet, venu enrichir notre équipe de matheux.

Merci à la sympathique équipe de thésards et compagnie, dans laquelle je suis "tombée" ! Merci pour la belle fête que vous m'avez faite après la soutenance ! Merci à Anne-Laure pour son accueil aussi bien à Lyon qu'à Sisteron, Elo et ses fréquents passages dans le bureau du fond, Magali et Jérôme nos sages voisins de bureau, Otmane la fusée, Benjamin notre Bisounours, Mathieu le Lyonnais Québécois, Ricounet le grognon au bon coeur, Michel le farceur, Tata Christine, Raouf le pressé de manger et Delphine ma défenseuse, Adeline et ses bons gâteaux, Judi et Marie-Liesse, les jeunes mariés, Sandhya la sportive, Céline du Sud, Olivier et Jean-Marie les fêtards, le petit Kamal, Guillaume, monsieur Pluviomètre et Caro, Arnaud mon moniteur de foot, Loïc le frappeur fou, Flora la gymnaste, Aline la courageuse, Fabien le VTTiste, Jean-Guillaume et Noémie les Martiniquais, l'attentionné Taha, la souriante Sophie, Karine la motarde, Emmanuelle la Toulousaine sans accent, Stéphane, aux allures de prof de philo, Jean-Phi, Oldrich et Fréd, les vieux thésards, Manu accroché au téléphone, Aurélien le nouveau.

Merci aussi à mes copains lyonnais : Mohammed, Jan, Michel Guillaud, Timothée, Sergio, Nico, Marc, Camille, Sébastien, Aurore, Aurélie, Geneviève, Claude-Pierre, Nicole, Mamie

Jeanne, Sixtine, Sasso, Mérète, Micheline. Merci à Robert Tantôt pour son aide. Merci à la joyeuse et courageuse Florence.

Enfin, merci à ma petite famille : Benoît, Christophe, Sandrine, Marie-Claire, mes Parents, Anthony et Albin !! Merci pour votre amour et votre patience !!

Merci à tous ceux que j'ai oubliés... Voilà, avec un tel entourage, on ne pouvait que faire au mieux !!

Table des matières

Résumé étendu

Différentes approches de la théorie des valeurs extrêmes ont été mises en oeuvre pour analyser les pluies extrêmes, à partir de séries pluviométriques mises à disposition par le Cemagref et Météo-France. Le point de vue adopté dans la thèse est local : les pluies sont analysées en un site donné, et nous laissons les points de vue régional et spatial en perspectives.

Dans une première partie, nous rappelons différents aspects de la théorie probabiliste des valeurs extrêmes uni-variées et bi-variées, et présentons des méthodes statistiques d'estimation des incertitudes de modèles probabilistes appliqués en hydrologie.

En hydrologie, les maxima saisonniers ou annuels sont souvent modélisés par une loi Gumbel. Cependant, la littérature hydrologique montre depuis peu un certain scepticisme vis-à-vis de la modélisation des maxima annuels ou saisonniers par une loi Gumbel. Il ne paraît en effet pas justifié d'utiliser la loi Gumbel, au lieu d'une loi GEV, plus générale. Nous avons analysé cette question dans la seconde partie de la thèse, à partir de longues séries de plus de 100 ans de mesures quotidiennes, après vérification de la stationnarité au second ordre des séries étudiées. Cette première approche a montré que dans un grand nombre de cas, la loi Gumbel traditionnellement utilisée par les hydrologues français est rejetée au profit de la loi GEV non bornée supérieurement et de paramètre de forme non nul, impliquant des quantiles de pluie extrême bien plus forts qu'avec une loi Gumbel. Cependant, une conclusion unique est difficile à établir car les résultats différent suivant les séries étudiées. En effet, compte tenu de la forte incertitude liée à l'estimation du paramètre de forme de la loi GEV, la loi Gumbel est acceptable dans un nombre non négligeable de cas, pour modéliser les maxima annuels.

Le problème du comportement des maxima annuels ou saisonniers a ensuite été étudié via une analyse multi-variée des maxima des cumuls de pluies mesurés sur différents pas de temps horaires (entre 1 heure et 72 heures) et journalier. D'un point de vue mathématique, nous avons utilisé un théorème sur les extrêmes ordonnés, ainsi que la théorie des valeurs extrêmes de séries stationnaires pour des variables mesurées à différentes fréquences. Plus précisément, les cumuls journaliers et les cumuls de pluie en 24 heures sont mesurés sur des pas de temps de 24 heures, mais selon un protocole expérimental différent (les cumuls journaliers sont mesurés une fois par jour à heure fixe, tandis que les cumuls de pluie en 24 heures sont mesurés par fenêtres glissantes de 1 heure). L'intérêt hydrologique du modèle a été de permettre l'estimation de la distribution des pluies maximales saisonnières mesurées sur différentes durées, grâce à un couplage de l'information horaire, souvent assez restreinte, avec une information journalière, en général mieux renseignée. Les courtes séries horaires peuvent en effet être pauvres en valeurs fortes, du fait de l'échantillonnage ou de valeurs manquantes. L'exemple étudié dans la thèse porte sur une série pluviographique de Marseille,

longue de 67 années de mesures au pas de temps 5 minutes, entre 1918 et 2002, associée à une série pluviométrique, longue de 122 années de mesures quotidiennes entre 1882 et 2003. Les valeurs de pluie les plus fortes sont absentes de la série pluviograhique, à cause d'un mauvais fonctionnement du pluviographe sous de très fortes intensités de pluie. En revanche, la série pluviométrique contient les valeurs fortes non mesurées par le pluviographe. Trois modèles de dépendance entre les pluies de différentes durées ont été proposés, en particulier en associant les connaissances hydrologiques sur les pluies de différentes durées avec la théorie des valeurs extrêmes uni-variée et bi-variée. La comparaison des trois modèles a montré que la dépendance des pluies de 24 heures et 72 heures, modélisée via une loi bi-variée des valeurs extrêmes du type logistique donne de bons résultats. Cette approche multi-variée a montré que les distributions des maxima de la saison pluvieuse de Marseille sont mieux modélisées par une distribution GEV non bornée supérieurement et de paramètre de forme non nul, que par la loi Gumbel. En particulier, une analyse uniquement marginale des données induit des quantiles sous-estimés et peu respectueux des contraintes physiques, du fait de la mauvaise qualité des séries pluviographiques en terme de valeurs extrêmes. En revanche, avec le modèle multi-varié, les quantiles sont revus à la hausse, et de manière plus cohérente avec la réalité. Les estimations de cette étude ont été réalisées dans un cadre bayésien, à l'aide d'algorithmes Monte Carlo de Chaînes de Markov (MCMC), qui ont permis l'estimation d'intervalles de crédibilité des distributions des extrêmes de pluies de différentes durées. Cette étude a fait l'objet d'un article, accepté à la revue Stochastic Environmental Research and Risk Assessment.

Dans la troisième partie, le comportement asymptotique de la distribution des pluies est étudié au travers de modèles stochastiques, générateurs de pluie. Un premier chapitre présente une bibliographie des différents types de générateurs, à des pas de temps journaliers ou horaires. Le second chapitre analyse plus particulièrement les incertitudes du modèle Shypre, développé au Cemagref. Dans sa version simplifiée, le modèle Shypre possède trois paramètres et simule des averses de pluies. Les paramètres de Shypre sont estimés par des moyennes d'un grand nombre de valeurs (nombre moyen d'événements pluvieux par an, moyenne de la pluie maximale d'un événement pluvieux, durée moyenne d'un événement pluvieux). Les résultats obtenus avec ce modèle ont été comparés en terme d'estimation et d'incertitudes des quantiles de pluie, avec les résultats donnés par une approche probabiliste classique des valeurs extrêmes. L'analyse des incertitudes a pris en compte l'incertitude d'échantillonnage des données (via une analyse fréquentielle) et l'incertitude des paramètres (via une analyse bayésienne). Cette étude montre d'une part que les estimations du modèle Shypre sont peu sensibles à l'échantillonnage, du fait de la paramétrisation par des paramètres moyens. D'autre part, l'étude montre que les incertitudes de modélisation sont très faibles : les intervalles de crédibilité des quantiles obtenus sont très étroits. L'étroitesse des intervalles de crédibilité obtenus suggère que la vraie valeur de quantile est recouverte avec une probabilité inférieure à celle du niveau de l'intervalle de crédibilité. Dans le cas de la loi des valeurs extrêmes, les conclusions observées sont à l'opposé de celles obtenues pour Shypre. La loi des valeurs extrêmes se montre fortement sensible à l'échantillonnage, essentiellement du fait de son paramètre de forme, et les intervalles de crédibilité obtenus peuvent être très larges, voire irréalistes pour des mesures de pluie. Cette étude fait l'objet d'un article en cours de finalisation, à soumettre dans la revue Hydrological Sciences Journal.

Dans les analyses précédentes, les valeurs extrêmes de pluies ont été étudiées via une extraction des maxima annuels ou saisonniers, ou via les dépassements de seuil. En revanche, le comportement temporel des valeurs de pluies successives a été ignoré. L'objet de la dernière

partie de la thèse est donc de proposer un modèle stationnaire, capable de modéliser une série temporelle de pluie, ainsi que ses valeurs extrêmes. L'enjeu particulier d'un tel modèle est de représenter la dépendance temporelle entre des valeurs extrêmes successives de pluie. Le modèle proposé est un processus markovien, il combine la théorie des valeurs extrêmes univariée pour définir la loi marginale du processus, et bi-variée pour définir la dépendance par la fonction de saut de la chaîne de Markov. Deux chapitres sont consacrés à l'étude d'un tel modèle. Le premier permet de valider le modèle sur un processus stationnaire théorique connu, dont la loi marginale est asymptotiquement une loi des valeurs extrêmes, et dont la dépendance temporelle est définie par une structure de dépendance de Morgenstern. Le modèle proposé est ensuite appliqué dans le second chapitre à des données réelles d'averses de Marseille. L'intérêt opérationnel du modèle est de pouvoir être inséré dans le modèle Shypre. En effet, parmi les différents générateurs de pluies existant, la spécificité du modèle Shypre est de modéliser les valeurs extrêmes, et plus particulièrement un phénomène de persistance des fortes averses. Ce phénomène de persistance est défini par le fait que lors d'un événement pluvieux extrême, les averses composant l'événement sont dépendantes les unes des autres. Cette persistance des fortes averses est actuellement modélisée dans Shypre de manière empirique. L'approche théorique proposée dans la thèse est donc une alternative à la modélisation empirique uilisée dans Shypre.

Introduction

Pour protéger les populations avec leurs habitations, les zones industrielles, etc., des inondations, plusieurs mesures peuvent être prises. Des ouvrages hydrauliques (barrages, digues, ponts, etc.) peuvent être construits, des Plans de Préventions des Risques d'Inondation peuvent être établis. Le dimensionnement des ouvrages hydrauliques, l'établissement des Plans de Prévention des Risques d'Inondation sont fondés sur le calculs de crues extrêmes de référence (de fréquence d'apparition de l'ordre de 10^{-2} à 10^{-4}). Par ailleurs, lorsqu'un événement pluvieux extrême survient, il est positionné sur une échelle d'intensité de la gravité de l'événement pluvieux (faible, moyenne, forte, exceptionnelle). En hydrologie, des méthodes ont été proposées pour estimer les crues extrêmes de référence, ainsi qu'une échelle d'intensité, à partir de la distribution des pluies extrêmes. En effet, l'information pluviométrique, mesurée avec des pluviométres ou pluviographes, est en général mieux renseignée que les mesures de débits de rivières, alors en crues pendant les événements extrêmes. Certaines de ces méthodes, couramment employées en France (par exemple, la méthode du Gradex (Guillot et Duband, 1967), (CFGB, 1994)), reposent sur l'hypothèse forte suivante : la distribution des pluies maximales annuelles ou saisonnières est une loi Gumbel. La loi Gumbel n'est cependant qu'un cas particulier d'une loi des valeurs extrêmes plus générale : la loi GEV. Quelques auteurs (Wilks, 1993; Koutsoyiannis et Baloutsos, 2000; Chaouche *et al.*, 2002; Coles *et al.*, 2003; Coles et Pericchi, 2003; Sisson *et al.*, 2006; Koutsoyiannis, 2004a,b; Bacro et Chaouche, 2006) ont récemment mis en cause la validité d'une loi Gumbel, et préféré l'usage de la loi GEV pour modéliser les maxima annuels ou saisonniers des pluies. Or la différence entre les quantiles estimés sous une hypothèse de loi Gumbel ou sous une hypothèse de loi GEV est considérable. Pour une fréquence donnée, les quantiles d'une loi GEV peuvent par exemple être deux ou trois fois plus grands que les quantiles d'une loi Gumbel. Ces remarques ne sont donc pas sans conséquence pour la sécurité des ouvrages hydrauliques, pour la validité des Plans de Prévention des Risques d'Inondation ou pour la définition de l'échelle d'intensité des événements extrêmes. En particulier, le Cemagref a introduit un assouplissement dans l'utilisation de la méthode du Gradex, en autorisant des lois présentant une courbure sur un graphique de probabilité à échelle logarithmique (Margoum *et al.*, 1994; Lang, 1997).

D'autre part, l'approche fréquentielle n'est pas la seule méthode proposée en hydrologie pour estimer les quantiles de pluies extrêmes, notamment lorsque les données observées sont peu nombreuses ou de mauvaise qualité. Une alternative est proposée avec des méthodes stochastiques : les estimations des distributions de pluies ou de débits reposent sur des simulations de très longues chroniques de pluie par un générateur de pluie, couplé d'un modèle pluie-débit transformant des données de pluie en données de débit. Au Cemagref, un tel générateur (nommé Shypre) a été en particulier développé dans le but d'estimer des quantiles de fréquence rare. Arnaud *et al.* (1998) ont montré, à partir d'observations sur une dizaine de postes pluviographiques du sud-est de la France, que la distribution simulée par le

modèle Shypre est proche d'une distribution GEV non bornée supérieurement, et estime des quantiles de pluies plus forts qu'une estimation par une loi Gumbel. Ces résultats rejoignent donc ceux de (Wilks, 1993; Koutsoyiannis et Baloutsos, 2000; Chaouche *et al.*, 2002; Coles *et al.*, 2003; Coles et Pericchi, 2003; Sisson *et al.*, 2006; Koutsoyiannis, 2004a,b; Bacro et Chaouche, 2006).

L'objet de cette thèse est donc de considérer la distribution de probabilité des pluies extrêmes, à partir d'observations sur des postes de mesures français. La définition même de pluie extrême est délicate. De manière classique, une pluie extrême peut être définie par une pluie dépassant un seuil élevé donné, ou par la pluie maximale observée en une année donnée. Les compétences de trois disciplines sont conjuguées dans cette thèse : les probabilités avec en particulier la théorie des valeurs extrêmes, les statistiques, et l'hydrologie et plus particulièrement l'étude fréquentielle des pluies.

Le comportement des valeurs extrêmes de pluie est décrit dans la thèse sous différents angles. Nous analysons les valeurs extrêmes de pluie via :(a) les maxima annuels ou saisonniers de pluie ; (b) l'analyse multi-variée des maxima annuels ou saisonniers d'un processus de pluie décrit à différents pas de temps de mesure ; (c) les valeurs dépassant un seuil élevé ; (d) un générateur stochastique de pluie, en particulier le modèle Shypre ; (e) enfin l'analyse temporelle d'un processus pluvieux comportant des événements extrêmes.

Les modèles proposés dans la thèse combinent différents aspects de la théorie uni-variée et bi-variée des valeurs extrêmes. Les modèles proposés sont appliqués à des processus de pluie réels, ou à des processus théoriques, définis par un générateur de pluie ou, dans la dernière partie de la thèse, par un processus markovien de fonction de saut déterminée par une structure de dépendance de Morgenstern. La structure de dépendance de Morgenstern a été utilisée dans un but exploratoire, et dans un but de validation d'un modèle stationnaire particulier proposé dans la thèse.

Des chroniques de mesures de pluie ont été mises à notre disposition par Météo-France et le Cemagref. Nous étudions plus particulièrement de longues chroniques, de plus de 100 ans de mesures quotidiennes. D'autre part, nous disposons de chroniques de mesures de pluie au pas de temps 1 heure, ce qui nous permet d'étudier le comportement des événements pluvieux extrêmes à des pas de temps entre 1 heure et 72 heures. Enfin, nous étudions le processus pluvieux par une chronique d'événements pluvieux décrits en termes d'averses.

Des méthodes d'estimations bayésiennes sont mises en oeuvre avec des algorithmes Monte Carlo de Chaînes de Markov pour estimer les quantiles et les incertitudes associées, correspondant aux modèles proposés.

La première partie de la thèse est un rappel des théories probabilistes des valeurs extrêmes uni-variées et bi-variées (*chapitre 1*), et une présentation des méthodes statistiques d'estimation des incertitudes des quantiles de modèles probabilistes ou stochastiques complexes (*chapitre 2*).

Dans la deuxième partie, nous donnons une analyse bibliographique sur les raisons du choix de la loi Gumbel en hydrologie et les récentes évolutions pour modéliser les maxima annuels ou saisonniers de pluie (*chapitre 3*). Nous étayons ensuite les résultats bibliographiques par une analyse des maxima annuels de longues séries de mesures en France, mises à notre disposition par Météo-France. Enfin, nous complétons cette étude par la proposition d'un modèle multivarié (*chapitre 4*) pour analyser conjointement les maxima annuels ou saisonniers de pluie mesurées sur différents pas de temps (entre 1 heure et 72 heures). L'une des nouveautés du

modèle proposé est l'analyse dans une même étude, des pluies mesurées sur plusieurs pas de temps mobiles multiples de l'heure et des pluies journalières, mesurées sur des pas de temps fixes de 24 heures.

Les processus pluvieux sont décrits dans la troisième partie, par les générateurs stochastiques de pluie. Un chapitre est consacré à une bibliographie des différents types de générateurs de pluie, aux pas de temps journalier ou inférieurs à 24 heures (*chapitre 5*). L'attention est plus particulièrement portée sur leur capacité à modéliser les valeurs extrêmes. Le chapitre suivant analyse plus particulièrement la distribution asymptotique des pluies simulées par une version simplifiée du générateur Shypre, ainsi que les incertitudes associées aux quantiles estimés (*chapitre 6*).

Enfin, la dernière partie de la thèse est consacrée à la description temporelle d'une série stationnaire présentant des valeurs extrêmes. Un modèle markovien au premier ordre est proposé, et validé sur deux cas d'étude : un cas d'étude théorique avec un processus markovien au premier ordre de structure de dépendance définie par un modèle de Morgenstern (*chapitre 7*), puis des données réelles d'averses de Marseille (*chapitre 8*).

Une conclusion générale dresse une synthèse des principaux résultats obtenus pendant la thèse et donne quelques perspectives possibles de recherches ultérieures.

Première partie

Approches probabilistes et analyse des incertitudes sur les valeurs extrêmes

Chapitre 1

Théorie probabiliste

1.1 Analyse exploratoire des valeurs extrêmes

Avant d'exposer la théorie des valeurs extrêmes, un premier aperçu du comportement des extrêmes est donné par une analyse graphique des données. Un QQ-plot représente les quantiles empiriques en fonction des quantiles d'une distribution G choisie. Une correspondance linéaire entre quantiles observés et théoriques indique une bonne adéquation des observations à la loi G. Le tracé d'un QQ-plot pour un échantillon ordonné $X_{1,n} \leq \ldots \leq X_{n,n}$, nécessite un estimateur de $P(X \leq X_{i,n})$, noté $\widehat{f_i}$ et appelé fréquence empirique.

Dans la littérature, différents estimateurs sont proposés pour les fréquences empiriques, voici quelques exemples :

$$\widehat{f_i} = i/n, \tag{1.1}$$

$$\widehat{f_i} = (i-a)/(n+1-2a), \tag{1.2}$$

où $a \in [0, 0.5]$. Le choix de la meilleure formule de fréquence empirique dépend de la loi sous-jacente de l'échantillon, modélisée par la distribution G. En hydrologie, elle est choisie en fonction de certains critères, établis par Cunnane (1978). En particulier, le QQ-plot tracé avec la distribution G doit permettre de juger si l'échantillon est tiré de la distribution G. Un tel estimateur $\widehat{f_i}$ est donné par $G(E(X_{i,n}))$, la fréquence de la moyenne de la i-ème statistique d'ordre d'un $n-$échantillon tiré dans la distribution G. Une approximation, compromis entre plusieurs distributions pour G (dont la loi Gumbel), est donnée par $\widehat{f_i} = (i-0.4)/(n+0.2)$ (Cunnane, 1978).

Guo (1990), cité par Naulet (2002), a comparé plusieurs formulations de la fréquence empirique, et montré que la formule de Cunnane (1978) est la moins biaisée sur les valeurs extrêmes. Dans la thèse, on utilisera donc la formule $\widehat{f_i} = (i-0.4)/(n+0.2)$, afin de se situer dans un cadre général.

1.2 Théorie des valeurs extrêmes

1.2.1 Analyse par les valeurs maximales

L'étude des extrêmes d'un processus passe naturellement par l'analyse du maximum d'un échantillon de taille n donnée : $M_n = \max\{X_1, \ldots, X_n\}$, où $X_1, \ldots, X_n$ est un échantillon i.i.d. de loi F. L'analyse des maxima d'échantillons de taille n est également appelée analyse des maxima par bloc. Si on note $x_F = \sup\{x, F(x) < 1\}$, alors $M_n \to x_F$ presque sûrement si $n \to \infty$. D'autre part, la distribution de M_n est connue exactement :

$$P(M_n \leq z) = (F(z))^n. \tag{1.3}$$

En pratique, F est inconnue, et la relation 1.3 n'est donc pas utilisable directement. De façon analogue au théorème central limite, la théorie des valeurs extrêmes montre qu'il existe des suites $\{a_n\}, \{b_n\}$, avec $a_n > 0, b_n \in \mathbb{R}$ telles que

$$P(\frac{M_n - b_n}{a_n} \leq x) \to G(x) \tag{1.4}$$

si $n \to \infty$, avec G non dégénérée. Deux questions se posent ici : quelle est la loi G ? et quelles conditions doit vérifier F pour qu'il existe des suites $\{a_n\}, \{b_n\}$ satisfaisant 1.4 ?

– Fisher et Tippett (1928), Gnedenko (1943), de Haan (1970) ont montré que les seules distributions limites non-dégénérées G possibles sont les distributions de valeurs extrêmes. Le théorème suivant résume ce résultat :

  Théorème : S'il existe des suites de réels $\{a_n\}, \{b_n\}$ satisfaisant la relation limite 1.4 lorsque $n \to \infty$ pour une distribution G non dégénérée, alors G appartient à la famille des distributions GEV (Generalized Extreme Value) :

$$G(z) = \exp\{-[1 - k(z - \beta)/\alpha]^{1/k}\} \tag{1.5}$$

  G est définie sur $\{z \in \mathbb{R} : 1 - k(z - \beta)/\alpha > 0\}$, avec $\beta \in \mathbb{R}, \alpha > 0, k \in \mathbb{R}$ les paramètres de position, échelle, forme de G.

  La distribution du maximum d'un échantillon de taille suffisamment élevée est donc approximativement une GEV.
  Certaines démonstrations de ce théorème utilisent les distributions max-stables (Leadbetter *et al.*, 1983), d'autres utilisent le théorème de Helly-Bray (Beirlant *et al.*, 2004). Suivant le signe de k, le comportement de G est différent, on définit trois types de lois GEV : les cas $k < 0, k = 0, k > 0$ sont respectivement appelés distributions de type Fréchet, Gumbel et Weibull. Le comportement des extrêmes est fortement influencé par le signe et la valeur de k. Si $k < 0$, $x_G = +\infty$, mais G a un support borné à gauche. Si $k = 0$ le support de la distribution G est $\mathbb{R}$. Si $k > 0$, $x_G < \infty$, et le support de G n'est pas borné à gauche. Les densités des trois types de distributions sont présentées dans la figure 1.1 (a). Les distributions de type Weibull ont les queues les moins lourdes, les distributions de type Fréchet ont les queues les plus lourdes, comme le montre le QQ-plot de la figure 1.1 (b).

– Pour répondre à la deuxième question, on définit le domaine d'attraction :

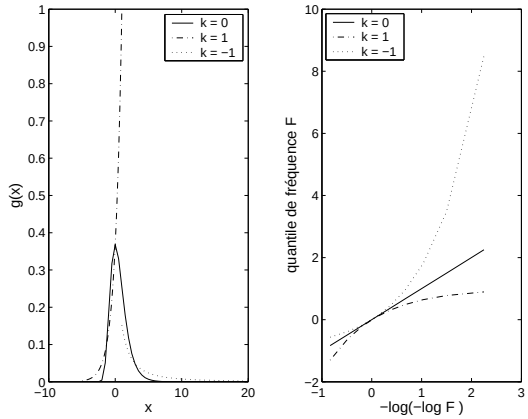


Fig. 1.1: (a). Allure des fonctions de densité des lois GEV et Gumbel pour différentes valeurs du paramètre de forme k. (b). Allure des QQ-plot dans un diagramme de Gumbel.

Définition : Une variable aléatoire X appartient au domaine d'attraction de la distribution de valeurs extrêmes G s'il existe des suites $\{a_n\}, \{b_n\}$ avec $a_n > 0, b_n \in \mathbb{R}$ satisfaisant la relation 1.4.
Avant de caractériser les domaines d'attraction, on définit les fonctions à variation lente.

Définition : Une fonction positive L est à variation lente en l'infini si $\lim_{x\to\infty} \frac{L(tx)}{L(x)} = 1$ pour tout $t > 0$. Une fonction positive f est à variation régulière en l'infini, et d'indice $c \in \mathbb{R}$ si $\lim_{x\to\infty} \frac{f(tx)}{f(x)} = t^c$, pour tout $t > 0$.

Soit F une distribution vérifiant la relation limite 1.4, trois cas de figure peuvent se produire sur la distribution G limite : le paramètre de forme k est nul, strictement positif ou strictement négatif. Le théorème suivant résume les différents cas :
Théorème :

- La distribution F appartient au domaine d'attraction de la loi GEV de paramètre de forme $k < 0$ si et seulement si $\bar{F}(x) = x^k L(x)$, avec L une fonction à variation lente en l'infini. Ces fonctions sont dites de type Pareto : $\bar{F}(x) \approx Kx^k$ lorsque $x \to \infty$, pour $K > 0$ (de Haan, 1970).
- La distribution F appartient au domaine d'attraction de la loi GEV de paramètre de forme $k > 0$ si et seulement si $x_F < \infty$ et $\bar{F}(x_F - x^{-1}) = x^k L(x)$, avec L une fonction à variation lente en l'infini (Resnick, 1987).
- La distribution F appartient au domaine d'attraction de la loi GEV de paramètre de forme $k = 0$ si et seulement s'il existe une fonction positive $\tilde{a}$ telle que

$$\lim_{x\to x_F} \frac{\bar{F}(x + t\tilde{a}(x))}{\bar{F}(x)} = \exp(-t), t \in \mathbb{R}.$$

  Donc la partie droite de la distribution de F décroît vers 0 plus vite que n'importe quelle fonction puissance (de Haan, 1970).

L'analyse des valeurs extrêmes via les maxima par blocs comporte certaines limites. Considérons par exemple les maxima annuels d'une série de pluies journalières, les blocs sont les années. Du fait de l'échantillonnage, il peut arriver qu'une année contienne plusieurs valeurs extrêmes, et qu'une autre année n'en contienne aucune. Dans le premier cas de figure, une seule réalisation de valeur extrême est conservée ; dans le deuxième cas de figure, le maximum n'est pas une réalisation de valeur extrême.

1.2.2 Analyse par les dépassements d'un seuil élevé

Une autre approche consiste à considérer les dépassements d'un seuil élevé (Leadbetter, 1983).

Théorème : Soit F une fonction de distribution de la variable aléatoire X, telle que F appartienne au domaine d'attraction de la loi GEV de paramètres de position, échelle et forme notés β, α, k. Alors pour un seuil $u \to \infty$, la distribution des dépassements $Y = X - u$, conditionnelle à $X > u$, est approximativement :

$$H(y) = \begin{cases} 1 - (1 - ky/\alpha_u)^{1/k} & y \in (0, \infty), \text{si } k < 0 \\ 1 - \exp(-y/\alpha_u) & y \in (0, \infty), \text{si } k = 0 \\ 1 - (1 - ky/\alpha_u)^{1/k} & y \in (0, \alpha_u/k), \text{si } k > 0 \end{cases} \tag{1.6}$$

avec

$$\alpha_u = \alpha - k(u - \beta). \tag{1.7}$$

La distribution H est appelée Generalized Pareto Distribution (GPD), de paramètre d'échelle α_u et de forme k. En particulier $\alpha_u + ku$ est une constante indépendante de u.

La méthode Peak Over Threshold (POT) (Beirlant *et al.*, 2004) consiste à modéliser les dépassements de seuil extraits d'un échantillon par une loi GPD.

La modélisation des dépassements de seuil par une loi GPD est aussi justifiée, dans un cadre plus général, par une caractérisation des dépassements de seuil avec un processus ponctuel (Beirlant *et al.*, 2004). Soit $X_1, \ldots, X_n$ un échantillon de variables aléatoires, de distribution F appartenant au domaine d'attraction de la loi GEV de paramètre de forme k. Soit le processus bi-dimensionnel $P_n = \{\frac{i}{n+1}, \frac{X_i - b_n}{a_n}; i = 1, \ldots, n\}$, où a_n, b_n normalisent $\max\{X_1, \ldots, X_n\}$. Alors le processus P_n converge en probabilité vers un processus de Poisson bi-dimensionnel, d'intensité Λ telle que $\Lambda((t_1, t_2) \times (x, \infty)) = (t_2 - t_1)(1 - kx)^{1/k}$.

Choix du seuil

Il existe différentes méthodes pour choisir le seuil u au dessus duquel les valeurs sont modélisées par une loi GPD. En effet, le seuil doit être suffisamment élevé pour satisfaire le caractère asymptotique du modèle, mais pas trop élevé pour garder un nombre suffisant de dépassements pour estimer les paramètres du modèle. Coles (2001) présente deux méthodes, qui ont été implémentées en langage R par Ribatet (2006) avec le package POT. La première méthode est exploratoire, et ne nécessite pas d'estimer les paramètres du modèle. La seconde méthode repose sur la stabilité des estimateurs des paramètres du modèle sur une gamme de seuils. Les deux méthodes reposent sur le fait que si la loi GPD est valide pour les dépassements d'un seuil u_0, avec des paramètres d'échelle et de forme notés α_{u_0}, k_{u_0}, alors la loi GPD est encore valide pour les dépassements de seuils $u > u_0$, et les paramètres vérifient $\alpha_u = \alpha_{u_0} - k(u - u_0)$ et $k_{u_0} = k_u$, d'après 1.7. Plus particulièrement, la première méthode est fondée sur la moyenne de la loi GPD :

$$E(X - u_0 | X > u_0) = \frac{\alpha_{u_0}}{1 + k_{u_0}}, \tag{1.8}$$

avec $k_{u_0} > -1$. Pour des dépassements d'un seuil $u > u_0$, on a

$$E(X - u|X > u) = \frac{\alpha_u}{1 + k_u} = \frac{\alpha_{u_0} - k(u - u_0)}{1 + k_{u_0}} \tag{1.9}$$

d'après 1.7. Ainsi, $E(X - u|X > u)$ est une fonction linéaire en u, et est estimée simplement par la moyenne empirique des dépassements de u. La recherche du seuil u_0 à partir duquel il est justifié de modéliser les dépassements par une loi GPD est donc réalisée graphiquement, par la recherche d'une linéarité dans la fonction $E(X - u|X > u)$ estimée empiriquement. Cette méthode est délicate car le graphe de d'estimation de $E(X - u|X > u)$ est souvent d'allure saccadée, et donc difficile à interpréter. La deuxième méthode proposée par Coles (2001) cherche u_0 de telle sorte que $\alpha^{modifié} = \alpha_u + k_u u$ et k_u soient constants à partir de $u > u_0$. En pratique, on considère les estimateurs $\hat{\alpha}^{modifié}, \hat{k}_u$ et leurs intervalles de confiance pour différents seuils u. C'est cette deuxième méthode que nous utiliserons dans la thèse pour chercher le seuil d'une loi GPD.

1.2.3 Estimation

Les paramètres des lois GEV et GPD peuvent être estimés par maximum de vraisemblance, moments pondérés (Greenwood *et al.*, 1979), méthode des moments (Christopeit, 1994) ou encore par des méthodes bayésiennes (Lye *et al.*, 1993). Hosking *et al.* (1985) et Hosking et Wallis (1987) comparent les estimateurs du maximum de vraisemblance et des moments pondérés et montrent que la méthode des moments pondérés donne de meilleurs résultats en terme d'estimation des paramètres et des quantiles, pour des tailles d'échantillon inférieures à 100 pour la loi GEV et 500 pour la loi GPD, avec un paramètre de forme $k < 0$ (c'est-à-dire une loi GEV ou GPD non bornée supérieurement). Il existe également des estimateurs non paramétriques : l'estimateur de Hill (1975) (pour le cas des distribution de type Fréchet), l'estimateur de Pickands (1975), l'estimateur de Dekkers-Einmahl-de Hann (Dekkers *et al.*, 1989). Ces estimateurs non paramétriques sont comparés par Embrechts *et al.* (1997) : la question du meilleur estimateur du paramètre de forme k n'a pas de réponse tranchée, et les propriétés des estimateurs varient en fonction de la valeur de k. L'estimateur de Pickands est en particulier difficile à utiliser car il présente une forte instabilité.

1.3 Théorie multivariée des valeurs extrêmes

L'analyse des processus extrêmes peut bénéficier d'informations fournies par d'autres processus. Le comportement des pluies extrêmes peut par exemple être mieux expliqué par l'étude simultanée des pluies de différentes durées, ou des pluies de différents postes pluviographiques voisins. L'intérêt de l'analyse multivariée des processus est donc de mieux représenter la complexité des phénomènes étudiés et d'éclairer chaque processus marginal par l'apport des autres processus.

Pour simplifier, et puisque les modèles de valeurs extrêmes présentés dans la thèse sont uniquement bi-variés, on présente seulement le cas bi-varié. On travaille donc ici avec un échantillon bi-varié $(X_1, Y_1), \ldots, (X_n, Y_n)$ de fonction de distribution $F(x, y)$. De même que dans le cas univarié, les valeurs extrêmes peuvent être étudiées par les maxima par bloc, par les dépassements de seuil, ou par un processus ponctuel. On s'intéressera ici au processus $(\max\{X_i\}, i = 1, \ldots, n; \max\{Y_i\}, i = 1, \ldots, n)$ des maxima par bloc. La théorie des valeurs

extrêmes s'applique donc aux distributions marginales. Pour simplifier les notations, on notera de la même manière la distribution bi-variée et la distribution marginale. On suppose que les variables X_i, Y_i suivent une loi de Fréchet standard, de distribution $F(z) = \exp(-1/z)$. Ceci n'est pas une perte de généralité puisque, si nécessaire, on peut se rapporter à une loi de Fréchet standard par le changement de variable $Z = -\frac{1}{\log F(X)}$. En effet,

$$P(Z \leq z) = P(F(X) \leq \exp^{-1/z}) = \exp^{-1/z} \tag{1.10}$$

car $F(X)$ suit une loi uniforme sur [0,1].

1.3.1 Lois bi-variées extrêmes

Dans le cadre présenté précédemment, on a le théorème :

Théorème : Soit $(X_1, Y_1), \ldots, (X_n, Y_n)$ un échantillon de variables i.i.d., de loi marginale Fréchet standard. Si

$$P(\max\{X_i\}/n \leq x, \max\{Y_i\}/n \leq y) \to G(x, y), \tag{1.11}$$

lorsque $n \to \infty$, où G est une fonction de distribution non-dégénérée, alors G est de la forme

$$G(x, y) = \exp(-V(x, y)), \tag{1.12}$$

$x > 0, y > 0$ où

$$V(x, y) = 2 \int_0^1 \max(\frac{w}{x}, \frac{1-w}{y}) dH(w), \tag{1.13}$$

et H est une fonction de distribution sur [0,1] satisfaisant la contrainte

$$\int_0^1 w dH(w) = 1/2. \tag{1.14}$$

Les fonctions limites G de l'équation 1.11 sont appelées distributions extrêmes bi-variées. L'ensemble de ces fonctions est infini, et ne possède pas de paramétrisation finie. Huit familles principales ont été identifiées :

- La famille logistique, avec un paramètre $0 < \Phi \leq 1$.

$$G(x, y) = \exp(-(x^{-1/\Phi} + y^{-1/\Phi})^{\Phi}), x > 0, y > 0. \tag{1.15}$$

  Si $\Phi \to 0, G(x, y) = \exp(-\max(x^{-1}, y^{-1}))$: ce cas correspond au cas de variables parfaitement dépendantes. Si $\Phi \to 1, G(x, y) = \exp(-(x^{-1} + y^{-1}))$: ce cas correspond au cas de variables indépendantes.
- La famille logistique asymétrique généralise la famille logistique, avec les paramètres $0 < \Phi \leq 1, \Psi_i \in [0, 1]$.

$$G(x, y) = \exp\{-(1-\Psi_1)/x - (1-\Psi_2)/y - ((\Psi_1/x)^{1/\Phi} + (\Psi_2/y)^{1/\Phi})^{\Phi}\}, \tag{1.16}$$

  $x > 0, y > 0$. On a indépendance si $\Phi = 1$ ou $\Psi_1 = 0$ ou $\Psi_2 = 0$. On a dépendance complète si $\Psi_1 = \Psi_2 = 1$ et $\Phi \to 0$.

– La distribution de Husler-Reiss (Husler et Reiss, 1989) :

$$G(x,y) = \exp\{-x\phi(\Phi^{-1} + \Phi\log(x/y)/2) - y\phi(\Phi^{-1} + \Phi\log(y/x)/2)\} \tag{1.17}$$

avec $\Phi > 0$ le paramètre de dépendance, et $\phi(.)$ la distribution de probabilité de la loi normale centrée réduite. Le cas d'indépendance est obtenu pour $\Phi \to 0$, et le cas de dépendance complète est obtenu pour $\Phi \to \infty$.

– La distribution bi-variée logistique négative (Galambos, 1975) :

$$G(x,y) = \exp\{-x - y + [x^{-\Phi} + y^{-\Phi}]^{-1/\Phi}\} \tag{1.18}$$

avec $\Phi > 0$. C'est un cas particulier de la distribution bi-variée logistique négative asymétrique. L'indépendance est obtenue pour $\Phi \to 0$, et la dépendance complète pour $\Phi \to \infty$.

– La distribution bi-variée logistique négative asymétrique (Joe, 1990) :

$$G(x,y) = \exp\{-x - y + [(\Psi_1 x)^{-\Phi} + (\Psi_2 y)^{-\Phi}]^{-1/\Phi}\} \tag{1.19}$$

avec $\Phi > 0, 0 < \Psi_1, \Psi_2 \leq 1$. $\Psi_1 = \Psi_2 = 1$ correspond au cas de la distribution bi-variée logistique négative. L'indépendance est obtenue à la limite si Φ, Ψ_1, ou Ψ_2 tend vers 0. La dépendance complète est obtenue à la limite si $\Psi_1 = \Psi_2 = 1$ et $\Phi \to \infty$.

– La distribution bilogistique, de paramètres Φ_1, Φ_2 (Smith, 1990) :

$$G(x,y) = \exp\{-xq^{1-\Phi_1} - y(1-q)^{1-\Phi_2}\} \tag{1.20}$$

où $q = q(x, y; \Phi_1, \Phi_2)$ est la racine de l'équation :

$$(1-\Phi_1)x(1-q)^{\Phi_2} - (1-\Phi_2)yq^{\Phi_1} = 0, \tag{1.21}$$

avec $0 < \Phi_1, \Phi_2 < 1$. Si $\Phi_1 = \Phi_2$, le modèle bilogistique est équivalent au modèle logistique de paramètre de dépendance $\Phi_1 = \Phi_2$. L'indépendance est obtenue si $\Phi_1 = \Phi_2$ et approchent 1, ou lorsque l'un des deux paramètres Φ_1, Φ_2 est fixé, et le second tend vers 1. La dépendance complète est obtenue si $\Phi_1 = \Phi_2$ et approchent 0.

– La distribution bilogistique négative, de paramètres Φ_1, Φ_2 (Coles et Tawn, 1994) :

$$G(x,y) = \exp\{-x - y + xq^{1+\Phi_1} + y(1-q)^{1+\Phi_2}\} \tag{1.22}$$

où $q = q(x, y; \Phi_1, \Phi_2)$ est la racine de l'équation :

$$(1+\Phi_1)xq^{\Phi_1} - (1+\Phi_2)y(1-q)^{\Phi_2} = 0, \tag{1.23}$$

avec $\Phi_1 > 0, \Phi_2 > 0$. Si $\Phi_1 = \Phi_2$, le modèle bilogistique négatif est équivalent au modèle logistique, de paramètre $\Phi = 1/\Phi_1 = 1/\Phi_2$. L'indépendance est obtenue si $\Phi_1 = \Phi_2$ tendent vers ∞, ou si Φ_1 ou Φ_2 est fixé et l'autre tend vers ∞. La dépendance complète est obtenue à la limite si $\Phi_1 = \Phi_2$ approchent 0.

– La distribution proposée par Coles et Tawn (1991)

$$G(x,y) = \exp\{-x[1 - B(q; \Phi_1 + 1, \Phi_2)] - yB(q; \Phi_1, \Phi_2 + 1)\} \tag{1.24}$$

où $q = \Phi_1 y/(\Phi_1 y + \Phi_2 x)$ et $B(q; \Phi_1, \Phi_2)$ est la distribution Beta, évaluée en q de paramètres (Φ_1, Φ_2). L'indépendance est obtenue si $\Phi_1 = \Phi_2$ tendent vers 0, ou si l'un des deux paramètres est fixé et le second tend vers 0. La dépendance complète est obtenue à la limite si $\Phi_1 = \Phi_2$ tendent vers ∞.

La structure de dépendance de toute distribution bi-variée extrême G peut être décrite de diverses manières. Une méthode souvent utilisée est l'étude de la fonction de dépendance de Pickands A, satisfaisant certaines propriétés (Beirlant *et al.*, 2004). La fonction de dépendance de Pickands $A(t)$ est définie pour $t \in [0,1]$, par

$$A(t) = -\ln(G[G_1^{-1}(\exp(-1+t)), G_2^{-1}(\exp(-t))]), \tag{1.25}$$

où G_1, G_2 sont les deux lois marginales. Une distribution bi-variée extrême G est complètement déterminée par ses marginales G_1, G_2 et sa fonction de dépendance de Pickands, via l'équation 1.25. A peut être estimée par des méthodes non paramétriques (Pickands, 1981), (Pickands, 1989), (Capéraà *et al.*, 1997), ou par des méthodes paramétriques (maximum de vraisemblance par exemple), en supposant un modèle paramétrique pour G. La comparaison des estimateurs paramétriques et non paramétriques de A permet de valider un modèle paramétrique particulier.

Par exemple, la fonction de distribution bi-variée asymétrique mélangée, de paramètres Φ_1, Φ_2 a une fonction de dépendance de Pickands définie par :

$$A(t) = 1 - (\Phi_1 + \Phi_2)t + \Phi_1 t^2 + \Phi_2 t^3 \tag{1.26}$$

avec $\Phi_1 > 0, \Phi_1 + 3\Phi_2 > 0, \Phi_1 + \Phi_2 \leq 1, \Phi_1 + 2\Phi_2 \leq 1$ (Tawn, 1988). Ces contraintes impliquent que $\Phi_1 \in [0, 1.5], \Phi_2 \in [-0.5, 0.5]$. Le degré de dépendance augmente avec Φ_1, à Φ_2 fixé. La dépendance complète ne peut pas être atteinte, l'indépendance est obtenue si les deux paramètres sont nuls.

1.3.2 Autre formalisation des lois bi-variées

Nous verrons dans la section suivante que les lois bi-variées extrêmes sont asymptotiquement dépendantes, sauf dans le cas de l'indépendance. Cette hypothèse ne convient pas pour modéliser les processus asymptotiquement indépendants. Ledford et Tawn (1996, 1997) ont introduit la modélisation suivante pour la loi jointe d'un couple (X, Y) de mêmes marginales :

$$P(X > x, Y > y) \approx \mathcal{L}(\frac{-1}{\log F(x)}, \frac{-1}{\log F(y)})(-\log F(x))^{c_1}(-\log F(y))^{c_2}, \tag{1.27}$$

avec $c_i > 0, c_1 + c_2 \geq 1$, et $\mathcal{L}$ une fonction à variation lente bi-dimensionnelle. Dans le cas de marginales Fréchet standard, on a :

$$P(X > x, Y > y) \approx \mathcal{L}(x, y)x^{-c_1}y^{-c_2} \tag{1.28}$$

L'intérêt de cette formulation est qu'elle permet de modéliser des processus asymptotiquement dépendants, mais également des processus asymptotiquement indépendants. Le paramètre $0 < \eta = 1/(c_1 + c_2) \leq 1$ mesure la dépendance entre les queues de distributions marginales. La distribution 1.27 vérifie les propriétés suivantes :

- Si $0 < \eta < 1$ alors les variables sont asymptotiquement indépendantes.
- Si $1/2 < \eta \leq 1$, les variables marginales sont positivement associées : la probabilité qu'une variable soit extrême est plus forte si la seconde variable est extrême, que sans

condition sur la seconde variable. De même, si $0 < \eta < 1/2$, elles sont négativement associées. En effet, soit $z \to \infty$, et $w = F(z) \to 1$, on a :

$$\begin{aligned} P(Y > z|X > z)/P(X > z) &= \mathcal{L}(-\frac{1}{\log F(z)}, -\frac{1}{\log F(z)})\frac{(-\log F(z))^{1/\eta}}{\bar{F}(z)^2} \\ &= \mathcal{L}(-\frac{1}{\log w}, -\frac{1}{\log w})\frac{(-\log w)^{1/\eta}}{(1-w)^2} \\ &\approx \mathcal{L}(\frac{1}{1-w}, \frac{1}{1-w})(1-w)^{1/\eta-2}. \end{aligned} \tag{1.29}$$

Or si $\eta < 1/2$ et si l'on considère le logarithme de l'expression ci-dessus, que l'on divise par $\log(\frac{1}{1-w}) > 0$, on a[1] :

$$\log(\mathcal{L}(\frac{1}{1-w}, \frac{1}{1-u}))/\log(\frac{1}{1-w}) - (1/\eta - 2) \to -(1/\eta - 2) < 0. \tag{1.30}$$

Donc $P(Y > z|X > z)/P(X > z) < 1$ si $0 < \eta < 1/2$ et le contraire si $1/2 < \eta < 1$.

- Si les variables marginales sont indépendantes, alors $\eta = 1/2$.
- Les variables sont asymptotiquement dépendantes si et seulement si $\eta = 1$ et $\mathcal{L}(x) \to c \neq 0$ quand $x \to \infty$.

Pour l'estimation de η, on peut utiliser les techniques uni-variées : en notant $T = \min(X, Y)$, on a (dans le cas Fréchet standard, auquel on peut se ramener par changement de variables) : $P(T > x) = P(X > x, Y > x) \approx \mathcal{L}(x)x^{-1/\eta}$, avec $\mathcal{L}$ fonction à variation lente uni-variée. η est le paramètre de forme d'une distribution uni-variée, et peut donc être estimé par l'estimateur de Hill (1975), ou d'autres méthodes référencées dans (Embrechts *et al.*, 1997).

1.3.3 Autres paramétrisations de la dépendance des extrêmes

On considère un couple de variables aléatoires (X, Y), de loi jointe $F(x, y)$, de loi marginale $F(x)$ (on suppose que X et Y sont de mêmes lois marginales). Dans cette section, certains des résultats ne sont vrais que si la loi marginale est une loi de Fréchet standard. Cette restriction n'est pas une perte de généralité puisque la transformation $Z = -\frac{1}{\log F(X)}$ permet de se ramener à une loi de Fréchet standard. Un grand nombre des résultats présentés ci-dessous sont repris de (Bacro, 2005).

Le coefficient χ

Joe (1993) a introduit la mesure χ :

$$\chi = \lim_{x \to x^*} P(Y > x|X > x), \tag{1.31}$$

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

χ mesure la relation de dépendance asymptotique entre deux variables, et s'appele coefficient de dépendance de queue. Plus précisément, $\chi \in [0, 1]$ est la probabilité que, si l'une des variables est extrême, l'autre variable le soit aussi.

[1] d'après la formule de Resnick (1987) : si $\mathcal{L}$ est une fonction à variation lente, $\log \mathcal{L}(x)/\log(x) \to 0$ si $x \to \infty$.

Dans le cas de l'indépendance stricte (c'est-à-dire X et Y indépendants), $\chi = 0$. Cependant, χ peut être nul sans que X et Y soient indépendants. Dans ce cas, X et Y sont asymptotiquement indépendants.

Si $\chi > 0$, les X et Y sont asymptotiquement dépendants, et le degré de dépendance augmente avec χ. Dans le cas de la dépendance totale, $\chi = 1$.

Partant de (X, Y), on se ramène à (U, V) de même lois marginales uniformes par la transformation $U = F(X), V = F(Y)$. La structure de dépendance de (X, Y) est conservée dans celle de (U, V) : en effet,

$$P(X \leq x, Y \leq y) = P(U \leq F(x), V \leq F(y)). \tag{1.32}$$

On a donc

$$\chi = \lim_{u \to 1} P(V > u | U > u). \tag{1.33}$$

Or,

$$\begin{array}{rcl} P(V > u|U > u) & = & \frac{P(U>u)-(P(V<u)-P(U<u,V<u))}{P(U>u)} \\ & = & 2 - \frac{1-P(U<u,V<u)}{P(U>u)} \\ & \approx & 2 - \log(P(U < u, V < u))/\log(P(U < u)) \end{array} \tag{1.34}$$

On pose $\chi(u) = 2 - \log(P(U < u, V < u))/\log(P(U < u))$ et on a finalement :

$$\chi = \lim_{u \to 1} \chi(u). \tag{1.35}$$

La fonction $\chi(u)$ et le paramètre χ vérifient certaines propriétés :

- Dans le cas de la dépendance totale, $\chi(u) = 1, \forall u \in [0, 1]$. Dans le cas de l'indépendance stricte $\chi(u) = 0, \forall u \in [0, 1]$.
- Pour une distribution bi-variée extrême, $\chi(u) = 2 - V(1, 1)$ est constant en u (pour u assez grand) (Coles, 2001), et $\chi > 0$, sauf dans le cas de l'indépendance stricte. En d'autre termes, la notion d'indépendance asymptotique n'existe pas dans le cas des distributions bi-variées extrêmes. En effet,

$$\begin{array}{rl} 2 - V(1,1) & = 2(1 - \int_0^1 \max(w, 1-w) dH(w)) = \\ & = 2 \int_0^1 \min(w, 1-w) dH(w) > 0. \end{array} \tag{1.36}$$

- Dans le cas d'une distribution bi-variée donnée par l'équation 1.27, si X et Y sont asymptotiquement dépendants, alors $\mathcal{L}(z, z) \to c > 0$ lorsque $z \to \infty$ et $\eta = 1$, et l'on a $c = \chi$ (Coles *et al.*, 2002).
- La fonction $\chi(u)$ est en général complexe. Par exemple, un couple gaussien de corrélation $0 < \rho < 1$ est asymptotiquement indépendant. Or, Coles *et al.* (2002) montrent que $\chi(u)$ augmente avec ρ, mais que lorsque $u \to 1$, $\chi(u)$ converge vers 0 de manière lente et asymptotiquement abrupte. Cet exemple montre que sur la base d'un graphique de $\chi(u)$, estimé empiriquement, il est possible de conclure à tort que des extrêmes sont asymptotiquement dépendants. Cette remarque est importante puisque nous serons amenés à choisir des modèles pour des processus asymptotiquement indépendants et des processus asymptotiquement dépendants. Par exemple, les modèles d'extrêmes multivariés ne sont pas adaptés au cas de variables asymptotiquement indépendantes, et reflètent donc mal le comportement des extrêmes multivariés en dehors d'une dépendance asymptotique.

χ est donc une mesure de la dépendance asymptotique, intéressante si les variables sont asymptotiquement dépendantes. Elle n'est pas suffisante pour discriminer les variables asymptotiquement indépendantes. C'est pourquoi on introduit une autre mesure $\bar{\chi}$, qui mesure le degré de dépendance pour des variables asymptotiquement indépendantes (Coles *et al.*, 2002).

Le coefficient $\bar{\chi}$

Pour $0 \leq u \leq 1$, la fonction $\bar{\chi}(u)$ est définie par :

$$\bar{\chi}(u) = \frac{2 \log P(U > u)}{\log P(U > u, V > u)} - 1, \tag{1.37}$$

où (U, V) sont les lois uniformes du paragraphe précédent sur χ. La fonction $\bar{\chi}(u)$ vérifie certaines propriétés :

- $\forall u \in [0, 1], -1 < \bar{\chi}(u) \leq 1$ (-1 est impossible, par la définition de $\bar{\chi}(u)$).
- Si, au delà de u, on a indépendance stricte, alors $\bar{\chi}(u) = 0$, ou si on a dépendance totale, $\bar{\chi}(u) = 1$.
- Si $0 < \bar{\chi}(u) < 1$, alors $P(V > u|U > u) > P(V > u)$: les extrêmes sont associés positivement. Cela signifie qu'on a une plus grande probabilité de dépassement de l'une des variables lorsque l'autre est extrême, que sous l'indépendance.
  Si $-1 < \bar{\chi}(u) < 0$, $P(V > u|U > u) < P(V > u)$: les extrêmes sont associés négativement. En effet, Bacro (2005) montre que

$$P(V > u|U > u) = P(U > u)^{\frac{1-\bar{\chi}(u)}{1+\bar{\chi}(u)}}. \tag{1.38}$$

- On déduit également de 1.38 que $|\bar{\chi}(u)|$ augmente avec la dépendance.

Par analogie avec χ, on définit :

$$\bar{\chi} = \lim_{u \to 1} \bar{\chi}(u). \tag{1.39}$$

Le coefficient $\bar{\chi}$ complète l'information donnée par χ. Dans le cas de l'indépendance asymptotique ($\chi = 0$), $\bar{\chi}$ permet de mesurer le degré de dépendance. Par exemple, dans le cas d'un couple gaussien de corrélation ρ, $\bar{\chi} = \rho$, et $\bar{\chi}(u)$ est approximativement linéaire pour $0.5 < u < 1$: on peut alors conclure à l'indépendance asymptotique, contrairement à ce qu'aurait pu laisser croire une interprétation directe de χ.

$\bar{\chi}$ vérifie les propriétés suivantes :

- $-1 < \bar{\chi} \leq 1$,
- le cas $\bar{\chi} = 1$ correspond à la dépendance asymptotique de X et Y,
- et le cas $\bar{\chi} \in]-1, 1[$ correspond à l'indépendance asymptotique de X et Y (Coles, 2001).

En résumé, χ, $\bar{\chi}$

Le couple $(\chi, \bar{\chi})$ permet de caractériser la dépendance des extrêmes :

$$\begin{array}{ll} \chi \in [0, 1], & \text{et le cas } \chi \in]0, 1] \text{ correspond à la dépendance asymptotique} \\ \bar{\chi} \in]-1, 1], & \text{et le cas } \bar{\chi} \in]-1, 1[\text{ correspond à l'indépendance asymptotique.} \end{array} \tag{1.40}$$

Ainsi,

– le cas ($\chi \in]0,1], \bar{\chi} = 1$) correspond à la dépendance asymptotique, et χ détermine le degré de la dépendance ;
– le cas ($\chi = 0, \bar{\chi} \in]-1,1[$) correspond à l'indépendance asymptotique, et $\bar{\chi}(u)$ donne une mesure du degré de dépendance à des niveaux u élevés.

On peut utiliser les mesures $\chi(u), \bar{\chi}(u)$ de manière exploratoire. En pratique, ces coefficients sont estimés empiriquement, en se ramenant à des lois uniformes.

Coles *et al.* (2002) donnent d'autres méthodes pour estimer χ et $\bar{\chi}$. Par exemple dans le cas des lois bi-variées extrêmes, $\chi = 2 - V(1,1)$ et $\chi > 0$ sauf dans le cas de l'indépendance stricte, l'estimation de $V(1,1)$ permet donc d'estimer χ. La formalisation des lois bi-variées présentée dans la section précédente introduit un paramètre de dépendance de queue η. On peut montrer une relation entre η et $\bar{\chi}$ (Bacro, 2005). En reprenant les notations de 1.27, on a pour x grand :

$$P(X > x, Y > x) \approx \mathcal{L}(x) P(X > x)^{1/\eta}, \tag{1.41}$$

où $\mathcal{L}$ est une fonction à variation lente en dimension 1, et $0 < \eta \leq 1$. D'autre part, si x est grand, on a aussi $P(X > x) = 1 - \exp(-1/x) \approx 1/x$. Alors

$$\bar{\chi}(u) \approx \frac{2\log(1-u)}{\log \mathcal{L}(1-u)^{-1} + \log(1-u)/\eta} - 1 \to 2\eta - 1,$$

d'après le résultat suivant (Resnick, 1987) : si $\mathcal{L}$ est une fonction à variation lente, $\log \mathcal{L}(x)/\log(x) \to 0$ si $x \to \infty$. On a donc

$$\bar{\chi} = 2\eta - 1. \tag{1.42}$$

L'estimation de η assure alors celle de $\bar{\chi}$.

Enfin, si X et Y sont asymptotiquement dépendantes, on a $\mathcal{L}(z,z) \to \chi$ lorsque $z \to \infty$ et $\bar{\chi} = 1$.

1.4 Théorie des valeurs extrêmes d'une série temporelle stationnaire

En général, sur des séries temporelles telles que les pluies journalières, l'hypothèse d'indépendance dans un échantillon n'est plus vérifiée. En particulier, les conditions extrêmes peuvent persister sur plusieurs observations consécutives. On généralise le cas des séries i.i.d. par le cas des séries stationnaires.

Définition : (stationnarité stricte)

$X_1, X_2, \ldots$ est une série stationnaire si pour tout j, k, l, $X_j, \ldots, X_{j+k}$ a la même loi que $X_{l+j}, \ldots, X_{l+j+k}$.

On fera attention à ne pas confondre cette définition avec celle de la stationnarité physique d'une série temporelle (stationnarité d'ordre deux), dans le contexte des changements climatiques par exemple. Une série temporelle de réalisations d'une grandeur aléatoire, à un pas de temps donné, est dite stationnaire d'ordre deux si ses réalisations sont issues d'un même processus aléatoire dont les paramètres (moyenne, variance, asymétrie, auto-corrélation...) restent constants au cours du temps.

1.4.1 Indice extrémal

La convergence en loi des maxima par blocs est encore vérifiée pour une série stationnaire, si celle-ci satisfait une

Condition $\mathcal{D}(u_n)$ de non-dépendance à long terme : (Leadbetter, 1974) Une série stationnaire $X_1, X_2, \ldots$ satisfait la condition $\mathcal{D}(u_n)$ au seuil u_n si, pour tous entiers

$i_1 < \ldots < i_p < j_1 < \ldots < j_q$ avec $j_1 - i_p > l_n$,

$$|P(X_{i_1} \leq u_n, \ldots, X_{i_p} \leq u_n, X_{j_1} \leq u_n, \ldots, X_{j_q} \leq u_n) - P(X_{i_1} \leq u_n, \ldots, X_{i_p} \leq u_n)P(X_{j_1} \leq u_n, \ldots, X_{j_q} \leq u_n)| \leq \alpha(n, l_n), \quad (1.43)$$

avec $\alpha(n, l_n) \to 0$ pour une suite l_n telle que $l_n/n \to 0$ quand $n \to \infty$.

Un théorème de Leadbetter (1974) montre que, si une série stationnaire vérifie d'une part la relation limite 1.4, et d'autre part la condition $\mathcal{D}(u_n)$ avec $u_n = a_n x + b_n$, a_n, b_n étant les suites de la relation 1.4, pour tout x tel que $G(x) > 0$, alors G est encore une distribution des valeurs extrêmes. Ce résultat implique que si une série stationnaire vérifie la condition $\mathcal{D}(u_n)$ dans les extrêmes, alors les maxima par bloc suivent le même type de distribution limite que dans le cas d'une série indépendante. Néanmoins, les paramètres de la distribution limite sont affectés par la dépendance dans la série. C'est ce qu'exprime le :

théorème (Leadbetter, 1983) : Soit $X_1, X_2, \ldots$ un processus stationnaire et $X_1^*, X_2^*, \ldots$ une suite de variables aléatoires i.i.d. de même distribution marginale que le processus stationnaire. On définit $M_n = \max\{X_1, \ldots, X_n\}$ et $M_n^* = \max\{X_1^*, \ldots, X_n^*\}$. S'il existe des constantes $a_n > 0$, $b_n \in \mathbb{R}$ et une distribution non-dégénérée telle que

$$P(\frac{M_n^* - b_n}{a_n} \leq x) \to G_1(x), \quad (1.44)$$

quand $n \to \infty$ et si le processus stationnaire vérifie d'une part la condition $\mathcal{D}(u_n)$ avec $u_n = a_n x + b_n$ pour tout x tel que $G_1(x) > 0$, et d'autre part

$$P(\frac{M_n - b_n}{a_n} \leq x) \to G_2(x), \quad (1.45)$$

quand $n \to \infty$, alors $G_2(x) = G_1(x)^{IE}$ où $0 \leq IE \leq 1$ est appelé indice extrémal.

Le cas $IE = 0$ correspond à un cas pathologique (Leadbetter, 1983). On ne s'intéressera qu'au cas où $0 < IE \leq 1$. Les paramètres de position, échelle et forme des distributions G_1, G_2 sont donc liés par :

$$\beta_1 = \beta_2 + \alpha_2(1 - IE^{-k_2})/k_2 \quad (1.46)$$

$$\alpha_1 = \alpha_2 IE^{-k_2} \quad (1.47)$$

$$k_1 = k_2 \quad (1.48)$$

si $k_1 \neq 0$, ou

$$\beta_1 = \beta_2 + \alpha_2 \log IE \quad (1.49)$$

$$\alpha_1 = \alpha_2 \quad (1.50)$$

$$k_1 = k_2 = 0 \quad (1.51)$$

si $k_1 = 0$.

Le paramètre *IE* donne une mesure de la dépendance à court terme des extrêmes. Il indique la tendance des extrêmes à se regrouper en clusters. Il n'existe pas de méthode spécifique pour identifier les clusters, et l'identification des clusters peut donc être assez arbitraire. Pour représenter un cluster, prenons un échantillon $X_1, \ldots, X_n$ de taille n d'une série stationnaire de loi F, r_n une suite d'entiers avec $r_n = o(n)$, et u_n une suite de seuils réels tel que $n\bar{F}(u_n) = O(1)$.

- L'identification des clusters par bloc consiste à partager les données en blocs de longueur r_n. On suppose alors que les dépassements de u_n appartiennent au même cluster si elles appartiennent au même bloc.
- Une autre manière, plus stable, consiste à considérer que des dépassements séparés de moins de r_n dépassements appartiennent au même cluster.

Soit $N_{r_n}(u_n) = \sum_{i=1}^{r_n} \mathbb{1}(X_i > u_n)$. La distribution π_n de la taille d'un cluster est définie par

$$\pi_n(j; u_n, r_n) = P\{N_{r_n}(u_n) = j | N_{r_n}(u_n) > 0\}, \tag{1.52}$$

pour $j = 1, \ldots, r_n$, et $\mathbb{1}(.)$ est la fonction indicatrice. Sous certaines conditions de sommabilité sur π_n, données par Hsing *et al.* (1988), on a

$$IE^{-1} = \lim_{n\to\infty} \sum_{i=1}^{r_n} j\pi_n(j; u_n, r_n), \tag{1.53}$$

c'est-à-dire la limite de la taille moyenne des clusters.

D'autre part, O'Brien (1987) montre que $\theta(u_n, r_n)$ défini par

$$\theta(u_n, r_n) = P(\max\{X_2, \ldots, X_{r_n}\} < u_n | X_1 > u_n) \tag{1.54}$$

vérifie $\lim_{n\to\infty} \theta(u_n, r_n) = IE$ avec r_n une suite d'entiers satisfaisant $r_n = o(u_n)$ et $u_n \to \infty$.

Pour des variables aléatoires $X_1, X_2, \ldots$, indépendantes, $IE = 1$, mais *IE* peut être égal à 1, sans que les variables aléatoires soient indépendantes. Dans ce cas, il s'agit de processus asymptotiquement indépendants. Plus *IE* est proche de 1, plus la dépendance est faible.

Il existe d'autres méthodes pour estimer l'indice extrémal *IE*, quelques une sont présentées dans (Beirlant *et al.*, 2004). Ancona-Navarrete et Tawn (2000) comparent différentes méthodes. Nous présentons ici l'estimateur utilisé dans la thèse (appelé 'runs estimator' dans la littérature anglaise), choisi d'après les études de biais asymptotiques, réalisés par Smith et Weissman (1994), Weissman et Novak (1998) :

$$\widehat{\theta}(u_n, r_n) = \frac{(n - r_n)^{-1} \sum_{i=1}^{n-r_n} \mathbb{1}(X_i > u_n, M_{i,i+r_n} \leq u_n)}{n^{-1} \sum_{i=1}^{n} \mathbb{1}(X_i > u_n)}, \tag{1.55}$$

avec $M_{i,i+r} = \max\{X_i, \ldots, X_{i+r}\}$, ce que l'on peut écrire aussi :

$$\widehat{\theta}(u_n, r_n) = \frac{(n - r_n)^{-1} n_c(u_n, r_n)}{n^{-1} n_u(u_n)}, \tag{1.56}$$

où $n_c(u_n, r_n)$ est le nombre de clusters, et $n_u(u_n)$ est le nombre de dépassements du seuil u_n. Cet estimateur est l'estimateur empirique de $\theta(u_n, r_n)$. Il permet de donner une représentation pratique des clusters : un cluster commence lorsque une valeur dépasse le seuil u_n, et se termine lorsque r_n valeurs consécutives sont sous le seuil u_n. Cette description des clusters correspond à la seconde méthode d'identification des clusters présentée plus haut.

Enfin, l'utilisation des processus ponctuels fournit une méthode élégante pour décrire le phénomène de clusterisation des extrêmes des processus stationnaires. Soit $\{X_i\}$ un processus stationnaire de loi marginale F, et d'indice extrémal $IE > 0$. On suppose que le processus vérifie une condition de non-dépendance à long terme (voir (Beirlant *et al.*, 2004) pour plus de précision sur cette condition), pour des seuils u_n vérifiant $n\bar{F}(u_n) \to \tau \in (0, \infty)$. Hsing *et al.* (1988) montrent que si la distribution π_n de la taille des clusters converge vers une distribution π, alors le processus d'arrivée des clusters, associé au processus de la taille des clusters converge en distribution vers un processus de Poisson composé d'intensité $IE \cdot \tau$ et la taille des clusters converge en loi vers la taille limite des clusters, de loi π.

En conséquence du théorème de Hsing *et al.* (1988), un intervalle de temps normalisé $(0, 1]$ compte en moyenne $IE \cdot \tau$ clusters, de tailles indépendantes et de loi π. La taille des clusters est en moyenne égale à $1/IE$.

De même, Leadbetter (1991) montre que sous des conditions semblables à celles du théorème de Hsing *et al.* (1988), le processus d'arrivée des clusters, associé au processus du pic maximum du cluster, converge vers un processus de Poisson composé d'intensité $IE \cdot \tau$ et dont les maxima des pics des clusters suivent une loi GPD. Le théorème de Leadbetter (1991) donne le fondement mathématique de la méthode POT (Peak over Threshold).

1.4.2 Approches par chaînes de Markov

Jusqu'à présent, la structure de dépendance entre les variables successives d'une série stationnaire n'a été décrite que par l'hypothèse de non-dépendance à long terme, et le degré IE de dépendance à court terme. Les chaînes de Markov sont un modèle intuitif pour représenter la dépendance temporelle de la série, et donc la structure interne des clusters. On s'intéresse ici à des chaînes de Markov d'ordre 1. On note $F(x_1, x_2), f(x_1, x_2)$ les fonctions jointes de répartition et de densité de (X_1, X_2), deux variables consécutives de la chaîne de Markov. Pour simplifier les notations, on note de la même manière les fonctions de densité bi-variées et marginales. Les fonctions de répartition et de densité marginales sont notées $F(x), f(x)$. Alors, la densité jointe d'un échantillon est

$$f(x_1, \ldots, x_n) = f(x_1) \prod_{i=2}^{n} f(x_i|x_{i-1}) = \prod_{i=2}^{n} f(x_{i-1}, x_i) / \prod_{i=2}^{n-1} f(x_i). \tag{1.57}$$

On suppose que la distribution marginale appartient au domaine d'attraction d'une loi GEV. Alors la distribution des dépassements d'un seuil u élevé est modélisée par une distribution GPD : $P(X > x|X > u) = (1 - k(x-u)/\alpha)^{1/k}$. Un calcul simple montre que

$$F(x) = 1 - \lambda(1 - k(x-u)/\alpha)_+^{1/k}, \tag{1.58}$$

pour $x > u$, $\lambda = P(X > u)$, et $(x)_+ = \max(x, 0)$.

On modélise les extrêmes de la chaîne par une hypothèse sur la loi jointe $f(x_1, x_2)$. On peut supposer, comme Smith *et al.* (1997), que la loi jointe bi-variée est dans le domaine d'attraction d'une distribution bi-variée extrême. Dans ce cas, plusieurs auteurs Coles et Tawn (1991), Joe *et al.* (1992), Ledford et Tawn (1996) montrent l'approximation :

$$F(x, y) \approx \exp(-V[-1/\log(F(x)), -1/\log(F(y))]), x > u, y > u, \tag{1.59}$$

V est une fonction définie par la relation 1.13. Dans le cas d'une dépendance par loi bi-variée extrême, on rappelle que la chaîne de Markov est asymptotiquement dépendante (sauf cas d'indépendance). De plus, la dépendance entre dépassements consécutifs, mesurée par χ, est constante avec le seuil. Cette modélisation est donc seulement adaptée au cas des processus asymptotiquement dépendants. Un exemple classique est la distribution logistique avec $V(x, y) = ((-1/x)^{1/\Phi} + (-1/y)^{1/\Phi})^{\Phi}$.

Un autre modèle, proposé par Ledford et Tawn (1996, 1997), avec les fonctions à variation lente, est cette fois adapté au cas des processus asymptotiquement indépendants. On modélise la distribution jointe par l'équation 1.27 pour $x, y > u$.

Exemples de lois bi-variées avec indépendance asymptotique pour la loi de transition de la chaîne de Markov

Différents modèles, avec indépendance asymptotique, ont été proposés par Ledford et Tawn (1997), Bortot et Tawn (1998), Sisson et Coles (2003), Ledford et Tawn (2003) et Heffernan et Resnick (2005). Nous en présentons ci-dessous quelques uns.

1. Bortot et Tawn (1998) et Ancona-Navarrete et Tawn (2000) particularisent le modèle de Ledford et Tawn (1997), donné par l'équation 1.28 avec $\mathcal{L}$ une fonction bi-variée à variation lente définie par

$$\mathcal{L}(x, y) = a + \frac{x + y}{(xy)^{1/2}}\{1 - \frac{xy}{x + y}V(x, y)\} \tag{1.60}$$

   et V est une fonction bi-variée définie dans l'équation 1.12. Par exemple, si V est donné par la loi bi-variée extrême logistique : $V(x, y) = (x^{-1/\Phi} + y^{-1/\Phi})^{\Phi}$, $x > 0, y > 0, 0 < \Phi \leq 1$. Ce modèle est asymptotiquement indépendant si $\eta = (c_1 + c_2)^{-1} < 1$ et $a + 2 - 2^{\Phi} > 0$ car alors $P(X > x, Y > x) = (a + 2 - 2^{\Phi})x^{-1/\eta} \to 0$.
2. Coles et Pauli (2002) et Sisson et Coles (2003) proposent un autre modèle :

$$P(X > x, Y > y) = [S_1(x)^{1-\xi} + S_2(y)^{1-\xi} - 1]^{\frac{q}{1-\xi}}[S_1(x)S_2(y)]^{1-q} \tag{1.61}$$

   où S_i sont des fonctions de survies marginales, $0 \leq q \leq 1, \xi \geq 1$. Ce modèle est asymptotiquement indépendant si $q \neq 1$. En effet $0 \leq S(x, x) \leq [2\max(S_1(x)^{1-\xi}, S_2(x)^{1-\xi}) - 1]^{\frac{q}{1-\xi}}[S_1(x)S_2(x)]^{1-q}$ et le terme de droite tend vers 0 lorsque $x \to \infty$.
3. La structure de dépendance de Morgenstern (Morgenstern, 1956) est définie par

$$F(x, y) = F(x)F(y)(1 - \alpha\bar{F}(x)\bar{F}(y)), \tag{1.62}$$

   où F et $\bar{F}$ désignent les fonctions de répartition et survie d'une loi de Fréchet standard, et $\alpha \in [-1, 1]$. On peut montrer facilement que $P(X > x, Y > x) \to 0$ lorsque $x \to \infty$: la structure de dépendance de Morgenstern est donc asymptotiquement indépendante.

$\alpha = 0$ correspond à l'indépendance. Ledford et Tawn (2003) montrent qu'un processus markovien dont la structure de dépendance est donnée par Morgenstern vérifie $IE = 1$. De plus $\theta(u_n, r_n)$ défini par l'équation 1.54, vérifie 1-$\theta(u_n, r_n) = O(u_n^{-1})$ lorsque $u_n \to \infty$. Dans le cas de la structure de dépendance de Morgenstern, $IE = 1$ et la convergence de $\theta(u_n, r_n)$ vers $IE = 1$ est rapide.

Estimation des paramètres de la chaîne de Markov

Bortot et Tawn (1998) et Sisson et Coles (2003) estiment les paramètres de la chaîne de Markov par une méthode de maximum de vraisemblance censurée. La vraisemblance de la série temporelle $X_i, i = 1, \ldots, n$ est donnée par l'équation 1.57, dans laquelle les valeurs des densités $f(x)$ et $f(x, y)$ sont censurées pour x et/ou y inférieurs à u.

Par exemple, si la queue de distribution de la marginale est modélisée par : $F(x) = 1 - \lambda\{1 - k(x-u)/\alpha\}^{1/k}$, avec $\lambda = P(X_1 > u)$ pour $x > u$, alors

si $x \geq u$	$f(x)$ est donné par $F'(x)$,
si $x < u$	$f(x)$ est remplacé par 1-λ

De même, si la queue de distribution de la loi jointe est modélisée par l'équation 1.27, soit :

$$F(x, y) = F(x) + F(y) - 1 + \mathcal{L}\{-1/\log F(x), -1/\log F(y)\}\{-\log F(x)\}^{c_1}\{-\log F(y)\}^{c_2}$$

pour $x, y > u$, alors

si $x < u, y < u$	$f(x, y)$ est remplacé par $F(u, u)$
si $x < u, y > u$	$f(x, y)$ est remplacé par $\frac{\partial F}{\partial y}(u, y)$
si $x > u, y < u$	$f(x, y)$ est remplacé par $\frac{\partial F}{\partial x}(x, u)$
si $x > u, y > u$	$f(x, y)$ est remplacé par $\frac{\partial F^2}{\partial x \partial y}(x, y)$

Les paramètres peuvent être également estimés avec une procédure bayésienne, en considérant la vraisemblance présentée ci-dessus.

Une autre méthode d'estimation est proposée par Ledford et Tawn (2003), avec une pseudo-vraisemblance censurée en utilisant la variable $T_\tau = \min\{X_1, X_{1+\tau}\}$. Ils proposent également une méthode fondée sur du bootstrap, pour calculer les intervalles de confiance des paramètres.

Simulation des clusters

Les méthodes proposées par Smith *et al.* (1997) et Bortot et Tawn (1998) supposent qu'à l'intérieur des clusters d'une série stationnaire, le processus est markovien de premier ordre. Ils ajustent alors un processus markovien à la série, par une méthode de vraisemblance censurée par exemple, puis simulent des clusters de taille r donnée, via le processus ajusté, avec r tel que deux dépassements de seuil u, séparées d'un pas de temps r soient quasiment indépendants. Ledford et Tawn (2003) proposent une méthode pour déterminer

un tel r. La simulation des extrêmes de la chaîne de Markov peut permettre ensuite d'estimer $\theta(u_n, r_n), \pi_n, \chi(u_n), \bar{\chi}(u_n)$ ou d'autres statistiques associées aux clusters.

La méthode générale de Smith *et al.* (1997), Bortot et Tawn (1998) et Sisson et Coles (2003) consiste à simuler les maxima des clusters, selon une loi GPD. Le choix de la loi GPD est justifié par un théorème de Leadbetter (1991) sur les processus ponctuels stationnaires, fondement mathématique de la méthode POT (Peak over Threshold), et cité par Beirlant *et al.* (2004). Le choix de la loi GPD est également justifié heuristiquement puisque les maxima des clusters sont en fait des dépassements de seuil.

Ensuite, pour chaque maximum de cluster X_0, on simule le contenu du cluster

$$X_1, \ldots, X_{[r/2]}, X_{-1}, \ldots, X_{-[r/2]}$$

($[x]$ est la partie entière de x) à partir de la distribution de la chaîne de Markov, en contraignant les valeurs simulées à être inférieures au maximum X_0. Le cluster correspond à toutes les valeurs X_i simulées à l'intérieur de la fenêtre du cluster simulé (de largeur $r+1$), et qui dépassent le seuil u. La taille du cluster correspond au nombre de dépassements à l'intérieur de la fenêtre du cluster simulé.

1.4.3 Comportement à l'intérieur des clusters

Le comportement à l'intérieur des clusters est décrit par :
- une approximation de *IE* à un seuil u_n donné : $\theta(u_n, r_n)$,
- la distribution de la taille des clusters $\pi_n(j; u_n, r_n)$ définie par l'équation 1.52,
- l'agrégation des dépassements de seuil u_n : $A(u_n, r_n) = \sum_{i=1}^{r_n}(X_i - u_n)_+$. Dans le cas des pluies, A correspond au cumul des dépassements d'un cluster.

Nous avons cité des théorèmes limites à la section 1.4.1. Dans le cas de chaînes de Markov d'ordre 1, $\theta(u_n, r_n), \pi_n, A(u_n, r_n)$ sont déterminés par la distribution jointe de deux termes consécutifs de la chaîne. Les analyses suivantes ont été réalisées dans le cas où les dépassements de seuil u_n sont modélisés par une chaîne de Markov au premier ordre.

Lorsque X_t, X_{t+1} sont asymptotiquement dépendants

$$\begin{aligned} \theta(u_n, r_n) &\to IE < 1 \\ \pi_n(1; u_n, r_n) &\to \pi(1) < 1 \end{aligned} \tag{1.63}$$

lorsque $n \to \infty$, d'après Smith (1992), Perfekt (1994), Smith *et al.* (1997) et Bortot et Tawn (1998).

Lorsque X_t, X_{t+1} sont asymptotiquement indépendants,

$$\begin{aligned} \theta(u_n, r_n) &\to 1 \\ \pi_n(j; u_n, r_n) &\to 1 \text{ si } j = 1 \text{ ou } 0 \text{ si } j > 1, \end{aligned} \tag{1.64}$$

lorsque $n \to \infty$. Asymptotiquement, les clusters sont donc constitués d'un seul dépassement, et sous certaines conditions, la distribution des agrégations des dépassements converge vers celle des dépassements :

$$P(A(u_n, r_n) \leq w | M_{r_n} - u_n > 0) \to \lim_{n\to\infty} P(M_{r_n} - u_n \leq w | M_{r_n} - u_n > 0), \tag{1.65}$$

où $M_r = \max\{X_1, \ldots, X_r\}$ (Bortot et Tawn, 1998).

Ces résultats limites ne donnent pas d'information sur le comportement des fonctions $\theta(u_n, r_n), \pi_n, A(u_n, r_n)$ pour des seuils u_n finis, ni la vitesse avec laquelle la taille des clusters s'affaiblit lorsque le seuil u_n augmente.

Cependant, en pratique, on s'intéresse au comportement des clusters au delà de seuils u_n élevés, mais finis. Par exemple, deux variables consécutives peuvent être asymptotiquement indépendantes, mais former des clusters au dessus de seuils finis. Beaucoup de recherches ont été menées sur *IE* et ses estimateurs $\widehat{\theta}(u_n, r_n)$ dépendants du seuil, sur $\pi_n(j; u_n, r_n)$, $A(u_n, r_n)$ notamment dans le cas des séries stationnaires modélisées par des chaînes de Markov (Smith *et al.*, 1997), (Bortot et Tawn, 1998), (Ancona-Navarrete et Tawn, 2000), (Ledford et Tawn, 2003).

Smith *et al.* (1997), Bortot et Tawn (1998) et Ledford et Tawn (2003) se sont intéressés à ce problème, et ont analysé le comportement des variables $\theta(u_n, r_n), \pi_n, A(u_n, r_n)$ en fonction de u_n et r_n. Smith *et al.* (1997) ont travaillé avec des processus asymptotiquement dépendants, en définissant la fonction de transition entre deux valeurs supérieures à u_n par une loi bivariée extrême. Ensuite, Bortot et Tawn (1998) et Ledford et Tawn (2003) se sont intéressés aux processus asymptotiquement indépendants, en définissant la fonction de transition entre deux valeurs consécutives supérieures à u_n par une loi bi-variée définie avec une fonction à variation lente d'équation 1.27 pour $x > u_n, y > u_n$. Bortot et Tawn (1998) ont proposé un équivalent de $\theta(u_n, r_n)$ lorsque u_n est grand, et des vitesses de convergence de π_n et $A(u_n, r_n)$. Les résultats de Bortot et Tawn (1998) permettent :

- d'identifier l'effet du paramètre η de la loi bi-variée, sur le comportement des clusters, en fonction du seuil u_n où on les extrait,
- d'estimer facilement $\theta(u_n, r_n)$.

Les résultats de Bortot et Tawn (1998) sont donnés sous la condition $\Delta^{(2)}(u_n, r_n)$ suivante :

Condition $\Delta^{(2)}(u_n, r_n)$:

$$\sum_{j=3}^{r_n} \frac{P(X_1 > u_n, \max\{X_2, \ldots, X_{j-1}\} \leq u_n, X_j > u_n)}{P(X_1 > u_n, X_2 > u_n)} \to 0 \tag{1.66}$$

lorsque $n \to \infty$.

Cette condition contrôle la forme des clusters. Bortot et Tawn (1998) font la conjecture que la condition $\Delta^{(2)}(u_n, r_n)$ est vérifiée pour les chaînes de Markov asymptotiquement indépendantes et positivement associées (c'est-à-dire avec $\frac{1}{2} \leq \eta < 1$).

Pour des chaînes de Markov asymptotiquement indépendantes vérifiant la condition $\Delta^{(2)}(u_n, r_n)$, on a, pour tout $r \geq 2$ fixé et $u_n \to \infty$,

$$1 - \theta(u_n, r_n) \sim 1 - \theta(u_n, r). \tag{1.67}$$

Donc pour toute suite r_n satisfaisant $\Delta^{(2)}(u_n, r_n)$, $\theta(u_n, r_n)$ peut être approché par $\theta(u_n) = \theta(u_n, r)$ pour n'importe quel entier r fixé.

Bortot et Tawn (1998) montrent ensuite que

$$1 - \theta(u_n) \sim \mathcal{L}(u_n, u_n) u_n^{1-1/\eta}. \tag{1.68}$$

L'estimation de $\theta(u_n, r_n), \pi_n, A(u_n, r_n)$ reste cependant délicate, car elle s'appuie sur des données extrêmes, supérieures à des seuils élevés u_n. Les valeurs extrêmes d'un échantillon

sont en effet peu nombreuses, relativement à la taille de l'échantillon. On voit ici l'intérêt de modéliser le processus stationnaire à l'intérieur des clusters par un processus markovien au premier ordre : Smith *et al.* (1997), Bortot et Tawn (1998) et Sisson et Coles (2003) ajustent un modèle markovien sur les données, et simulent des clusters avec le processus markovien ajusté. Ces simulations permettent d'obtenir un grand nombre de valeurs et ainsi mieux estimer $\theta(u_n, r_n), \pi_n, A(u_n, r_n)$. Remarquons que pour simuler un cluster, on doit auparavant fixer le seuil u_n et le nombre r_n de valeurs à l'intérieur du cluster. Comme Bortot et Tawn (1998) l'ont montré pour $\eta < 1$, les résultats obtenus sont robustes par rapport au choix de r_n. Pour de grandes valeurs de u_n, r_n doit néanmoins être suffisamment grand pour ne pas tronquer les clusters. D'autre part, si r_n est trop grand, le nombre de clusters sera petit et l'estimation incertaine.

À propos du choix de r_n, Ledford et Tawn (2003) ont donné une méthode pour estimer la valeur de r_n au delà de laquelle on peut accepter l'hypothèse de non-dépendance entre deux variables du processus séparées de r_n pas de temps. La méthode de Ledford et Tawn (2003) est fondée sur une hypothèse markovienne : la dépendance dans les clusters du processus est markovienne au premier ordre, et la fonction de transition entre deux variables consécutives supérieures à un seuil élevé est modélisée par une loi bi-variée exprimée à l'aide de fonction à variation lente. Une autre méthode pour selectionner r_n a été proposée par Ferro et Segers (2003).

Enfin, dans le cas de séries stationnaires, où la dépendance entre deux variables consécutives est modélisée via une loi bi-variée des valeurs extrêmes, la propriété de max-stabilité de la famille des distributions des valeurs extrêmes implique que la dépendance extrême, mesurée par $\theta(u_n, r_n)$ est constante à tous les niveaux u_n élevés.

1.5 Remarques et conclusions

Nous avons présenté différents modèles de valeurs extrêmes, utilisables en hydrologie. Les applications sont en effet nombreuses, pour modéliser par exemple des maxima annuels de pluie, des dépassements de seuils élevés par deux familles paramétriques de lois de valeurs extrêmes (GEV et GPD). En hydrologie, la loi des valeurs extrêmes la plus couramment employée en modélisation uni-variée est la loi Gumbel pour les maxima annuels et la loi exponentielle pour les dépassements de seuils. Cette utilisation est justifiée par la simplicité de la loi Gumbel et le nombre parfois réduit de données disponibles. Les problèmes hydrologiques plus complexes nécessitent une modélisation multivariée des valeurs extrêmes, pour modéliser plusieurs durées de pluie ou plusieurs variables comme la pluie, la température, les débits, etc. La difficulté de la modélisation multi-variée réside dans l'infinité des modèles multi-variés et dans le choix à réaliser entre modèles asymptotiquement dépendants et asymptotiquement indépendants. Certaines fonctions de dépendance des extrêmes, comme χ et $\bar{\chi}$ peuvent aider à différencier les deux cas. Des modélisations plus précises des processus pluvieux cherchent parfois à représenter non seulement les valeurs maximales annuelles ou les dépassements mais toute la chronique de pluie. Dans ce cas, on peut s'intéresser à la modélisation des clusters et en particulier à la dépendance à l'intérieur des clusters. La dépendance temporelle à l'intérieur des clusters peut être modélisée et simulée par un processus markovien, et étudiée par certains paramètres de dépendance de queue comme $\chi, \bar{\chi}, \eta$. Les estimations des paramètres des modèles peuvent être paramétriques ou non paramétriques, et permettent de déduire les quantiles de pluie pour des fréquences données.

Ces modèles permettent d'estimer des quantiles de manière ponctuelle. Le chapitre suivant s'intéresse à l'incertitude d'estimation des valeurs extrêmes.

Chapitre 2

Incertitudes sur les valeurs extrêmes

Les modèles probabilistes ou stochastiques utilisés pour représenter des phénomènes réels sont entachés d'erreur, qu'il est nécessaire de préciser. L'estimation de quantiles pour le dimensionnement d'ouvrages hydrauliques, ou pour les Plans de Prévention des Risques liés aux Inondations, requiert un certain niveau de confiance, que l'on doit prendre en considération. La connaissance de ces incertitudes est d'autant plus importante si l'on est dans le domaine d'extrapolation des données observées. L'objectif de cette section est de présenter différentes méthodes pour estimer les incertitudes des modèles et des quantiles estimés par les modèles.

En modélisation hydrologique, les incertitudes peuvent être dues aux données ou au modèle :

- incertitudes sur les données d'entrée du modèle (échantillonnage, erreurs de mesure, faible nombre de données en particuliers dans les extrêmes),
- incertitudes statistiques sur les paramètres du modèle lors de son calage,
- incertitudes de modélisation : la structure du modèle statistique ou stochastique choisi peut être représentative ou non du phénomène modélisé.

Dans l'état actuel de nos connaissances, aucune recherche ne traite explicitement l'analyse des incertitudes en considérant simultanément les différentes sources d'erreurs. Néanmoins, Beven et Binley (1992) ont proposé une méthode (GLUE, présentée dans la section 2.1.3) prenant en compte implicitement toutes les sources d'incertitude. D'après une analyse bibliographique de Engeland *et al.* (2005), plusieurs études montrent que les mesures de pluies causent de plus grandes incertitudes que les paramètres (Storm *et al.*, 1988), (Thorsen *et al.*, 2001) ou la structure du modèle (Krzysztofowicz, 1999).

En pratique, on présente les incertitudes par des intervalles permettant de quantifier la dispersion d'une variable ou d'un paramètre autour de la valeur attendue. Pour γ un paramètre, un quantile, ou toute fonction d'un paramètre, on présente deux types d'intervalles.

- En analyse fréquentielle, γ est supposé fixé, et estimé par un estimateur $\widehat{\gamma}$, aléatoire. On peut alors définir un intervalle de confiance de niveau de confiance $1-\alpha$ de γ par deux limites aléatoires G_1, G_2, telles que $[G_1, G_2]$ contient la vraie valeur γ avec une probabilité $1-\alpha$.

– Si X est une variable aléatoire, et si $[x_{\alpha/2}, x_{1-\alpha/2}]$ est un intervalle tel que $P(x_{\alpha/2} \leq X \leq x_{1-\alpha/2}) = 1-\alpha$, alors l'intervalle $[x_{\alpha/2}, x_{1-\alpha/2}]$ est également appelé intervalle de confiance de niveau $1-\alpha$ de la variable aléatoire X.
– Dans le cadre bayésien, les paramètres sont supposés aléatoires. On peut donc définir un intervalle de confiance des paramètres, à un niveau $1-\alpha$ par deux réels G_1, G_2 tels que γ appartient à $[G_1, G_2]$ avec probabilité $1-\alpha$. Pour différencier le cas bayésien du cas fréquentiel, on appelle intervalle de crédibilité de γ, l'intervalle de confiance ainsi défini.

Nous avons vu dans le chapitre précédent différentes méthodes pour estimer des quantiles de valeurs extrêmes. Or ces estimations sont soumises aux différentes incertitudes citées plus haut. Le présent chapitre donne donc, dans une première partie, des méthodes pour calculer les intervalles de confiance des quantiles de valeurs extrêmes.

Dans une deuxième partie, nous analysons l'incertitude des fréquences empiriques et des statistiques d'ordre, dont les plus fortes ont un rôle important dans la validation graphique des modèles. Cette analyse permet de relativiser les comparaisons graphiques des quantiles observés (dont les fréquences sont estimés par les fréquences empiriques) avec les quantiles estimés issus des modèles.

Enfin, puisque la loi GEV sera largement utilisée dans cette thèse, nous présentons une courte analyse des incertitudes relatives à l'estimation de ses paramètres.

2.1 Intervalles de confiance de quantiles de valeurs extrêmes

Suivant le modèle hydrologique utilisé (loi de probabilité ou modèle stochastique complexe), l'analyse des incertitudes des quantiles est menée avec différentes méthodes. On présente d'abord les méthodes adaptées pour des lois de probabilité, puis des méthodes spécifiquement proposées en hydrologie, et enfin des méthodes bayésiennes, utilisables pour de nombreux modèles probabilistes simples ou complexes.

2.1.1 Calculs d'intervalles de confiance dans le cas de modèles décrits par une loi de probabilité

Soit une variable modélisée par une loi de probabilité. Dans le cas paramétrique, l'incertitude est représentée par la variabilité aléatoire des estimateurs des paramètres, et se propage dans la variabilité aléatoire des estimateurs des quantiles. On présente ici, de manière non exhaustive, les méthodes les plus répandues en statistiques : une méthode non paramétrique (le bootstrap), puis des méthodes paramétriques fondées sur des simulations, ou sur des estimations par une méthode de maximum de vraisemblance, des moments, ou des moments pondérés, et enfin quelques méthodes développées pour les distributions des valeurs extrêmes.

1. Le bootstrap (Efron, 1979) est une méthode non paramétrique, elle consiste à tirer avec remise des valeurs de l'échantillon de départ, pour produire de nouveaux échantillons (appelés échantillons bootstrap) de même taille que l'échantillon de départ. Pour chacun des B échantillons Bootstrap ainsi créés, on estime le quantile de fréquence donnée, via une méthode paramétrique ou non paramétrique. On obtient par conséquent une suite de quantiles. Sous certaines conditions de régularité, la théorie montre que

la distribution de la suite de quantiles obtenus converge vers la réelle distribution d'échantillonnage du quantile. Si $q_{B,\alpha/2}, q_{B,1-\alpha/2}$ désignent les quantiles empiriques de fréquence $\alpha/2, 1-\alpha/2$ de la distribution empirique de q, simulée par bootstrap, $[q_{B,\alpha/2}, q_{B,1-\alpha/2}]$ est un intervalle de confiance bootstrap de niveau $1-\alpha$ autour du quantile q.

Il existe d'autres méthodes de bootstrap, en particulier la méthode t-bootstrap (Efron, 1981). La méthode échantillonne B échantillons bootstrap. Sur chaque échantillon bootstrap, on estime le quantile $\widehat{q}^b$ et l'écart-type de l'estimateur : $\widehat{\sigma}(q^b)$, pour $b = 1\ldots, B$. On note $T^b = \frac{\widehat{q}^b - \widehat{q}}{\widehat{\sigma}(q^b)}$, où $\widehat{q}$ est l'estimation du quantile q sur l'échantillon de départ. On obtient ainsi une suite de T^b, et si $T_{B,\alpha/2}, T_{B,1-\alpha/2}$ désignent les quantiles empiriques de fréquence $\alpha/2$ et $1-\alpha/2$ de la distribution de T, alors l'intervalle de confiance t-bootstrap de q est donné par

$$[\widehat{q} - T_{B,\alpha/2}, \widehat{q} - T_{B,1-\alpha/2}] \tag{2.1}$$

Si l'on s'intéresse aux valeurs extrêmes, la méthode du bootstrap a l'inconvénient de ne pas créer des échantillons avec des valeurs supérieures à la valeur maximale observée dans l'échantillon de départ. De plus, il est possible de tirer des échantillons avec un grand nombre de valeurs extrêmes, du fait du tirage avec remise dans l'échantillon de départ. Pour éviter ce dernier problème, Davison *et al.* (1986) proposent une méthode, appelée 'balanced resampling', dans laquelle chaque observation de l'échantillon de départ apparaît le même nombre de fois dans la réunion de tous les échantillons bootstrap créés. Le 'balanced resampling' est équivalent à concaténer B copies identiques à l'échantillon de départ (de taille n), permuter les nB valeurs de l'échantillon réuni, et ensuite diviser l'échantillon réuni en B échantillons de taille n. Cette méthode a été appliquée à l'étude des quantiles de crues extrêmes (Burn, 2003).

De manière plus théorique, différents auteurs ont proposé et vérifié des méthodes modifiées du bootstrap pour l'étude des quantiles extrêmes (Swanepoel, 1986), (Deheuvels *et al.*, 1993), (Angus, 1993), (Bacro et Brito, 1998), (Athreya et Fukuchi, 1997). La modification la plus connue consiste à prendre une taille m d'échantillons bootstrap, avec $m \to \infty$ et $m/n \to 0$ si $n \to \infty$, où n est la taille de l'échantillon de départ. Zelterman (1993) propose une méthode de bootstrap semi-paramétrique dans laquelle la loi des variables de l'échantillon appartient au domaine d'attraction de la loi Gumbel. La méthode semi-paramétrique proposée simule la loi jointe des j plus grandes valeurs d'un échantillon de grande taille.

Enfin, le nombre B d'échantillons bootstrap est souvent choisi (de manière conventionnelle) de l'ordre de 1000. Lee (2000) montre que ce nombre n'est pas toujours adéquat : la probabilité de recouvrement souhaitée peut être fausse. Il propose une méthode pour choisir B, fondée sur un choix adaptatif de B.

2. Dans un cadre théorique où l'on connaît la loi paramétrique F_θ dont est issu l'échantillon, il est possible de simuler la distribution d'échantillonnage des quantiles, et d'en déduire les intervalles de confiance des quantiles. Pour cela, on simule un grand nombre d'échantillons (de taille fixée au préalable) avec la distribution F_θ, on estime $F_{\widehat{\theta}}$, et on en déduit les quantiles correspondant, par inversion de $F_{\widehat{\theta}}$. Dans le cas d'un modèle non-paramétrique, un intervalle de confiance des quantiles peut également être estimé par un grand nombre de simulations des quantiles via le modèle non-paramétrique. Dans les cas paramétrique ou non-paramétrique, l'intervalle de confiance est jugé correct dès qu'il est stable par rapport au nombre de simulations.

3. Si les paramètres de la distribution ajustée aux observations sont estimés par maximum de vraisemblance, le résultat connu de normalité asymptotique des estimateurs des paramètres permet le calcul d'intervalles de confiance (Kendall et Stuart, 1987).
   On rappelle le théorème (voir (Coles, 2001)) :
   Théorème : Soient $x_1, \ldots, x_n$ des réalisations indépendantes d'une distribution appartenant à une famille paramétrique. Soient $l(.)$ et $\widehat{\theta}$ la fonction de log-vraisemblance et l'estimateur du maximum de vraisemblance du paramètre d-dimensionnel θ. Alors, sous certaines conditions de régularité, pour n grand,
   $$\widehat{\theta} \sim \mathcal{N}_d(\theta, I_E(\theta)^{-1}), \tag{2.2}$$
   où $I_E(\theta)$ est la matrice d'information de Fisher.
   En pratique, on peut donc approcher la loi de l'estimateur du paramètre par une loi gaussienne, centrée sur $\widehat{\theta}$ et de matrice de variance-covariance donnée par l'inverse de la matrice d'information de Fisher estimée en $\widehat{\theta}$.
   Par la suite, le calcul des intervalles de confiance des quantiles peut être réalisé d'au moins deux manières. Une première méthode est de simuler les paramètres, selon leur distribution normale approchée, puis d'en déduire les quantiles par inversion de la fonction de répartition. Un tri des quantiles simulés permet ensuite d'en déduire des intervalles de confiance pour les quantiles. Cette méthode a été appliquée par Seong et Lee (2003). Une seconde méthode est analytique, et généralement appelée delta méthode. Elle repose sur le théorème suivant :
   Théorème : On reprend les notations et hypothèses du théorème précédent. Soit $\Phi = g(\theta)$ une fonction réelle du paramètre θ, continûment dérivable. Soit $\nabla\Phi = [\frac{\partial\Phi}{\partial\theta_1}, \ldots, \frac{\partial\Phi}{\partial\theta_d}]^T$ évalué en $\widehat{\theta}$, on suppose $\nabla\Phi \neq 0$. Alors l'estimateur du maximum de vraisemblance de Φ satisfait $\widehat{\Phi} = g(\widehat{\theta})$ et $\widehat{\Phi} \sim \mathcal{N}(\Phi, V)$ où $V = \nabla\Phi^T I_E(\theta)^{-1} \nabla\Phi$.
   D'après ce théorème, et puisque les quantiles sont fonctions des paramètres, les intervalles de confiance des quantiles sont symétriques pour un échantillon de taille importante.
   Dans la première méthode, les intervalles de confiance des quantiles ne sont pas en général symétriques. Par exemple, dans le cas de la loi GEV, si les paramètres sont simulés dans leur distribution asymptotique gaussienne, les quantiles ont une dissymétrie positive, particulièrement forte pour les quantiles de fréquence élevée. Cette dissymétrie est plus réaliste dans le cas des quantiles extrêmes de pluie : on peut en effet penser, intuitivement, que l'incertitude est plus forte sur la borne supérieure des quantiles que sur la borne inférieure.
4. Si le paramètre $\theta = (\theta_1, \ldots, \theta_d)$ de la distribution étudiée est de dimension $d \geq 2$, il est possible de calculer des intervalles de confiance des paramètres marginaux θ_i avec la méthode précédente, en considérant les lois marginales de la loi normale asymptotique de l'estimateur de θ. Une autre méthode, habituellement plus précise, est fondée sur la vraisemblance profilée (Coles, 2001). On note la log-vraisemblance $l(\theta) = l(\theta_i, \theta_{-i})$, où θ_{-i} représente toutes les composantes de θ, excepté θ_i. La log-vraisemblance profilée du paramètre marginal θ_i est définie par
   $$l_p(\theta_i) = \max_{\theta_{-i}} l(\theta_i, \theta_{-i}). \tag{2.3}$$
   Pour toute valeur de θ_i, la log-vraisemblance profilée de θ_i est la log-vraisemblance maximisée sur toutes les autres composantes du paramètre θ. En d'autres termes, $l_p(\theta_i)$ est la projection de la surface de log-vraisemblance 'vue' de l'axe θ_i.

Cette définition se généralise : soit $\theta = (\theta^{(1)}, \theta^{(2)})$, avec $\theta^{(1)}$ de dimension k, et $\theta^{(2)}$ de dimension $d - k$. La log-vraisemblance profilée de $\theta^{(1)}$ est définie par

$$l_p(\theta^{(1)}) = \max_{\theta^{(2)}} l(\theta^{(1)}, \theta^{(2)}). \tag{2.4}$$

Si $k = 1$, cette définition se réduit à la définition précédente. Le théorème suivant (voir (Coles, 2001)) permet de définir une procédure pour définir un intervalle de confiance sur $\theta^{(1)}$.

Théorème de déviance : Soient $(x_1, \ldots, x_n)$ des réalisations indépendantes d'une distribution appartenant à une famille paramétrique. Soit $\widehat{\theta}$ l'estimateur du maximum de vraisemblance du paramètre à d dimensions $\theta = (\theta^{(1)}, \theta^{(2)})$ de cette distribution, avec $\theta^{(1)}$ de dimension k. Alors sous certaines conditions de régularité, et pour n assez grand,

$$D_p(\theta^{(1)}) = 2\{l(\widehat{\theta}) - l_p(\widehat{\theta^{(1)}})\} \sim \chi^2_k. \tag{2.5}$$

Ce théorème permet d'en déduire une région de confiance de niveau $1 - \alpha$ sur $\theta^{(1)}$: $\{\theta^{(1)} : D_p(\theta^{(1)}) \leq \chi^2_{k,\alpha}\}$, où $\chi^2_{k,\alpha}$ est le quantile de fréquence $1 - \alpha$ de la distribution χ^2 à k degrés de liberté.

Si un changement de variable permet de remplacer un paramètre par un quantile, on peut calculer des intervalles de confiance sur le quantile. Cette méthode est proposée par Coles et Tawn (1996), ainsi que Seong et Lee (2003) pour estimer un intervalle de confiance des quantiles d'une loi GEV. Les paramètres de position, échelle et forme de la loi GEV $\beta_{GEV}, \alpha_{GEV}$ et k_{GEV} sont remplacés par $\beta_{GEV}, k_{GEV}, q_T$, avec T une période de retour fixée. La région de confiance à deux dimensions en (k_{GEV},q_T) est déterminée par les valeurs k_{GEV} et q_T qui vérifient

$$2[l(\widehat{\beta_{GEV}}, \widehat{k_{GEV}}, \widehat{q_T}) - l_p(k_{GEV}, q_T)] \leq \chi^2_{2,\alpha}$$

où $l_p = \max_{\beta_{GEV}} l(\beta_{GEV}, k_{GEV}, q_T)$, et $\chi^2_{2,\alpha}$ représente le quantile d'ordre $1 - \alpha$ de la distribution du χ^2 à deux degrés de liberté.

Seong et Lee (2003) comparent la delta méthode et la méthode de la vraisemblance profilée, sur plusieurs séries de Corée du Sud. La différence relative entre les bornes supérieures des deux types d'intervalles de confiance des quantiles de pluie maximale annuelle en 30, 60 ou 120 minutes (modélisée par une loi GEV) est de l'ordre de 10%.

5. Lorsque la distribution de probabilité possède deux paramètres et qu'ils sont estimés par une méthode des moments, Colin (1977) a proposé une méthode pour le calcul des intervalles de confiance de quantiles. Cette méthode repose sur une hypothèse forte de normalité. Cependant, Lang (1995) a montré par simulations que ces intervalles de confiance ne sont pas recouverts avec la probabilité de recouvrement attendue pour les lois normale, Gumbel, Pearson III.

6. Des études ont été proposées pour des lois particulières souvent utilisées en hydrologie. Nous en présentons ici quelques unes, de manière non exhaustive. Moran (1957) a donné une approximation asymptotique de la variance des quantiles estimés par maximum de vraisemblance, dans le cas des lois normale, lognormale, gamma ; de même, Kite (1975) pour les lois à trois paramètres Pearson III et log-Pearson III. Stedinger (1983) montre que la méthode de Kite (1975) ne donne pas de bons résultats, en terme de taux de recouvrement du vrai quantile par l'intervalle de confiance. Il propose une

autre méthode pour calculer les intervalles de confiance des quantiles d'une loi Pearson III ou log-Person III, lorsque le coefficient d'asymétrie γ est connu, en utilisant les intervalles de confiance exacts des quantiles d'une loi normale. Si les quantiles x_p et y_p de fréquence p, des lois respectivement normales et Pearson III sont obtenus par la méthode des moments, alors un développement au premier ordre des variances des estimateurs des quantiles donne :

$$\lambda^2 = \frac{\mathrm{var}(\widehat{y}_p)}{\mathrm{var}(\widehat{x}_p)} \approx \frac{1 + \gamma K_p + \frac{1}{2}(1 + \frac{3}{4}\gamma^2)K_p^2}{1 + \frac{1}{2}z_p^2} \tag{2.6}$$

avec K_p, z_p les quantiles de fréquence p d'une loi de Pearson III centrée réduite, et d'une loi normale centrée réduite. Alors un intervalle de confiance des quantiles y_p, pour un échantillon $y_1, \ldots, y_n$ de réalisations d'une loi de Pearson III, est donné par

$$[\bar{y} + s_y\{K_p + \lambda(\zeta_\alpha(p) - z_p)\}, \bar{y} + s_y\{K_p + \lambda(\zeta_{1-\alpha}(p) - z_p)\}], \tag{2.7}$$

où $\zeta_\alpha(p), \zeta_{1-\alpha}(p)$ sont des quantiles d'une distribution de Student non centrée, dont les tables sont données dans (Stedinger, 1983).

Une autre méthode est proposée par Ashkar et Bobee (1988), et généralisée au cas où le coefficient d'asymétrie est inconnu. L'idée de Ashkar et Bobee (1988) reprend celle de Stedinger (1983), mais avec une méthode d'anamorphose vers une loi quelconque. Soit Y une variable aléatoire dont on ne sait pas construire d'intervalle de confiance pour les quantiles Y_p, et X une variable aléatoire dont on peut calculer les intervalles de confiance $[L_\alpha(p), U_\alpha(p)]$ des quantiles X_p à un niveau α. Alors si
- $p_1 = F_X(L_\alpha(p))$,
- $p_2 = F_X(U_\alpha(p))$,
- $Y_{p_i} = F_Y^{-1} o F_X(X_{p_i}), i = 1, 2$

$[Y_{p_1}, Y_{p_2}]$ est un intervalle de confiance de niveau α de Y_p. F_Y étant inconnu, on aura un intervalle de confiance approché pour Y_p en remplaçant F_Y par son estimateur $\hat{F}_Y$. Ashkar et Bobee (1988) appliquent cette méthode au cas où Y est de loi Pearson III ou log-Pearson III, et X de loi normale, lognormale à trois paramètres[1], exponentielle, Weibull à deux paramètres $a, b > 0$ (de densité, pour $x > 0 : f(x) = \frac{b}{a}(\frac{x}{a})^{b-1}\exp(-\frac{x}{a}^b)$), ou Gumbel. Ashkar et Bobee (1988) montrent par simulations Monte-Carlo que la distribution normale donne de meilleurs résultats dans le cas où le coefficient d'asymétrie γ de Y est connu. Dans le cas où γ est inconnu, la distribution de Weibull à deux paramètres donne de bons résultats, pour γ positif et dans un domaine de valeurs courantes en hydrologie, et pour une fréquence des quantiles étudiés supérieure à 0.9. Ces résultats sont intéressants pour l'étude des extrêmes en hydrologie, car le coefficient d'asymétrie des quantités de pluie est positif.

7. Dans le cas des lois GEV ou GPD, Hosking *et al.* (1985) et Hosking et Wallis (1987) donnent une approximation de la variance asymptotique des paramètres et des quantiles estimés par les moments pondérés, introduits par Greenwood *et al.* (1979). Ils montrent également que la distribution asymptotique des paramètres est normale. Il est donc possible d'utiliser cette propriété pour construire des intervalles de confiance des paramètres et des quantiles, par exemple par simulation Monte-Carlo des paramètres, puis calcul des quantiles correspondants. Hosking et Wallis (1987) comparent les taux

[1] X suit une loi lognormale à trois paramètres si il existe $x_0 \in \mathbb{R}$ tel que $Y = \log(X - x_0)$ suit une loi normale.

de recouvrements d'intervalles de confiance de paramètres et de quantiles de loi GPD, calculés avec des méthodes de maximum de vraisemblance, de moments ou de moments pondérés. Les conclusions varient avec les valeurs de la taille de l'échantillon, la valeur du paramètre de forme k, la fréquence du quantile estimé. Pour des quantiles de fréquence élevée, les intervalles de confiance obtenus avec la méthode des moments pondérés ont un meilleur taux de recouvrement qu'avec la méthode de maximum de vraisemblance, pour des tailles d'échantillons entre 15 et 500. Néanmoins, les résultats de Hosking et Wallis (1987) montrent que les intervalles de confiance sont souvent trop étroits (intervalles correspondant à un recouvrement de l'ordre de 80% au lieu de 90%, pour le quantile de fréquence 0.99, et une taille d'échantillon égale à 50).

Lu et Stedinger (1992) donnent d'autres estimateurs des variances des quantiles de la loi GEV de paramètre de forme connu ou inconnu, lorsque les paramètres sont estimés par une méthode des moments pondérés. Ils montrent que les formules de variance des quantiles de Hosking *et al.* (1985) peuvent être imprécises lorsque la taille d'échantillon n est inférieure à 10, ou que le quantile est de fréquence supérieure à 0.99 (période de retour 100 ans), ou que le paramètre de forme k est inférieur à -0.2. Ils montrent que la distribution des quantiles de la GEV, estimés par les moments pondérés, n'est pas normale pour une petite taille d'échantillon, mais le taux de recouvrement est cependant correct si la vraie valeur du paramètre de forme k est utilisée, et faux si le paramètre de forme est estimé.

Dans leur étude, Lu et Stedinger (1992) comparent la variabilité des estimateurs des quantiles, lorsque le paramètre de forme k de la GEV est supposé connu (par exemple, par une estimation régionale), et lorsqu'il est inconnu. Pour des échantillons de tailles 20 à 100, ils montrent que la variance d'échantillonnage des quantiles de fréquence supérieure à 0.95 (période de retour supérieure à 20 ans) est significativement réduite par l'estimation avec k supposé connu. Si on compare les erreurs quadratiques moyennes des quantiles, le biais introduit par une mauvaise représentation du paramètre de forme k est compensé par la réduction de variance d'échantillonnage de l'estimation à deux paramètres. C'est pourquoi, lorsque k est proche de 0, la loi Gumbel est préférée à la loi GEV à trois paramètres, pour des quantiles de fréquence supérieure à 1-1/n (période de retour supérieure à n), avec n la taille de l'échantillon. Ce résultat n'est pas surprenant puisque dans le cas d'une loi GEV, où le paramètre de forme est fixé, les moments d'ordre 1 et 2 suffisent à estimer les deux autres paramètres, tandis que si le paramètre de forme est inconnu, le coefficient d'asymétrie est nécessaire pour l'estimation des paramètres.

Cependant, pour des quantiles de fréquence inférieure à 0.9 (période de retour inférieure à 10 ans), et $k \leq 0$, les quantiles estimés avec une loi GEV à trois paramètres ont une variance d'échantillonnage plus faible que dans le cas d'une loi GEV avec un paramètre de forme k fixé. Selon Lu et Stedinger (1992), cela est dû au manque de flexibilité de l'estimateur à deux paramètres, par rapport aux valeurs extrêmes de l'échantillon.

2.1.2 Méthodes spécifiquement proposées en hydrologie

Classiquement, les paramètres des modèles hydrologiques sont estimés de manière à reproduire au mieux les observations. Cependant, les données utilisées pour estimer ou caler les modèles sont incertaines (du fait des erreurs de mesures, de l'échantillonnage), et les structures des modèles sont également incertaines. Une partie des hydrologues a donc rejeté la

notion de "paramètre optimal" au profit d'un ensemble de paramètres acceptables pour le modèle. Il semble que Young (1978), Spear et Hornberger (1980), Hornberger et Spear (1981) soient les premiers à avoir proposé cette idée. Plus tard, Beven et Binley (1992) introduisent la notion d'équifinalité : des paramètres différents peuvent donner des résultats similaires et corrects.

La méthode proposée par Young (1978), Spear et Hornberger (1980), Hornberger et Spear (1981) est une classification binaire des paramètres, avec la définition d'un ensemble de paramètres acceptables et d'un ensemble de paramètres non acceptables. Cette méthode est appelée RSA (Regionalized Sensitivity Analysis), et est constituée de trois étapes :

- Première étape : les observations disponibles pour l'estimation ou le calage des paramètres du modèle sont remplacées par des intervalles de valeurs, de manière à figurer (mais de façon qualitative et subjective) les incertitudes qui affectent ces données.
- Deuxième étape : plusieurs paramètres sont tirés aléatoirement dans une loi a priori des paramètres, puis insérés dans le modèle. Seuls les paramètres dont le modèle produit des sorties respectant les intervalles de valeurs observées (définies à l'étape 1) sont conservés. Après un nombre suffisant de tirages, on obtient un ensemble $\mathbf{A}$ de paramètres acceptables ou représentatifs du système réel, et son ensemble complémentaire $\overline{\mathbf{A}}$, formé des paramètres non représentatifs du système réel.
- Troisième étape : la fonction de répartition des paramètres est calculée sur chacun des ensembles $\mathbf{A}$ et $\overline{\mathbf{A}}$. Si la distribution cumulée des valeurs acceptables diffère significativement de la distribution cumulée des valeurs non acceptables, les sorties du modèles sont sensibles à la valeur du paramètre. Si les deux distributions sont similaires, les sorties du modèle sont peu sensibles aux valeurs du paramètre.

Pour des références bibliographiques d'utilisation de cette méthode, on renvoie à (Zin, 2002).

Dans la méthode RSA, aucune distinction n'est introduite entre les paramètres dits acceptables. Des modifications ont été proposées par la suite par différents auteurs, par exemple : Keesman et Van Straten (1990), Beven et Binley (1992), Klepper et Hendrix (1994), cités par Zin (2002).

Beven et Binley (1992) ont notamment proposé de pondérer les paramètres acceptables par une 'mesure de vraisemblance', relative à la capacité des paramètres à bien reproduire les observations. La méthode proposée par Beven et Binley (1992) est nommée GLUE (Generalized Likelihood Uncertainty Estimation). GLUE a été appliquée à la modélisation hydrologique par Freer *et al.* (1996), Romanowicz *et al.* (1996), Blazkova et Beven (1997), Cameron *et al.* (1999), Beven et Freer (2001), Hossain et Anagnostou (2005) et Mo *et al.* (2006). En particulier, elle a été appliquée sur des modèles de génération de pluies, pour estimer les paramètres de la loi GPD utilisée dans un modèle stochastique de pluies horaires du type Bartlett-Lewis (ce type de modèle est présenté dans le chapitre 5) par Cameron *et al.* (2000c). Elle est présentée avec plus de détails dans la section suivante, comme une généralisation d'une méthode bayésienne.

Ces deux méthodes (RSA, GLUE) nécessitent un balayage de l'espace des paramètres, par une loi a priori. Elles aboutissent à la spécification d'un ensemble de paramètres acceptables pour la modélisation, avec une approximation de leur distribution. Suivant la dispersion de la loi a priori, les temps de calcul pour ces méthodes peuvent être très longs.

2.1.3 Méthodes bayésiennes

Pour des modèles probabilistes ou stochastiques complexes, les incertitudes peuvent être analysées par la modélisation probabiliste des erreurs entre les valeurs observées et simulées (par exemple (Montanari et Brath, 2004), (Engeland *et al.*, 2005)). Les incertitudes sont alors quantifiées par une fonction objectif, la fonction de vraisemblance, et estimées par des méthodes de maximum de vraisemblance ou bayésiennes. L'approche bayésienne est un cadre intéressant pour l'étude des incertitudes puisqu'elle inclue les observations, l'information a priori sur les paramètres, et permet de comparer objectivement les sorties du modèle avec les observations (via la vraisemblance). Cette approche a été développée en hydrologie avec la notion d'équifinalité, introduite par Beven et Binley (1992) : des jeux de paramètres différents peuvent produire des résultats semblables. Les algorithmes Monte Carlo de chaînes de Markov et l'échantillonnage préférentiel (Gelman *et al.*, 1997) figurent parmi les méthodes bayésiennes les plus connues.

On considère un échantillon d'observations $\mathbf{x} = (x_1, \ldots, x_n)$, supposées être des réalisations d'une variable aléatoire dont la densité appartient à une famille paramétrique $\{f(x;\theta) : \theta \in \Theta\}$. La vraisemblance des observations est égale à $f(\mathbf{x}|\theta) = \prod_{i=1}^{n} f(x_i;\theta)$. Dans le cadre bayésien, on suppose qu'il est possible d'exprimer les connaissances sur les paramètres à travers une distribution a priori $f(\theta)$, indépendante des observations $(x_1, \ldots, x_n)$.

Alors le théorème de Bayes (1763) permet de calculer la distribution des paramètres, connaissant les observations. Pour cela, l'information a priori et les observations sont combinées de manière à définir la loi a posteriori $f(\theta|\mathbf{x})$:

$$f(\theta|\mathbf{x}) = \frac{f(\theta)f(\mathbf{x}|\theta)}{\int_\Theta f(\theta)f(\mathbf{x}|\theta)d\theta}. \tag{2.8}$$

Le cadre bayésien a ses avantages et ses inconvénients. L'intérêt de spécifier une loi a priori permet de compléter l'information détenue dans les observations, souvent très limitée, avec d'autres sources d'informations. En contrepartie, la loi a posteriori est influencée par la loi a priori. On doit donc reconnaître une part de subjectivité aux résultats a posteriori. Malgré la subjectivité du résultat a posteriori, la méthode bayésienne permet d'estimer ponctuellement et par intervalles aussi bien les paramètres que les variables prédictives, déduites des paramètres (par exemple les quantiles). La distribution d'une variable prédictive z, dépendante de θ, est donnée par : $f(z|\mathbf{x}) = \int_\Theta f(z|\theta)f(\theta|\mathbf{x})d\theta$, où $f(z|\theta)$ est la densité de z étant donné le modèle paramétré par θ. Cette densité prédictive reflète à la fois l'incertitude de la paramétrisation (par $f(\theta|\mathbf{x})$) et l'incertitude due à la variabilité de la variable z (par $f(z|\theta)$).

A propos de la loi GEV appliquée à l'étude des pluies extrêmes, Coles et Tawn (1996) montrent que les paramètres d'échelle et de forme dépendent des caractéristiques régionales du processus de pluie. La loi a priori de ces deux paramètres peut donc contenir des informations régionales. Coles et Tawn (1996) examinent l'effet d'une loi a priori non informative : pour des périodes de retour de l'ordre de la taille de l'échantillon, les estimations moyennes et les intervalles de confiance sont semblables dans le cas informatif et dans le cas non informatif. Pour des périodes de retour élevées, la moyenne des quantiles est plus faible et les intervalles de confiance sont plus larges dans le cas non informatif. Enfin, il peut être difficile de spécifier une loi a priori sur les paramètres, du fait du manque de représentativité physique des paramètres. Coles et Tawn (1996) pallient ce problème par un changement de variables, dans lequel ils remplacent les paramètres par des quantiles. En effet, la signification physique des quantiles permet aux experts en hydrologie de donner une loi a priori sur les quantiles.

En général, le dénominateur de la formule 2.8 est incalculable numériquement. Dans quelques rares cas, on peut utiliser des familles conjuguées de lois a priori et de vraisemblance, ce qui permet d'éviter le calcul de l'intégrale dans la formule 2.8. Deux techniques particulières permettent de dépasser cette difficulté : il s'agit des chaînes de Markov et des méthodes d'échantillonnade préférentiel.

Simulation Monte Carlo de chaînes de Markov

L'idée de l'algorithme de Metropolis est de simuler une loi de probabilité quelconque. Dans notre cas, nous nous sommes intéressés aux lois a posteriori des paramètres $f(\theta|\mathbf{x})$. L'algorithme simule θ grâce à une chaîne de Markov, dont la loi converge vers une loi stationnaire, en l'occurrence la loi a posteriori $f(\theta|\mathbf{x})$.

L'algorithme de Metropolis a été décrit pour la première fois par Metropolis et Ulam (1949) et Metropolis *et al.* (1953), et généralisé par Hastings (1970). Cet algorithme est fondé sur la théorie asymptotique des chaînes de Markov réversibles, indécomposables et apériodiques. C'est une méthode Monte Carlo de simulation de chaîne de Markov (MCMC), permettant de simuler la loi a posteriori des modèles, après convergence de la chaîne de Markov. Cette méthode est largement utilisée en statistique bayésienne (Chib et Greenberg, 1995), (Gelman *et al.*, 1997), (Geyer, 1992), et a été appliquée aux modèles hydrologiques (Engeland *et al.*, 2005), (Gaume *et al.*, 1998), (Kuczera et Parent, 1998), (Marshall *et al.*, 2004), (Renard *et al.*, 2006), (Thyer *et al.*, 2002), (Vrugt *et al.*, 2003).

La chaîne de Markov est simulée suivant les étapes suivantes :

1. On choisit un point de départ $\theta^{(0)} = (\theta_1^{(0)}, \ldots, \theta_d^{(0)})$, et une distribution de saut symétrique $J(\theta^{(i+1)}|\theta^{(i)}) = J(\theta^{(i)}|\theta^{(i+1)})$. Cette fonction de saut doit être facile à simuler. La loi normale centrée en $\theta^{(i)}$ est souvent utilisée.
2. A l'étape i :
   - On génère $\theta^{(i+1)}$ par la loi de saut $J(\theta^{(i+1)}|\theta^{(i)})$.
   - Le nouveau paramètre $\theta^{(i+1)}$ est conservé avec probabilité $\min\{1, f(\theta^{(i+1)}|\mathbf{x})/f(\theta^{i}|\mathbf{x})\}$. S'il est rejeté, $\theta^{(i+1)} = \theta^{(i)}$.

Remarque : il est possible de complexifier les étapes, en particulier d'utiliser une fonction de saut non symétrique, c'est le cas de l'algorithme de Metropolis-Hastings (Hastings, 1970). Il faut alors remplacer $\min\{1, f(\theta^{(i+1)}|\mathbf{x})/f(\theta^{i}|\mathbf{x})\}$ par $\min\{1, \frac{f(\theta^{(i+1)}|\mathbf{x})/J(\theta^{(i+1)}|\theta^{(i)})}{f(\theta^{(i)}|\mathbf{x})/J(\theta^{(i)}|\theta^{(i+1)})}\}$

La difficulté de cet algorithme est la vérification de sa convergence : au bout de combien d'itérations peut-on considérer que la chaîne de Markov a atteint sa loi stationnaire ? Pour répondre à cette question, Gelman *et al.* (1997) estiment une statistique R sur plusieurs chaînes de Markov simulées en parallèle avec des points de départs différents tirés dans la loi a priori des paramètres. Cette statistique mesure la variabilité des termes simulés à l'intérieur des séquences, et entre les séquences. Si $\widehat{R}$ est "proche" de 1, on peut accepter l'hypothèse de convergence de la chaîne.

La simulation d'un grand nombre de paramètres, suivant leur loi a posteriori, permet d'en déduire les distributions marginales, et les distributions de variables prédictives.

Méthode d'échantillonnage préférentiel, méthode GLUE

La méthode d'échantillonnage préférentiel consiste à explorer et évaluer la fonction objectif dans tout l'espace des modèles. Le but est de déterminer un ensemble de modèles "acceptables", au sens d'un critère sur la fonction objectif. Tanner (1992) présente la méthode d'échantillonnage préférentiel de manière exhaustive. La méthode d'échantillonnage préférentiel est simple, et permet dans certains cas d'estimer la valeur d'une intégrale : supposons que l'on cherche à calculer l'intégrale $E(h(\theta)|\mathbf{x})$, où θ est une variable aléatoire de densité $f(\theta|\mathbf{x})$. On considère le cas où l'on ne sait pas simuler θ selon sa loi.

Soit une densité g selon laquelle on sait simuler, alors on peut écrire

$$E[h(\theta|\mathbf{x})] = \int \frac{h(\theta)f(\theta|\mathbf{x})}{g(\theta)} g(\theta)d\theta \tag{2.9}$$

qui peut être estimé par

$$\frac{1}{L}\sum_{i=1}^{L} h(\theta^i)f(\theta^i|\mathbf{x})/g(\theta^i), \tag{2.10}$$

avec L tirages $\theta^1, \ldots, \theta^L$ sous la loi $g(\theta)$. Selon Gelman *et al.* (1997), il n'existe pas de méthode pour estimer la précision d'une estimation par échantillonnage préférentiel. Néanmoins, si g est tel que les ratios $h(\theta^i)f(\theta^i|\mathbf{x})/g(\theta^i)$ varient fortement, ou si les ratios sont faibles avec une forte probabilité et très forts avec une faible probabilité, alors l'estimation sera peu précise, voire fausse. Tanner (1992) et Gelman *et al.* (1997) observent que le succès de la méthode dépend fortement du choix de g. Pour une meilleure estimation, g doit être choisie proche de $f(.|\mathbf{x})$.

Si on ne connaît pas la forme approximative de la distribution $f(.|\mathbf{x})$, il est possible de manquer des θ^i dans l'équation 2.10, dont les ratios sont de très faible probabilité, mais de valeurs très élevées. Pour déceler les éventuels problèmes, on regarde souvent la distribution des ratios. En pratique, on regarde l'histogramme des logarithmes des plus grands ratios : les estimations ne seront pas correctes si les plus grands ratios sont trop grands par rapport aux autres. Le comportement des petits ratios est moins important puisque ils influencent peu le calcul de l'estimateur. Néanmoins, s'ils sont très nombreux, l'estimation peut être mauvaise.

Cette méthode peut être appliquée dans le cadre bayésien, pour le calcul des densités a posteriori marginales et pour le calcul des distributions prédictives.

Estimation des densités a posteriori marginales des paramètres Soit $\theta = (\lambda, \phi)$ une partition de l'espace des paramètres, avec la factorisation correspondante de la densité : $f(\theta|\mathbf{x}) = f(\gamma|\phi, \mathbf{x})f(\phi|\mathbf{x})$. On suppose que l'on sait approcher la densité a posteriori conditionnelle de γ par $f_{approx1}(\gamma|\phi, \mathbf{x})$. Alors la densité a posteriori marginale de ϕ est, pour toute valeur du paramètre ϕ :

$$f(\phi|\mathbf{x}) = \int f(\gamma, \phi|\mathbf{x})d\gamma \tag{2.11}$$

$$= E_{approx1}\left(\frac{f(\gamma, \phi|\mathbf{x})}{f_{approx1}(\gamma|\phi, \mathbf{x})}\right) \tag{2.12}$$

$$\approx \frac{1}{L}\sum_{i=1}^{L} \frac{f(\gamma^i, \phi|\mathbf{x})}{f_{approx1}(\gamma^i|\phi, \mathbf{x})}, \tag{2.13}$$

où $E_{approx1}$ est la moyenne sur γ sous la loi $f_{approx1}(\gamma|\phi, \mathbf{x})$, et γ^i est simulé sous la loi $f_{approx1}(\gamma|\phi, \mathbf{x})$.

Estimation des distributions a posteriori prédictives Soit z une variable prédictive (aléatoire ou non), dépendante de θ (par exemple un quantile associé à θ). On suppose que l'on sait approcher la distribution a posteriori $f(\theta|\mathbf{x})$ par $f_{approx2}(\theta)$, facile à simuler. Alors la distribution prédictive a posteriori de z est :

$$P(z \leq x|\mathbf{x}) = \int P(z \leq x|\theta) f(\theta|\mathbf{x}) d\theta \tag{2.14}$$

$$= E_{approx2}\left(\frac{P(z \leq x|\theta) f(\theta|\mathbf{x})}{f_{approx2}(\theta|\mathbf{x})}\right) \tag{2.15}$$

$$\approx \frac{1}{L} \sum_{i=1}^{L} \frac{P(z \leq x|\theta^i) f(\theta^i|\mathbf{x})}{f_{approx2}(\theta^i|\mathbf{x})}, \tag{2.16}$$

où θ^i est simulé sous la loi $f_{approx2}(\theta|\mathbf{x})$.

Si $f_{approx2}(\theta|\mathbf{x})$ est la loi a priori de θ, notée $f(\theta)$, alors l'approximation 2.16 est proportionnelle à :

$$\frac{1}{L} \sum_{i=1}^{L} P(z \leq x|\theta^i) v(\mathbf{x}|\theta^i), \tag{2.17}$$

avec $f(\theta|\mathbf{x}) \propto f(\theta) v(\mathbf{x}|\theta)$, d'après le théorème de Bayes, où $v(\mathbf{x}|\theta)$ est la vraisemblance de $\mathbf{x}$ étant donné le paramètre θ.

La méthode d'échantillonnage préférentiel a été généralisée par la méthode GLUE (Generalized Likelihood Uncertainty Estimation) introduite par Beven et Binley (1992). Elle généralise l'approximation 2.17 en remplaçant la vraisemblance par une "mesure de vraisemblance", non nécessairement issue de modèles probabilistes. Par exemple, cette mesure de vraisemblance peut être le critère de Nash et Sutcliffe, avec un facteur de forme $N > 0$: $(1 - \sigma_e^2/\sigma_0^2)^N$ où σ_e est la variance résiduelle et σ_0 la variance des observations. La mesure de vraisemblance peut également être une mesure quelconque calculant l'écart entre les observations et les résultats du modèle : par exemple $(\sigma_e^2)^{-N}$, ou $\exp(-N\sigma_e^2)$, avec un paramètre de forme $N > 0$.

La méthode GLUE est intéressante pour observer le comportement de fonctions objectifs particulières sur l'espace des paramètres ; mais les bases mathématiques de cette méthode sont encore mal connues. En particulier, les résultats sont conditionnés par le choix de la mesure de vraisemblance. Les propriétés des mesures de vraisemblances 'non classiques' sont moins connues. D'autre part, si la "mesure de vraisemblance" définie dans la méthode GLUE n'est pas associée à une distribution de probabilité, la théorie bayésienne ne permet plus d'affirmer que la distribution prédictive ainsi estimée correspond à la distribution prédictive a posteriori classique. L'usage de ces mesures nécessite donc d'abord d'en comprendre le comportement statistique. On préférera la vraisemblance classique chaque fois que c'est possible. Par exemple, si la sortie du modèle est complexe, mais peut s'ajuster à une loi de probabilité connue, on pourra prendre la vraisemblance associée à la loi de probabilité ajustée à la sortie

du modèle. Différentes mesures de vraisemblance ont été comparées par Freer *et al.* (1996) : les résultats obtenus sont nettement influencés par le choix de la mesure de vraisemblance.

La méthode GLUE introduite par Beven et Binley (1992) inclue également une procédure de rejet de paramètres non acceptables. Un paramètre est dit non-acceptable si la mesure de vraisemblance associée est trop faible. Cette procédure de rejet nécessite la définition d'un seuil de mesure de vraisemblance en dessous duquel les paramètres sont rejetés. La définition d'un tel seuil est subjective. Cameron *et al.* (2000c) proposent un critère fondé sur le théorème de déviance (voir (Saporta, 1990) ou (Coles, 2001)) pour choisir le seuil de manière objective.

Kuczera et Parent (1998) comparent une méthode MCMC avec une méthode d'échantillonnage préférentiel, sur une étude de cas théorique. Dans le cas étudié, ils montrent que la méthode MCMC conduit à des intervalles de crédibilité proches des intervalles de crédibilité théoriques, tandis que les intervalles de crédibilité de l'échantillonnage préférentiel montrent de forts écarts avec les intervalles de crédibilité théoriques.

On doit garder en mémoire que les résultats d'incertitudes obtenus par ces méthodes, dépendent des données d'entrée du modèle, des réponses du modèle, du choix de la distribution a priori des paramètres, du choix de la mesure de vraisemblance, et des observations impliquées dans le calcul de la vraisemblance.

Comparaison des deux méthodes sur un exemple

On compare les taux de recouvrement des intervalles de confiance obtenus par les méthodes MCMC et échantillonnage préférentiel, et pour un nombre donné d'itérations.

La comparaison est menée sur la base d'un échantillon de taille $n = 10$, et issu d'une population de loi GEV de paramètres $\alpha = 1, \beta = 0, k = -0.05$. Dans les deux cas, les lois a priori des paramètres sont les suivantes : loi lognormale de paramètres (0,100) pour α, loi normale de paramètres (0,100) pour β, et loi uniforme sur [-1,1] pour k. L'algorithme MCMC utilisé ici est l'algorithme de Metropolis (algorithme dans sa version simple, sans mise à jour de la variance).

Théoriquement, les intervalles de crédibilité obtenus avec les deux méthodes doivent être identiques, puisque les lois a priori et les vraisemblances utilisées dans les deux cas sont identiques. Néanmoins, dans le cas de l'échantillonnage préférentiel, il est nécessaire d'échantillonner de manière suffisante l'espace des paramètres pour obtenir le taux théorique de recouvrement. Dans le cas de l'algorithme MCMC, le problème est différent : on doit vérifier la convergence de l'algorithme pour s'assurer du fait que la distribution simulée par l'algorithme est bien la distribution a posteriori.

Les résultats sont présentés dans le tableau 2.1. Les taux de recouvrement ne sont pas exactement égaux à 90% puisque les intervalles de crédibilité des quantiles sont des intervalles de confiance des quantiles, considérés comme des variables aléatoires : ce ne sont pas des intervalles de confiance de l'estimateur des quantiles. Rien ne justifie que le taux de recouvrement de la vraie valeur d'un quantile (considéré réel fixé), par un intervalle de crédibilité du quantile (considéré aléatoire) à 90% soit égal à 90%.

Pour comparer les résultats, on suppose que le taux de recouvrement théorique est estimé lorsqu'il est stable avec le nombre d'itérations. La stabilité est atteinte entre 20000 et 30000

méthode	nombre d'itérations	$T = 2$	$T = 56$	$T = 10$	$T = 100$	$T = 1000$
MCMC	5000	0.087	0.083	0.084	0.087	0.088
MCMC	10000	0.089	0.083	0.084	0.088	0.091
MCMC	20000	0.090	0.085	0.087	0.089	0.091
MCMC	30000	0.091	0.086	0.088	0.090	0.092
MCMC	50000	0.091	0.086	0.088	0.090	0.092
MCMC	70000	0.091	0.087	0.088	0.090	0.092
Echantillonnage préférentiel	5000	0.082	0.074	0.073	0.079	0.081
Echantillonnage préférentiel	10000	0.087	0.082	0.081	0.085	0.087
Echantillonnage préférentiel	20000	0.090	0.085	0.085	0.088	0.090
Echantillonnage préférentiel	30000	0.090	0.084	0.086	0.088	0.090
Echantillonnage préférentiel	50000	0.091	0.086	0.087	0.089	0.090
Echantillonnage préférentiel	70000	0.091	0.087	0.087	0.090	0.091
Echantillonnage préférentiel	90000	0.091	0.086	0.087	0.090	0.091

Tab. 2.1: Taux de recouvrement des quantiles de période de retour 2, 5, 10, 100, 1000 ans, de la loi GEV de paramètres $\alpha = 1, \beta = 0, k = -0.05$, par les intervalles de crédibilité à 90% estimés par un algorithme de Metropolis, et par échantillonnage préférentiel.

itérations pour l'algorithme de Metropolis et entre 50000 et 70000 itérations pour l'algorithme de l'échantillonnage préférentiel. Cette comparaison montre que la convergence de la méthode de l'échantillonnage préférentiel nécessite un plus grand nombre d'itérations que la méthode MCMC. En terme de temps de calcul, les deux méthodes sont équivalentes. Dans l'échantillonnage préférentiel, les paramètres sont tirés dans leur loi a priori, et les vraisemblances correspondantes sont calculées. Dans la méthode MCMC, les paramètres sont tirés par une loi de saut, et acceptés ou rejetés en fonction de leur probabilité a posteriori.

D'autre part, dans la méthode de l'échantillonnage préférentiel, il est difficile de déterminer si le nombre d'itérations effectuées est suffisant. Au contraire, dans la méthode MCMC, il existe une statistique R permettant de vérifier la convergence de l'algorithme (Gelman *et al.*, 1997). Après ces considérations, nous avons décidé d'utiliser les méthodes MCMC pour analyser les incertitudes des modèles. Dans la suite, nous utiliserons une méthode MCMC proposée par Renard *et al.* (2006), combinant un algorithme de Gibbs-Metropolis (pour son efficacité) et un algorithme de Metropolis (pour sa rapidité). Cette combinaison d'algorithmes est également présentée dans l'article (Muller *et al.*, 2006) en annexe.

2.2 Intervalles de confiance pour les statistiques d'ordre

Nous nous intéressons à présent aux statistiques d'ordre, dont les plus fortes sont particulièrement importantes dans l'étude des valeurs extrêmes.

Statistiques d'ordre d'un échantillon

On note X la variable aléatoire étudiée, de fonction de répartition F_X, et de densité f_X. Soient $X_{1,n} \leq X_{2,n} \leq \ldots \leq X_{n,n}$ les n statistiques d'ordre d'un échantillon de variables aléatoires i.i.d. de distribution F_X. Les densités $f_{X_{i,n}}$ et les fonctions de répartition $F_{X_{i,n}}$ des statistiques d'ordre $X_{i,n}, i = 1, \ldots, n$ sont (Embrechts *et al.*, 1997), (Naulet, 2002) :

$$f_{X_{i,n}}(x) = \frac{n!}{(i-1)!(n-i)!} F_X(x)^{i-1}[1 - F_X(x)]^{n-i} f_X(x), \tag{2.18}$$

$$F_{X_{i,n}}(x) = \sum_{k=0}^{n-i} C_n^k F_X(x)^{n-k}[1 - F_X(x)]^k \tag{2.19}$$

Notons $x_{i,n,\alpha/2}, x_{i,n,1-\alpha/2}$, les quantiles d'ordres $\alpha/2$ et $1-\alpha/2$ de $X_{i,n}$. $[x_{i,n,\alpha/2}, x_{i,n,1-\alpha/2}]$ est donc un intervalle de confiance de $X_{i,n}$, de fréquence de recouvrement $1-\alpha$.

Statistiques d'ordre des fréquences

On peut aussi considérer les fréquences associées aux statistiques d'ordre : $F_{i,n} = F_X(X_{i,n})$ (Weibull, 1939). $F_{i,n}$ suit une loi Beta de paramètres $(i, n-i+1)$.

En effet : soit $u \in [0,1]$

$$\begin{aligned} P(F_{i,n} \leq u) &= \int_{\mathbb{R}} \mathbb{1}_{F_X(x)\leq u} \frac{n!}{(i-1)!(n-i)!} F_X(x)^{i-1}[1-F_X(x)]^{n-i} f_X(x)dx \\ &= \int_0^1 \mathbb{1}_{y\leq u} \frac{n!}{(i-1)!(n-i)!} y^{i-1}(1-y)^{n-i}dy. \end{aligned} \tag{2.20}$$

Soit $f_{i,n,\alpha/2}$ et $f_{i,n,1-\alpha/2}$ les quantiles d'ordres $\alpha/2$ et $1-\alpha/2$ de $F_{i,n} = F_X(X_{i,n})$. On montre que $F_X(x_{i,n,p}) = f_{i,n,p}$ pour tout $p \in [0,1]$.

En effet :

$$\begin{aligned} P\{F_{i,n} \leq F_X(x_{i,n,p})\} &= P\{F_X(X_{i,n}) \leq F_X(x_{i,n,p})\} \\ &= P(X_{i,n} \leq x_{i,n,p}) = p \end{aligned} \tag{2.21}$$

car F_X est croissante.

On a donc les résultats suivants :

– la loi de $F_{i,n} = F_X(X_{i,n})$ est indépendante de la loi de X. Dans une série ordonnée d'observations, on peut alors associer à chaque statistique d'ordre $X_{i,n}$ un intervalle de confiance sur sa fréquence, sans connaître la loi que suit la variable observée.

–

$$\begin{gathered} F_X([x_{i,n,\alpha/2}, x_{i,n,1-\alpha/2}]) = [f_{i,n,\alpha/2}, f_{i,n,1-\alpha/2}] \\ \text{ou} \\ [x_{i,n,\alpha/2}, x_{i,n,1-\alpha/2}] = F_X^{-1}([f_{i,n,\alpha/2}, f_{i,n,1-\alpha/2}]) \end{gathered} \tag{2.22}$$

l'image par F_X de l'intervalle de confiance de la i-ème statistique d'ordre $X_{i,n}$ ne dépend donc pas de la loi de X.

Connaissant F_X, on peut calculer les intervalles $[x_{i,n,\alpha/2}, x_{i,n,1-\alpha/2}]$ de manière analytique (via la relation 2.22), ou par simulation Monte-Carlo d'échantillons de loi F_X. Si F_X est inconnue, une méthode (non paramétrique) du bootstrap consiste à créer des échantillons de même taille n que l'échantillon de départ, par tirage avec remise dans l'échantillon de départ. Si B échantillons ont été créés, on dispose de B valeurs pour chaque statistique d'ordre. L'intervalle de confiance de chaque statistique d'ordre est calculé en ôtant les valeurs les plus faibles et les plus fortes.

On dispose donc de deux types d'intervalles de confiance, pour la description des statistiques d'ordre :

– les intervalles de confiance sur la fréquence de chaque statistique d'ordre, ils ne dépendent pas de la loi suivie par les valeurs de la série,

– les intervalles de confiance sur les statistiques d'ordre.

2.2.1 Exemple d'application

Soit un échantillon de taille $n = 100$. On calcule d'abord les intervalles de confiance des fréquences empiriques $[f_{i,n,\alpha/2}, f_{i,n,1-\alpha/2}]$, à l'aide des quantiles de la loi Beta. Supposons que la loi de l'échantillon est une loi Gumbel standard ($\alpha = 1, \beta = 0, k = 0$). On calcule alors les intervalles de confiance des statistiques d'ordre via la relation 2.22. Les intervalles de confiance obtenus ne reflètent que la variabilité des statistiques d'ordre d'un échantillon de taille $n = 100$ d'une loi Gumbel standard.

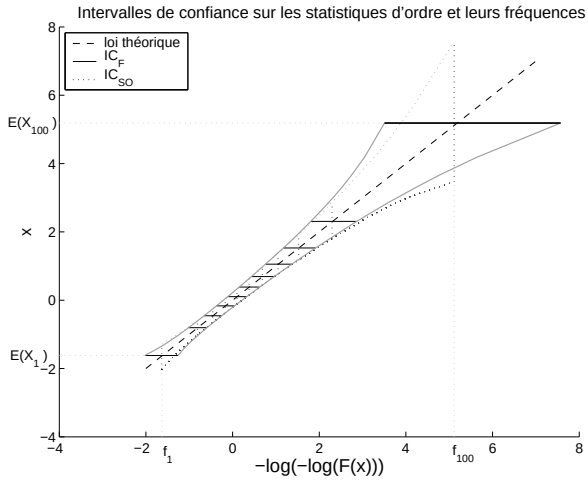
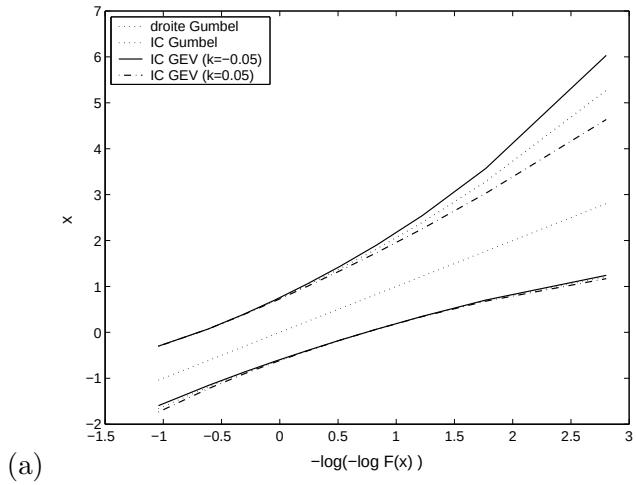


Fig. 2.1: Intervalles de confiance à 90% pour les statistiques d'ordre $X_{i,n}$: IC_{SO}, et pour leurs fréquences associées $F_{i,n}$: IC_F.

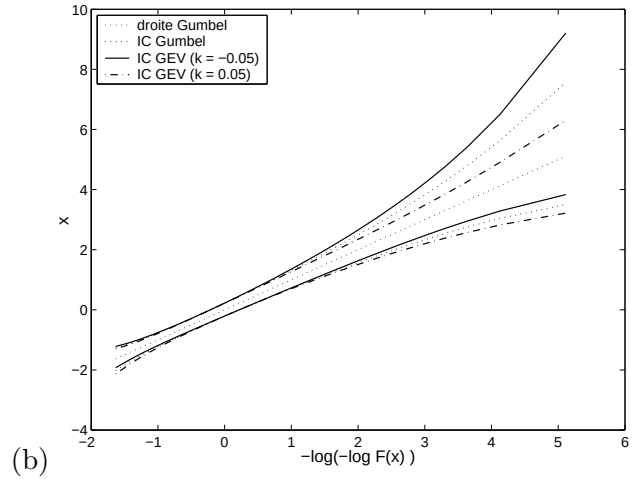
La figure 2.1 illustre les intervalles de confiance à 90% des statistiques d'ordre $X_{i,n}$ et de leurs fréquences associées $F_{i,n}$ pour un échantillon de taille n=100 d'une population de loi Gumbel. La largeur des intervalles de confiance sur les fréquences $F_{i,n}$ montre la forte incertitude des fréquences associées aux événements extrêmes observés. C'est pourquoi dans un graphe, on n'accordera pas une trop forte importance au positionnement des observations extrêmes. De la même manière, la largeur des intervalles de confiance des statistiques d'ordre montre la grande incertitude sur les valeurs observées les plus fortes. L'intervalle de confiance à 90% de la valeur la plus forte de l'échantillon est [3.5, 7.6], alors que $E(X_{100,100}) = 5.2$. L'écart relatif entre la valeur moyenne de $X_{100,100}$ et la borne supérieure de son intervalle de confiance à 90% est prés de 50%. De manière générale, on peut associer un *rectangle de confiance* à une statistique d'ordre, montrant l'incertitude relative à sa fréquence et sa valeur.

On compare ensuite les intervalles de confiance des statistiques d'ordre issues d'échantillons de loi Gumbel standard ($\alpha = 1, \beta = 0, k = 0$), de loi GEV de paramètres $\alpha = 1, \beta = 0, k = -0.05$, et de loi GEV de paramètres $\alpha = 1, \beta = 0, k = 0.05$. Les intervalles de confiance sont comparés pour des tailles d'échantillon $n = 10$ et 100.

La figure 2.2 montre la forte influence du choix du paramètre de forme de la GEV. Sauf pour les très petites fréquences ($F_X(x) \leq 0.37$, c'est-à-dire $T \leq 1.58$ ans), la largeur des intervalles de confiance augmente lorsque k diminue. On retrouve la remarque de Lu et Stedinger (1992) : sur les petites fréquences et pour un paramètre de forme négatif, la variance du quantile de la loi GEV est moins forte que celle du quantile de la loi Gumbel.



(a)



(b)

Fig. 2.2: Intervalles de confiance à 90% pour les statistiques d'ordre d'un échantillon de taille (a) n=10, (b) n=100, d'une population de loi GEV, pour différents paramètres de forme (Gumbel standard, GEV de paramètres $\alpha = 1, \beta = 0, k = -0.05$, GEV de paramètres $\alpha = 1, \beta = 0, k = 0.05$).

Bâ *et al.* (2001) utilisent cette méthode en remplaçant la loi théorique F par la loi $\widehat{F}$ estimée sur un échantillon. Cela permet de construire des intervalles de confiance des statistiques d'ordre d'un échantillon donné, en supposant que la loi estimée de l'échantillon est la loi de l'échantillon. Supposer que la loi estimée est la loi réelle de l'échantillon est une hypothèse forte. L'incertitude des statistiques d'ordre peut être mieux décrite, si l'on accepte de reconnaître le caractère incertain des paramètres. On présente cela dans le paragraphe suivant, dans un cadre bayésien.

Enfin, la figure 2.3 compare les intervalles de confiance des statistiques d'ordre avec les intervalles de confiance des quantiles : les intervalles de confiance des statistiques d'ordre deviennent plus larges que les intervalles de confiance des quantiles, lorsque l'on s'approche des plus grandes statistiques d'ordre. Ceci est dû au fait que dans le cas des quantiles, on fait l'hypothèse supplémentaire sur la loi de probabilité de la variable échantillonnée.

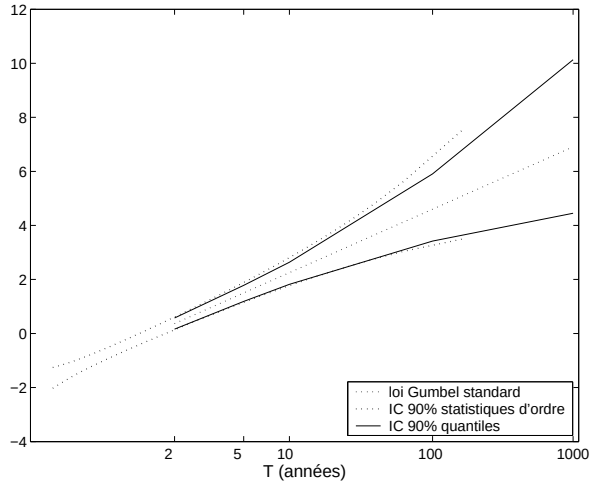


Fig. 2.3: Comparaison des intervalles de confiance des quantiles de la loi Gumbel standard avec les intervalles de confiance des statistiques d'ordre (pour un échantillon de taille $n = 100$). Les fréquences empiriques de l'échantillon sont calculées par la formule de Cunnane (1978) :$\widehat{f}_i = \frac{i-0.4}{n+0.2}$, d'où une fréquence $\widehat{f}_{100,100}$= 0.994 pour la valeur maximale de l'échantillon de taille 100, et une période de retour $\frac{1}{1-\widehat{f}_{100,100}}$=167 ans.

Intervalle de confiance "global" pour les statistiques d'ordre On définit l'intervalle de confiance "global" par la réunion des intervalles de confiance à 90% des $X_{i,n}$, pour $i = 1, \ldots, n$. Le but de ce paragraphe est simplement de montrer que la réunion des intervalles de confiance de niveau 1-α des statistiques d'ordre ne constitue pas un intervalle de confiance de niveau 1-α de l'échantillon. Ce résultat est intuitif, puisque les statistiques d'ordre d'un échantillon ne sont pas indépendantes.

Par exemple, sur 5000 simulations d'échantillons de taille 100 et de loi Gumbel standard, le taux de simulations pour lesquelles les 100 statistiques d'ordre sont toutes dans leurs intervalles de confiance à 95% respectifs est de 45%, ce qui est largement supérieur à $95\%^{100} = 0.6\%$, mais bien inférieur au taux local de 95%. Pour des n plus grands (n=200,500,800), il semble que ce pourcentage converge vers 35 %. Si le niveau des intervalles de confiance est de

99%, le taux de simulations pour lesquelles les n statistiques d'ordre sont toutes dans leurs intervalles de confiance est de 91% si $n = 20$, 86% si $n = 50$, 82% si $n = 100$.

Comme attendu intuitivement, le taux de l'intervalle de confiance "global" pour toutes les statistiques d'ordre n'est donc pas égal au produit des taux des intervalles de confiance des statistiques d'ordre.

La densité jointe des statistiques d'ordre (Embrechts *et al.*, 1997)

$$f_{X_{1,n},\ldots,X_{n,n}}(x_1,\ldots,x_n) = n!\prod_{i=1}^{n} f_X(x_i) \text{ si } x_1 \leq,\ldots,\leq x_n \tag{2.23}$$

$$= 0 \text{ sinon,} \tag{2.24}$$

peut permettre de calculer l'intervalle de confiance pour l'ensemble des statistiques d'ordre, par simulation, par exemple avec une technique de Metropolis (Hastings, 1970).

2.2.2 Distribution a posteriori des statistiques d'ordre

Soit un échantillon d'une loi F_θ, paramétrée par θ. On cherche à étudier l'incertitude liée au choix du paramètre θ. Si θ est connu, la densité $f_{X_{i,n}|\theta}$ de $X_{i,n}$ est donnée par l'expression 2.18. Dans ce paragraphe, on suppose θ aléatoire. Étant donné l'échantillon $\mathbf{x}$, on note $f(\theta|\mathbf{x})$ la loi a posteriori de θ. La densité a posteriori de $X_{i,n}$ est alors :

$$f_{X_{i,n}}(x|\mathbf{x}) = \int f_{X_{i,n}|\theta}(x,\theta) f(\theta|\mathbf{x}) d\theta. \tag{2.25}$$

On peut calculer $P(X_{i,n} \leq x|\mathbf{x})$ de manière analytique :

$$\begin{aligned} P(X_{i,n} \leq x|\mathbf{x}) &= \int_0^x f_{X_{i,n}}(u|\mathbf{x}) du \\ &= \int_0^x \int f_{X_{i,n}|\theta}(u,\theta) f(\theta|\mathbf{x}) d\theta du \\ &= \int f(\theta|\mathbf{x}) P(X_{i,n} \leq x|\theta) d\theta \\ &= \int f(\theta|\mathbf{x}) P\{B_{i,n} \leq F_\theta(x)\} d\theta, \end{aligned} \tag{2.26}$$

avec $B_{i,n}$ une variable aléatoire de loi Beta$(i, n-i+1)$. On peut également calculer $P(X_{i,n} \leq x|\mathbf{x})$ par simulation, d'après 2.26, on a :

$$P(X_{i,n} \leq x|\mathbf{x}) = E_{\theta|\mathbf{x}}\{P(X_{i,n} \leq x|\theta)\} \tag{2.27}$$

$$\approx \frac{1}{L}\sum_{l=1}^{L} P(X_{i,n} \leq x|\theta_l) \tag{2.28}$$

$$\approx \frac{1}{L}\sum_{l=1}^{L}\frac{1}{L'}\sum_{l'=1}^{L'} \mathbb{1}_{X_{l,l',i,n} \leq x} \tag{2.29}$$

$$\tag{2.30}$$

avec θ_l tiré dans la loi a posteriori de θ, et $X_{l,l',i,n}$ un tirage de la $i$ème statistique d'ordre d'un échantillon de taille n, et de loi F_{θ_l}.

On compare les incertitudes liées uniquement à l'échantillonnage (méthode de Bâ *et al.* (2001)) avec les incertitudes liées à l'échantillonnage et l'estimation des paramètres (méthode présentée ici). Dans notre exemple, F_θ est la famille des distributions de valeurs extrêmes (GEV). On considère un échantillon de taille n=10, ou 100. On applique dans un premier temps la méthode de Bâ *et al.* (2001) : on suppose que la loi de $X_{i,n}$ est donnée par l'équation 2.19, avec la distribution estimée de l'échantillon. Dans un deuxième temps, on ne suppose plus que θ est fixé par l'estimation, mais on suppose que θ suit une loi a posteriori, simulée par un algorithme MCMC. La loi a priori choisie pour les paramètres est large : loi lognormale de paramètres (0,100) pour α, loi normale de paramètres (0,100) pour β, loi uniforme sur [-1,1] pour k. La convergence de l'algorithme MCMC est vérifiée avec la statistique R de Gelman *et al.* (1997). On vérifie que l'algorithme a convergé après 6000 itérations de l'algorithme MCMC dans le cas $n = 10$ et après 100 itérations dans le cas n=100. On utilise 10000 valeurs simulées après convergence de l'algorithme. On présente les résultats dans les cas où n=10 et 100, sur la figure 2.4.

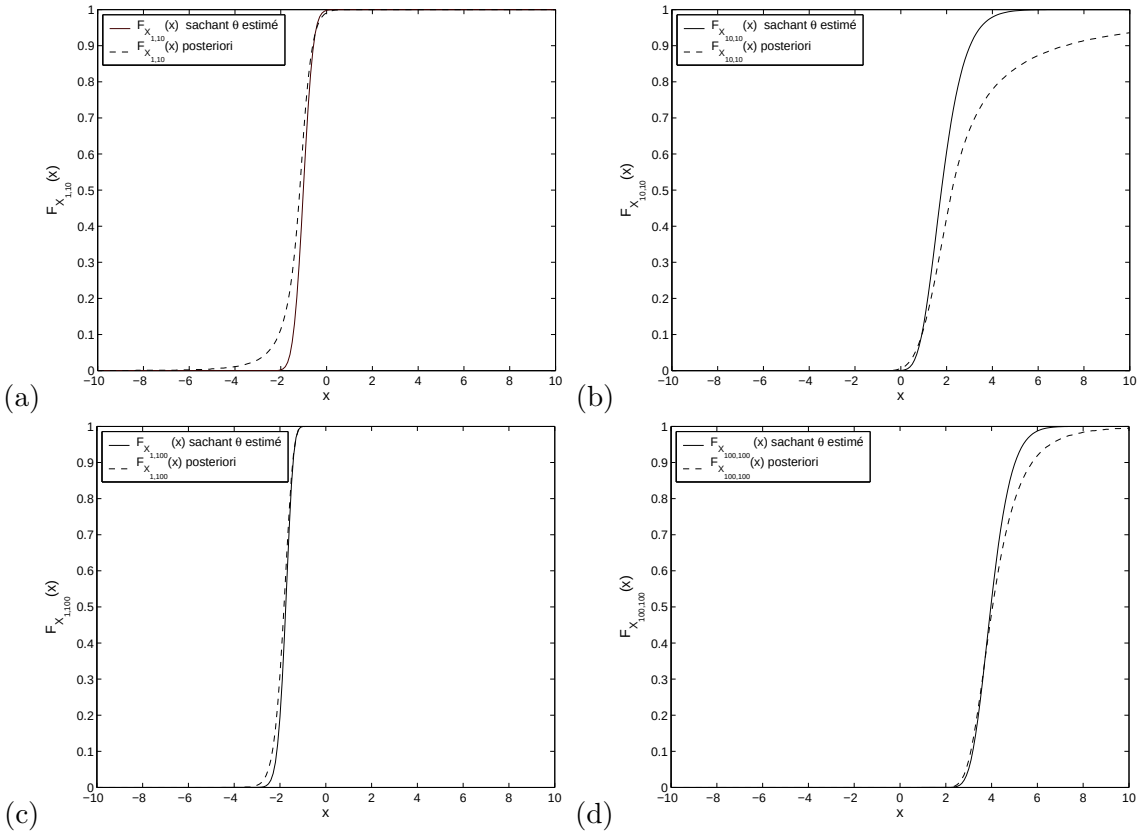


Fig. 2.4: Estimation de la fonction de répartition des statistiques d'ordre (a) $X_{1,10}$ et (b) $X_{10,10}$ d'un échantillon de taille $n = 10$, (c) $X_{1,100}$, et (d) $X_{100,100}$ d'un échantillon de taille $n = 100$. Comparaison du cas où la distribution de l'échantillon est supposée fixée à la distribution GEV ajustée à l'échantillon, et du cas où la distribution de l'échantillon est seulement supposée appartenir à la famille GEV, avec des paramètres aléatoires suivant la loi a posteriori.

L'allure de la fonction de répartition du cas bayésien, sans loi d'échantillon $F_{\widehat{\theta}}$ fixée au départ, reflète une plus grande incertitude que dans le cas de la méthode avec $F_{\widehat{\theta}}$ fixée.

2.3 Incertitudes d'estimation des paramètres de la loi GEV

Les incertitudes d'estimation des paramètres de la GEV sont très grandes sur le paramètre de forme, même pour une taille d'échantillon égale à 100. Pour illustrer ce fait, on a calculé les intervalles de confiance et les médianes des estimateurs de maximum de vraisemblance des paramètres de la GEV de paramètres $\alpha = 1, \beta = 0, k = -0.05$, pour des tailles n d'échantillon allant de 10 à 5000. Pour le calcul de ces quantités, on a procédé, pour chaque taille n, par simulation de 1000 échantillons de taille n. Le tableau 2.2 donne les résultats de cette étude. La distribution semble symétrique pour les grands échantillons, confirmant le résultat connu de normalité asymptotique des estimateurs de maximum de vraisemblance. Pour les petits échantillons, la distribution des estimateurs de maximum de vraisemblance est asymétrique. En particulier, la médiane de $\widehat{k}$ est inférieure à la moyenne de $\widehat{k}$. La même étude est reproduite avec $k = 0$ (tableau 2.3) et $k = 0.05$ (tableau 2.4). Qu'il soit négatif, positif ou nul, le paramètre de forme k présente une asymétrie positive (la médiane est à gauche du milieu des intervalles de confiance), comme le montrent les résultats des tableaux 2.2, 2.3 et 2.4.

n	$\hat{\alpha}$	$\hat{\beta}$	$\hat{k}$
5000	1.00(0.98,1.01)	0.00(-0.02,0.02)	-0.05(-0.06,-0.03)
1000	1.00(0.96,1.04)	0.00(-0.06,0.06)	-0.05(-0.08,-0.01)
100	0.98(0.85,1.13)	0.01(-0.17,0.19)	-0.05(-0.17,0.08)
10	0.88(0.43,1.92)	0.02(-0.55,0.67)	-0.01(-0.72,1.16)

Tab. 2.2: Médiane (avec intervalles de confiance à 90%) des estimateurs de maximum de vraisemblance des paramètres de la GEV de paramètres $\alpha = 1, \beta = 0, k = -0.05$, pour différentes tailles n d'échantillon. Valeurs calculées par simulation de 1000 échantillons de taille n.

n	$\hat{\alpha}$	$\hat{\beta}$	$\hat{k}$
1000	1.00(0.96,1.04)	0.00(-0.06,0.06)	0.00(-0.03,0.03)
100	0.98(0.85,1.13)	0.01(-0.17,0.19)	0.00(-0.11,0.13)
10	0.87(0.44,1.69)	0.01(-0.58,0.74)	0.04(-0.64,1.09)

Tab. 2.3: Médiane (avec intervalles de confiance à 90%) des estimateurs de maximum de vraisemblance des paramètres de la loi Gumbel standard ($\alpha = 1, \beta = 0, k = 0$), pour différentes tailles n d'échantillon. Valeurs calculées par simulation de 1000 échantillons de taille n.

n	$\hat{\alpha}$	$\hat{\beta}$	$\hat{k}$
1000	1.00(0.96,1.04)	0.00(-0.06,0.06)	0.05(0.02,0.08)
100	0.98(0.86,1.12)	0.01(-0.17,0.19)	0.05(-0.06,0.18)
10	0.89(0.45,2.20)	-0.01(-0.54,0.64)	0.10(-0.58,1.18)

Tab. 2.4: Médiane (avec intervalles de confiance à 90%) des estimateurs de maximum de vraisemblance des paramètres de la loi Gumbel standard ($\alpha = 1, \beta = 0, k = 0.05$), pour différentes tailles n d'échantillon. Valeurs calculées par simulation de 1000 échantillons de taille n.

2.4 Conclusions

Nous avons présenté des méthodes statistiques pour analyser les incertitudes d'échantillonnage et d'estimation dans un contexte paramétrique.

- Les incertitudes des estimateurs des quantiles de lois de probabilités peuvent être évaluées différemment, suivant la méthode d'estimation des paramètres (par maximum de vraisemblance, moments, moments pondérés). Des théorèmes de normalité asymptotiques ont été formulés dans le cas des estimations par maximum de vraisemblance, mais également dans le cas particulier des lois des valeurs extrêmes dont les paramètres sont estimés par la méthode des moments pondérés. Un théorème de déviance permet, dans le cas de l'estimation par maximum de vraisemblance, d'estimer des régions de confiance d'un sous-ensemble de paramètres. Il existe également des méthodes non paramétriques données par le bootstrap et ses variantes. Nous avons proposé une revue bibliographique sur les estimations des variances de quantiles pour des lois particulières utilisées en hydrologie.
- D'un point de vue hydrologique, des méthodes ont été proposées pour des modèles plus complexes que de simples lois de probabilité (méthodes RSA et GLUE). Ces méthodes sont des généralisations de méthodes bayésiennes, justifiées de manière heuristique. A notre connaissance, la justification mathématique n'est pas encore démontrée.
- Contrairement aux approches fréquentielles, les méthodes bayésiennes supposent que les paramètres des modèles sont aléatoires. L'intérêt de ces méthodes est de permettre d'inclure des informations a priori sur les paramètres (lorsque l'on en possède) et de considérer à la fois les incertitudes d'échantillonnage (dans la vraisemblance) et de modélisation (par la loi a priori). De plus, les méthodes bayésiennes ne reposent pas sur des hypothèses asymptotiques, comme dans le cas de la normalité asymptotique des estimateurs du maximum de vraisemblance. Malgré ces qualités, les méthodes bayésiennes ont quelques limites : puisque la loi a priori apporte des informations qui influencent la loi a posteriori, les résultats obtenus a posteriori sont dépendants de la loi a priori, d'où une certaine subjectivité des résultats. D'autre part, les lois a posteriori sont généralement impossibles à simuler de façon simple, et l'on doit recourir à des méthodes assez lourdes pour simuler les lois a posteriori. Deux grands types de méthodes permettent de mettre en pratique les estimations bayésiennes : les algorithmes MCMC et l'échantillonnage préférentiel. Un exemple d'application des deux méthodes est donné. D'après les résultats obtenus, et en accord avec d'autres études données dans la bibliographie, nous préférons utiliser les méthodes MCMC.
- Dans l'étude des valeurs extrêmes, nous sommes amenés à considérer les valeurs les plus fortes des échantillons et des graphiques en fréquence des observations et des quantiles estimés. Les statistiques d'ordre et les fréquences empiriques ont donc un rôle descriptif

important pour les valeurs extrêmes. Comme on peut s'y attendre les statistiques d'ordre les plus fortes d'un échantillon et les fréquences des plus fortes valeurs d'un échantillon sont très incertaines, et nous avons quantifié cette incertitude.

– Enfin, puisque la loi GEV est particulièrement importante dans la théorie des valeurs extrêmes, nous avons considéré l'incertitude d'échantillonnage des estimateurs de ses paramètres, lorsque ceux-ci sont estimés par maximum de vraisemblance. L'incertitude des estimateurs est considérable, notamment l'incertitude de l'estimateur du paramètre de forme. Le paramètre de forme étant déterminant pour le calcul des quantiles extrêmes, l'incertitude sur le paramètre de forme se propage dans l'incertitude sur les quantiles extrêmes. Cette forte incertitude du paramètre de forme a conduit un grand nombre d'hydrologues à préférer la loi Gumbel à la loi GEV, en particulier avec les tailles d'échantillons couramment employés en hydrologie.

Deuxième partie

Applications à l'analyse des valeurs extrêmes de longues séries pluviométriques

Chapitre 3

Diagnostic sur le comportement asymptotique des pluies

3.1 Bibliographie

En hydrologie, la distribution Gumbel a été, et est encore largement utilisée pour estimer les quantiles de valeurs extrêmes. Lu et Stedinger (1992) ont montré que les quantiles de la loi Gumbel avaient une variance d'échantillonnage plus faible que les quantiles d'une loi GEV, pour des échantillons de tailles courantes en hydrologie ($n = 20$ à 100), et pour des fréquences supérieures à 0.95. Ils affirment que les quantiles de période de fréquence supérieure à 1-$1/n$ sont mieux estimés par une loi Gumbel. A l'inverse, les quantiles de période de retour inférieure à 10 ans sont estimés de façon plus précise (c'est-à-dire avec une variance plus faible) par la loi GEV à trois paramètres, plus flexible que la loi Gumbel. C'est aussi ce que nous avions remarqué en examinant la largeur des intervalles de confiance de la figure 2.2.

De plus, la plupart des distributions utilisées en hydrologie (exponentielle, gamma, Weibull, normale, lognormale) appartiennent au domaine d'attraction de la loi Gumbel. Le domaine d'attraction de la loi de Fréchet ($k < 0$) contient des distributions moins utilisées en hydrologie (Pareto, Cauchy, log-gamma).

L'usage de la loi Gumbel s'explique aussi par la simplicité de cette loi, et par son tracé linéaire dans un diagramme de Gumbel, avec une échelle en $-\ln(-\ln(F))$, où F est la probabilité cumulée. D'autre part, du fait des tailles d'échantillons disponibles en hydrologie (souvent une trentaine d'années), les tracés des points empiriques ont souvent une allure alignée sur les diagrammes de Gumbel, suggérant une loi Gumbel.

Dans le cas des pluies, plusieurs auteurs ont montré que la loi Gumbel pouvait sous-estimer les quantités de pluies les plus fortes : Wilks (1993), Koutsoyiannis et Baloutsos (2000), Chaouche *et al.* (2002), Coles *et al.* (2003), Coles et Pericchi (2003), Sisson *et al.* (2006), Koutsoyiannis (2004a,b) et Bacro et Chaouche (2006). Cet inconvénient est très important du point de vue de l'ingénierie, dans la construction des ouvrages prévus pour résister à des événements de périodes de retour 100 ans, 1000 ans, voire 100 000 ans dans certains cas. Koutsoyiannis (2004a) montre même qu'utiliser la loi Gumbel au lieu de la loi GEV revient à donner le risque maximum aux structures d'ingénierie : *'Normally, this would be a sufficient reason to avoid the use of EV1 in engineering studies'* écrit Koutsoyiannis (2004a).

Koutsoyiannis (2004a) analyse la vitesse de convergence de la distribution du maximum d'un échantillon de taille n (cf. équation 1.4), si la loi F de l'échantillon appartient au domaine d'attraction de la loi Gumbel. Avec une loi normale centrée réduite ou une loi Weibull pour F, la courbure de la distribution du maximum dans un diagramme de Gumbel est encore visible pour n de l'ordre de 10^6. En hydrologie, on ne peut pas analyser le maximum d'un si grand nombre d'événements dans une année ou une saison. On n'est donc pas dans les conditions asymptotiques, pour ces deux distributions F évoquées par Koutsoyiannis (2004a). Pour les valeurs de n rencontrées en hydrologie, la distribution GEV semble donc plus adaptée à l'allure convexe de la distribution du maximum dans le diagramme de Gumbel. Koutsoyiannis (2004a) reproduit la même étude, mais avec une distribution F appartenant au domaine d'attraction de la GEV, avec $k < 0$. La distribution du maximum d'un échantillon de taille n finie est convexe dans un diagramme de Gumbel : la loi GEV, bien qu'avec des paramètres différents de ceux de la loi GEV limite, semble encore adaptée pour modéliser la distribution du maximum d'un échantillon de taille n finie.

Par une étude des biais des estimateurs du paramètre de forme k, avec différentes méthodes (moments, moments pondérés, maximum de vraisemblance), Koutsoyiannis (2004a) montre qu'il est difficile de rejeter l'hypothèse $k = 0$ sur des échantillons de tailles 20 à 50, voire 150.

A l'échelle mondiale, Koutsoyiannis (2004b) utilise 169 séries longues de 100 à 154 années de mesures, et examine la variabilité des estimateurs du paramètre de forme k de la GEV. En comparant la variabilité des 169 estimateurs de k avec celle des 169 estimateurs de k sur des séries simulées par une loi GEV dont le paramètre de forme et le rapport des paramètres de position et d'échelle sont fixés, il conclue que la variabilité observée sur les 169 séries mondiales correspond à la variabilité statistique du paramètre de forme k, et que celui-ci pourrait être constant sur la partie du monde contenant les 169 séries (Etats-Unis, Royaume-Uni, France, Italie, Grèce). La valeur constante proposée par Koutsoyiannis (2004b) est de -0.15. D'autre part, sur 90% des 169 séries, le paramètre de forme k est estimé négatif. Si au lieu de considérer une loi Gumbel, on considère une loi GEV de paramètre de forme fixé à $k = -0.15$, on retrouve les avantages de la loi Gumbel : la variabilité des estimateurs est plus faible car k est fixé, et on peut créer des diagrammes de type diagrammes de Gumbel dans lesquels la distribution GEV de paramètre de forme $k = -0.15$ est une droite.

Enfin, dans une étude de Koutsoyiannis et Baloutsos (2000) sur la longue série pluviométrique d'Athènes (136 années), on constate que la loi Gumbel n'est pas adaptée aux maxima annuels de la série de 136 années, tandis qu'elle paraît appropriée si l'on ne considère par exemple que les 34 dernières années.

3.2 Cas d'étude

Météo-France a mis à notre disposition un grand nombre de séries pluviométriques : 308 séries de la base de données des Séries Quotidiennes de Référence (SQR), et 47 séries de la Banque Pluvio. Les Séries Quotidiennes de Référence sont des séries dont certains critères de qualité ont été vérifiés :

- le taux de données quotidiennes manquantes est inférieur à 10%
- l'amplitude de la plus grosse rupture[1] en valeur absolue est inférieure à 10%

[1]Une rupture est un changement brutal de la valeur d'un paramètre descriptif d'une série chronologique. Dans l'étude des pluies journalières, une rupture peut être due à une variation du climat, un changement de capteur, un déplacement, un changement de l'environnement de mesure, etc.

– la somme des amplitudes des ruptures est inférieure à 10%
– le plus grand déplacement horizontal est inférieur à 10 km
– le plus grand déplacement vertical est inférieur à 50 m
– le numéro du poste ne doit pas avoir subi de changements.

Pour l'étude du comportement asymptotique des pluies, on choisit de travailler sur des séries avec plus de 100 ans de mesures. On a posé ce choix en raison des difficultés d'estimation du paramètre de forme de la loi GEV sur des petites séries (voir le chapitre 1).

On retient donc 39 séries de plus de 100 ans, réparties sur 14 départements français (02, 08,12,13,19,21,25,33,40,46,48,75,87,89) : 7 Séries Quotidiennes de Référence et 32 séries de la Banque Pluvio.

Les séries sont échantillonnées par maximum annuel, sur l'année civile (janvier-décembre). Avant d'aller plus loin dans l'étude des séries, on vérifie que celles-ci sont stationnaires au second ordre. En effet, sur des périodes de mesures de plus de 100 ans, les séries peuvent être affectées d'une variabilité climatiques ou d'erreurs métrologiques.

Un test de stationnarité simple consiste à appliquer le test du rapport de vraisemblance, ou test de déviance (Coles, 2001) entre deux distributions emboîtées. Les distributions emboîtées utilisées dans notre cas sont les suivantes : le modèle stationnaire est une distribution GEV de paramètres d'échelle, position, forme notés α_s, β_s, k_s. Le modèle non stationnaire est une distribution GEV avec une tendance linéaire sur ses paramètres de position et d'échelle, et un paramètre de forme fixe : $\alpha_{tend} = \alpha_s + t\alpha', \beta_{tend} = \beta_s + t\beta', k_{tend} = k_s$. Le modèle stationnaire est un sous modèle du modèle non-stationnaire. La modélisation non-stationnaire est empruntée à celle de Coles (2001).

Après application du test, on ne garde que les séries dont l'hypothèse de non-stationnarité ($\alpha' = 0, \beta' = 0$) n'est pas rejetée, au niveau 5%. On garde ainsi 6 Séries Quotidiennes de Référence, et 21 séries de la Banque Pluvio. En fait, 5 Séries Quotidiennes de Référence sont issues de postes déjà présents dans les 21 séries conservées de la Banque Pluvio. Pour ne pas répéter l'information de deux mêmes postes, on ne garde que 22 postes, dont 6 sont référencés dans les Séries Quotidiennes de Référence. Les 22 postes sont répartis dans 11 départements français : 02, 13, 21, 25, 33, 40, 46, 48, 75, 87, 89. La répartition des postes par département est présentée dans le tableau 3.1.

département	nombre de séries dans le département
02	1
13	2
21	4
25	1
33	2
40	1
46	1
48	6
75	1
87	2
89	1

Tab. 3.1: Répartition des 22 longues séries stationnaires dans les départements français.

Les maxima annuels de ces postes sont modélisés par une loi GEV (stationnaire), et analysés dans un cadre bayésien. Les lois a priori choisies pour les paramètres sont larges, et courantes en hydrologie (Coles et Pericchi, 2003) : la loi lognormale de paramètres (0,100)

pour α_s, la loi normale de paramètres (0,100) pour β_s, et la loi uniforme sur [-1,1] pour k. Les résultats sont présentés sur la figure 3.1.

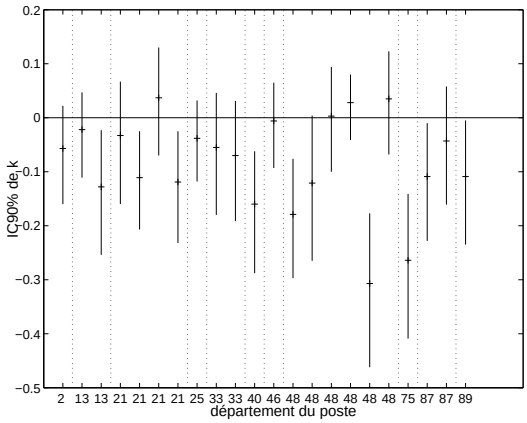


Fig. 3.1: Intervalles de crédibilité à 90% du paramètre de forme k de la loi GEV, modèle des maxima annuels journaliers de 22 postes possédant une longue chronique de mesures (plus de 100 ans).

Sur les 22 séries, 9 séries ne contiennent pas la valeur 0 dans l'intervalle de crédibilité du paramètre de forme k, et aucune série ne possède un intervalle de crédibilité entièrement inclus dans $\mathbb{R}^+$. Donc tester $k < 0$ revient à tester si 0 n'appartient pas à l'intervalle de crédibilité de k. Du fait de la dépendance des postes, on ne peut pas conclure que le taux $p_{k<0}$ de postes vérifiant $k < 0$ est égal à $9/22 \approx 41\%$. Pour avoir une idée de la valeur de ce taux, et pour s'affranchir de la dépendance entre postes du même département, on échantillonne un seul poste par département.

Si on échantillonne les séries en ne gardant qu'un seul poste par département, on peut créer $1 \cdot 2 \cdot 4 \cdot 1 \cdot 2 \cdot 1 \cdot 1 \cdot 6 \cdot 1 \cdot 2 \cdot 1 = 192$ échantillons de 11 séries. On considère que les postes de différents départements sont indépendants. On peut alors estimer sur chaque échantillon le taux de postes $p_{k<0}$ tels que $k < 0$ par $p_{k<0} \approx \sum_{i=1}^{11} \mathbb{1}_{0 \notin IC_i(k)}/11$ où $IC_i(k)$ désigne l'intervalle de crédibilité de k d'un poste d'un des 11 départements, représenté sur la figure 3.1. Les valeurs du taux $p_{k<0}$ ainsi obtenues sont entre 27.3% et 63.6%, et leur répartition est représentée dans l'histogramme de la figure 3.2. Ces résultats montrent qu'une partie non négligeable des séries vérifie $k < 0$. Mais comme le remarquait Koutsoyiannis (2004a), l'autre partie, non négligeable, des séries vérifie $0 \in IC(k)$. Notons que sur les 22 postes, 14 postes vérifient $-0.15 \in IC(k)$, où -0.15 est la valeur du paramètre de forme k proposée par Koutsoyiannis (2004b) pour la loi GEV des pluies maximales annuelles, dans l'Hémisphère Nord, et $IC(k)$ est l'intervalle de crédibilité du paramètre de forme.

L'analyse aurait pu être améliorée, en conservant les séries avec tendance, et en reproduisant l'analyse bayésienne avec des paramètres de position et d'échelle à tendance linéaire. Néanmoins, par manque de temps, cette analyse n'a pas été réalisée. Les résultats montrés jusqu'ici laissent penser que cette analyse aboutirait certainement à la même conclusion : le paramètre de forme k est significativement négatif sur un nombre important de postes.

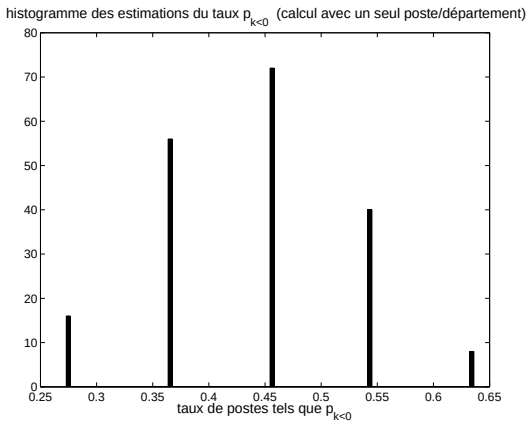


FIG. 3.2: Histogramme du taux $p_{k<0}$ de postes tels que $k < 0$, estimé sur les 192 échantillons de 11 séries appartenant à des départements français différents.

3.3 Conclusions

Même si la loi Gumbel présente certains avantages qui ont provoqué un large usage de cette loi en hydrologie (simplicité, faible variabilité des estimateurs des quantiles extrêmes), de nombreux auteurs ont récemment montré l'inadéquation de cette loi avec la distribution des maxima annuels de pluie. L'enjeu d'un tel débat entre la loi de Gumbel et la loi GEV est considérable, puisqu'il est directement lié à la sécurité des structures hydrauliques. Cependant, la variabilité de l'estimateur du paramètre de forme k de la loi GEV est telle (comme montré au chapitre 2), que les tests d'hypothèse nulle $k = 0$ contre l'hypothèse alternative $k < 0$ sont souvent acceptés, en particulier avec les tailles d'échantillons fréquemment rencontrés en hydrologie. Une autre approche, proposée par Koutsoyiannis (2004b), est de considérer une loi GEV avec k fixé à -0.15, valeur fréquente sur des estimations de 169 longues séries de l'Hémisphère Nord.

Nous avons également étudié le problème de la valeur du paramètre de forme k à partir de 22 séries pluviométriques de plus de 100 ans de mesures, fournies par Météo-France. Sur les 22 postes considérés, il apparaît clairement que l'hypothèse $k < 0$ est acceptée pour un nombre important de postes. Cependant, du fait de la variabilité de l'estimateur de k, l'hypothèse $k = 0$ est également vérifiée pour un nombre important de postes. Cela souligne la difficulté de rejeter la loi Gumbel au profit de la loi GEV.

Chapitre 4

Analyse des extrêmes de pluie de différentes durées

4.1 Introduction

L'analyse du comportement asymptotique des pluies a été présentée jusqu'ici par l'analyse des valeurs extrêmes de pluie mesurée au pas de temps journalier. Or, le comportement des extrêmes de pluie est mieux expliqué si l'on considère les extrêmes de pluie mesurée sur d'autres pas de temps (plus fins mais aussi plus grossiers). En hydrologie, le terme couramment employé pour désigner le pas de temps de mesure est 'durée'. Les courbes Hauteur-Durée-Fréquence sont un outil largement utilisé en ingénierie hydrologique, notamment dans les systèmes de drainage urbains. Elles représentent les distributions des extrêmes de pluie (échantillonnés par maxima annuels ou saisonniers, ou par valeurs supérieures à un seuil élevé). L'objectif des courbes Hauteur-Durée-Fréquence est donc d'estimer les quantiles de pluie mesurée sur un pas de temps donné, et pour une fréquence ou période de retour donnée. Si l'on dispose d'une série de mesures à un pas de temps fin (par exemple, des mesures de cumuls au pas de temps 1 h), on peut produire des séries de mesures des cumuls en 2 h, 3 h, etc. À partir de ces séries, on peut produire des séries de maxima annuels (ou saisonniers) ou de valeurs supérieures à un seuil élevé, pour différentes durées (supérieures au pas de temps de mesure initiale), et ajuster une distribution de probabilité marginale à chaque durée. Cependant, l'étude des distributions des pluies de différentes durées par leurs distributions marginales peut aboutir à des contradictions entre les estimateurs de pluie : les quantiles de pluie d'une durée $d < d'$ peuvent être supérieurs aux quantiles de pluie de durée d', ce qui est absurde.

Une analyse multivariée des échantillons de pluie de différentes durées fournit une approche plus complète. En particulier, l'analyse Hauteur-Durée-Fréquence empêche l'intersection des distributions de quantiles de différentes durées.

La première relation entre durées est apparue en 1932 (Bernard, 1932). L'approche classique de construction des courbes Hauteur-Durée-Fréquence est constituée de trois étapes (Chow *et al.*, 1988). Dans la première étape, une distribution de probabilité est ajustée à chaque échantillon de chaque durée. Dans la seconde étape, on calcule les quantiles correspondant à plusieurs périodes de retour, en utilisant la distribution ajustée à l'étape 1. Enfin, on détermine les courbes Hauteur-Durée-Fréquence en ajustant une équation paramétrique

pour chaque période de retour, par exemple avec des procédés de régression entre les quantiles et les durées. L'inconvénient de cette procédure est qu'elle nécessite un grand nombre de paramètres, et une régression fondée sur des valeurs dépendantes (puisque les quantiles estimés proviennent de la même série observée, mais avec des cumuls sur différents pas de temps).

Plusieurs modèles empiriques ont été proposés, Garcia-Bartual et Schneider (2001) donnent des références bibliographiques. Plus récemment, Burlando et Rosso (1996), de Lima et Grasman (1999), Veneziano et Furcolo (2002), Borga *et al.* (2005) ont proposé des approches empruntées aux processus multi-fractals. Toutes ces approches nécessitent moins de paramètres que l'approche classique, mais il reste encore à modéliser la dépendance entre les différentes durées.

Dans ce chapitre, on présente deux modèles : un modèle empirique classique, et un modèle empirique amélioré, modélisant conjointement les pluies journalières et les pluies mesurées sur des pas de temps multiples de l'heure.

Avant de présenter les modèles, on pose quelques notations pour fixer les idées.
- Si on note $I(t)$ l'intensité de la pluie à l'instant t, alors $Y_i(\delta) = \int_i^{i+\delta} I(t)dt$ est le cumul de pluie précipitée entre les instants i et $i+\delta$.
- Les séries de pluies de 1 h, 24 h et journalières sont donc notées respectivement $\{Y_i(1)\}$, $\{Y_i(24)\}$et $\{Y_{24i}(24)\}$.
- Les variables étudiées ici sont les cumuls maximum en d heures : $H_d = \max\{Y_i(d)\}$, et les cumuls maximum journaliers : $H_J = \max\{Y_{24i}(24)\}$. Les maxima peuvent être considérés sur l'année ou sur la saison.

On note la différence entre pluies de 24 h et pluies journalières : les pluies de 24 h $\{Y_i(24)\}$ sont mesurées toutes les heures sur des pas de temps de 24 h. Les pluies journalières $\{Y_{24i}(24)\}$ sont mesurées quotidiennement, à heure fixe (en général à 6 h T.U.) sur des pas de temps de 24 h. Pour différencier les pluies de type $\{Y_i(d)\}$ mesurées toutes les heures, des pluies journalières $\{Y_{24i}(24)\}$ mesurées une seule fois par jour, on appelle les premières **pluies horaires**, et les secondes **pluies journalières**. Les premières sont données dans une série **pluviographique**, les secondes dans une série **pluviométrique**.

4.2 Présentation des modèles

D'après la théorie des valeurs extrêmes, on considère que les distributions marginales G_d des cumuls maximum H_d en d heures sont des lois GEV. L'analyse conjointe des différentes durées est présentée dans les deux paragraphes suivants.

4.2.1 Modèle Hauteur-Durée-Fréquence pour les pluies horaires

Garcia-Bartual et Schneider (2001) présentent neuf modèles empiriques, à deux ou trois paramètres. Koutsoyiannis *et al.* (1998) généralisent les modèles proposés :

$$I_d(T) = a(T)/b(d), \tag{4.1}$$

où $I_d(T)$ désigne le quantile de période de retour T années, de l'intensité maximale annuelle (ou saisonnière) mesurée sur un pas de temps de d heures ; $b(d) = (d+\vartheta)^\epsilon$ avec $\vartheta > 0, \epsilon \in (0,1)$ et $a(T) = F^{-1}(1-1/T)$ avec F une distribution de probabilité cumulée (par exemple GEV,

lognormale, gamma, log-Pearson III, GPD) du processus d'intensité normalisé $I_d(T)b(d)$. Pour des raisons théoriques, on considère ici que F est la loi GEV, de paramètres α, β, k. Alors H_d suit une distribution GEV, dont les quantiles sont donnés par :

$$H_d(T) = dI_d(T) = d(\beta + \frac{\alpha}{k}[1 - (-\log(1 - 1/T))^k])/(d + \vartheta)^\epsilon. \tag{4.2}$$

Les paramètres d'échelle, position et forme α_d, β_d, k_d de la distribution de H_d s'expriment simplement en fonction de $\alpha, \beta, k, \vartheta, \epsilon$:

$$\alpha_d = d\alpha/(d + \vartheta)^\epsilon, \beta_d = d\beta/(d + \vartheta)^\epsilon, k_d = k. \tag{4.3}$$

Avant d'appliquer un tel modèle, on doit vérifier s'il est adapté pour l'ensemble des durées que l'on souhaite modéliser. Le cas de figure suivant peut également se réaliser : le modèle 4.2 est adapté, mais avec des paramètres différents sur différentes plages de durées.

4.2.2 Modèle Hauteur-Durée-Fréquence pour les pluies horaires et journalières

Ce second modèle complète le précédent en incluant la modélisation des pluies journalières. Le lien entre les pluies horaires et journalières est modélisé grâce à la relation entre les pluies de 24 h, appartenant aux pluies horaires, et les pluies journalières, mesurées au même pas de temps que les pluies de 24 h. En hydrologie, une relation empirique est couramment utilisée pour estimer les maxima de pluies de 24 h, à partir des maxima des pluies journalières (Hershfield, 1961), (Weiss, 1964) :

$$H_{24} = c \cdot H_J \tag{4.4}$$

où c est le facteur Hershfield. Weiss (1964) donne une estimation du facteur de Hershfield : $c = 1.14$. van Montfort (1997) propose une méthode statistique pour estimer ce facteur. On présente ci-dessous une relation plus générale, fondée sur la théorie des valeurs extrêmes des séries stationnaires, proposée par Robinson et Tawn (2000).

Les deux séries de pluies de 24 h et de pluies journalières sont des séries de variables aléatoires non pas i.i.d., mais seulement i.d. On en déduit donc la présence de clusters, c'est-à-dire de regroupements dans le temps de valeurs fortes, comme le montre la figure 4.1. Heuristiquement, un cluster sur la série de pluies journalières formera également un cluster

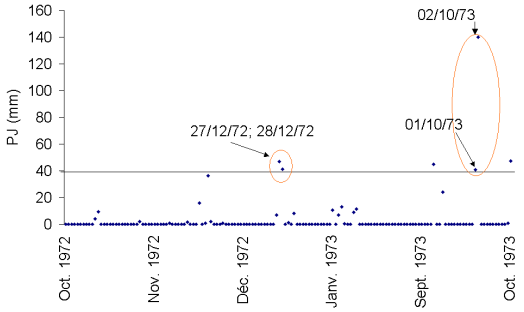


Fig. 4.1: Exemple de clusters sur la série journalière de Marseille.

sur la série de pluie de 24 h, et le cluster des pluies de 24 h aura une taille plus grande que

le cluster de pluies journalières. Cela est dû à la forte fréquence avec laquelle sont mesurées les pluies de 24 h. La taille des clusters reflète la dépendance des extrêmes, or la pluie de 24 h mesurée à l'heure h est fortement dépendante de la pluie de 24 h mesurée à l'heure $h+1$.

Dans la théorie des valeurs extrêmes des séries stationnaires, l'indice extrémal (noté IE, $IE \in [0,1]$) peut s'interpréter comme une mesure de la dépendance des extrêmes dans la série. On a vu au premier chapitre de la thèse que ce paramètre correspond à l'inverse de la taille moyenne des clusters.

On reprend les notations introduites précédemment, en les précisant.

– Le cumul de pluie maximum précipité en 24 heures est noté

$$H_{24} = \max\{Y_1(24), Y_2(24), \ldots, Y_m(24)\}$$

où m désigne le nombre de fenêtres mobiles de 24 heures mesurées dans une année. On note IE_{24} l'indice extrémal des pluies de 24 heures.

– Le cumul de pluie maximum précipité en un jour est noté

$$H_J = \max\{Y_{24}(24), Y_{48}(24), \ldots, Y_{24[m/24]}(24)\}$$

où $[m/24]$ est la partie entière de $m/24$ et désigne le nombre de jours de mesures dans l'année. On note IE_J l'indice extrémal des pluies journalières.

– On note F la distribution des $Y_i(24)$. On suppose qu'elle est identique pour tout i, et on suppose que les séries de pluies de 24 h et journalières sont stationnaires au sens strict. Cela peut être justifié si on considère des séries sur des saisons homogènes. D'autre part, la condition $\mathcal{D}(u_n)$ de non-dépendance à long terme, est une hypothèse physiquement raisonnable dans le cas des pluies.

Alors, d'après la théorie des valeurs extrêmes des séries stationnaires, on a, en reprenant les notations du théorème de Leadbetter (1983) de la section 1.4.1 :

$$P(H_{24} \leq x) \approx P(H_{24}^* \leq x)^{IE_{24}} = F(x)^{mIE_{24}} \tag{4.5}$$

$$P(H_J \leq x) \approx P(H_J^* \leq x)^{IE_J} = F(x)^{[m/24]IE_J}. \tag{4.6}$$

On a donc la relation

$$P(H_{24} \leq x) \approx P(H_J \leq x)^{24IE_{24}/IE_J}, \tag{4.7}$$

montrée par Robinson et Tawn (2000). Dans la suite, on note $\Theta = 24IE_{24}/IE_J$.

La relation 4.7 implique des relations entre les paramètres marginaux des lois GEV des deux distributions (Ancona-Navarrete et Tawn, 2000), (Coles, 2001) :

si $k_J = 0$:

$$\beta_{24} = \beta_J + \log(\Theta)\alpha_J, \qquad \alpha_{24} = \alpha_J, \qquad k_{24} = k_J, \tag{4.8}$$

si $k_J \neq 0$:

$$\beta_{24} = \beta_J + \alpha_J(1 - \Theta^{-k_j})/k_J, \qquad \alpha_{24} = \alpha_J\Theta^{-k_J}, \qquad k_{24} = k_J. \tag{4.9}$$

On a donc défini un nouveau modèle, de paramètres $\alpha_J, \beta_J, k_J, \Theta, \vartheta, \epsilon$. Les quantiles des pluies de différentes durées sont donnés par :

$$H_d(T) = (d/24)(\beta_J + \alpha_J/k_J(1 - \Theta^{-k_J}(-\log(1-1/T))^{k_J}))((24+\vartheta)/(d+\vartheta))^{\epsilon}. \tag{4.10}$$

Tous les paramètres des GEV marginales des variables H_d se déduisent de ces six paramètres. Par exemple, si $k_J \neq 0$, on a

$$\begin{aligned} \alpha_d &= \frac{d(24+\vartheta)^\epsilon}{24(d+\vartheta)^\epsilon}\alpha_J\Theta^{-k_J} \\ \beta_d &= \frac{d(24+\vartheta)^\epsilon}{24(d+\vartheta)^\epsilon}(\beta_J + \alpha_J(1-\Theta^{-k_J})/k_J) \\ k_d &= k_J \end{aligned} \tag{4.11}$$

Dans le modèle, le paramètre de forme k_d est constant sur les différentes durées, et égal au paramètre de forme des pluies journalières k_J. Cette conséquence du modèle empirique de Koutsoyiannis *et al.* (1998) est justifiée théoriquement par une étude de Nadarajah *et al.* (1998). Nadarajah *et al.* (1998) ont montré par une étude multivariée des extrêmes ordonnés, que si $H_d, H_{d'}$ suivent des distributions GEV, alors la relation

$$d \leq d' \Rightarrow H_d \leq H_{d'} \leq \frac{d'}{d}H_d, \tag{4.12}$$

impose certaines restrictions sur les marginales. En particulier, les paramètres de forme vérifient

$$k_d = k_{d'} \leq 0 \text{ ou } k_d > 0, k_{d'} > 0. \tag{4.13}$$

Dans le cas des pluies horaires, la relation 4.12 est vérifiée, et les quantités de pluies sont supposées non bornées supérieurement, donc leurs paramètres de formes sont négatifs. D'autre part, la relation théorique 4.7 entraîne $k_J = k_{24}$. Le paramètre de forme est donc constant sur toutes les durées considérées ici.

4.2.3 Choix de deux plages de durées

Puisque les cumuls de pluie sur des durées courtes ou longues sont issus de processus pluvieux différents, on sépare les durées en deux plages : les 'petites' durées et les 'grandes' durées. On suppose que les pluies journalières appartiennent aux pluies de grandes durées, le modèle défini par l'équation 4.10 est donc choisi pour décrire les pluies de grandes durées. Le modèle empirique donné par l'équation 4.2 est choisi pour décrire les petites durées. Soit d_f la durée frontière séparant les grandes des petites durées. Pour assurer une cohérence entre les quantiles de petites et grandes durées, les paramètres estimés sur les deux plages doivent satisfaire des propriétés de continuité en d_f. Le paramètre de forme reste constant sur les deux plages (égal à k_J), selon l'étude de Nadarajah *et al.* (1998).

Soit $f_d(x; \alpha_d, \beta_d, k_d)$ la densité de la loi GEV du maximum annuel (ou saisonnier) de la pluie mesurée en d heures. Les relations entre les paramètres sont les suivantes :

- pour les petites durées $d \leq d_f$, on note $\alpha_p, \beta_p, k_J, \vartheta_p, \epsilon_p$ les paramètres de l'équation 4.2, alors pour $d \leq d_f$:

$$\alpha_d = d\alpha_p/(d+\vartheta_p)^{\epsilon_p}, \beta_d = d\beta_p/(d+\vartheta_p)^{\epsilon_p}, k_d = k_J, \tag{4.14}$$

- pour les grandes durées $d \geq d_f$, on note $\alpha_J, \beta_J, k_J, \Theta, \vartheta, \epsilon$ les paramètres de l'équation 4.10.

Les hypothèses de continuité à la durée frontière d_f impliquent que les paramètres $\alpha_{d_f}, \beta_{d_f}$ soient égaux dans les deux équations 4.11 et 4.14. Cela implique :

$$\beta_p = \alpha_p \beta_{24} / \alpha_{24}, \qquad \epsilon_p = \frac{\ln[24\alpha_p(d_f + \vartheta)^\epsilon] - \ln[\alpha_{24}(24 + \vartheta)^\epsilon]}{\ln(d_f + \vartheta_p)}. \tag{4.15}$$

Avec deux plages de durées, huit paramètres $(\alpha_J, \beta_J, k_J, \Theta, \vartheta, \epsilon, \alpha_p, \vartheta_p)$ suffisent donc à calculer les paramètres α_d, β_d, k_d pour tout d appartenant aux plages de durées.

4.2.4 Estimation bayésienne des paramètres

Le modèle dont on dispose possède huit paramètres. Ces paramètres peuvent être estimés par maximum de vraisemblance, si un modèle de vraisemblance est donné. Des méthodes asymptotiques typiquement associées aux estimateurs de vraisemblance (matrice de Fisher, delta-méthode, profil de vraisemblance : voir chapitre 1) permettent d'estimer un intervalle de confiance des estimateurs des paramètres et des quantiles. Cependant, ces méthodes sont asymptotiques : dans le cas de notre modèle à huit paramètres, on ne connaît pas la vitesse de convergence des lois des estimateurs vers leurs lois asymptotiques.

D'autre part, les distributions étudiées ici vérifient un certain nombre de propriétés physiques, qu'il est intéressant de prendre en compte dans la procédure d'estimation. Des études régionales fournissent également des ordres de grandeurs des quantiles de pluie des différentes durées.

Pour ces raisons, il paraît intéressant d'utiliser une méthode bayésienne pour estimer les paramètres et leur distribution conjointe.

En pratique, nous avons utilisé un algorithme MCMC combinant un algorithme de Gibbs-Metropolis avec un algorithme de Metropolis. Cet algorithme a été proposé par Renard *et al.* (2006). L'intérêt de cette combinaison est de profiter de l'efficacité de l'algorithme de Gibbs-Metropolis, et la rapidité de l'algorithme de Metropolis. Les détails de ces algorithmes sont donnés dans l'article soumis (Muller *et al.*, 2006), en annexe.

Modèles de vraisemblances

Pour estimer les paramètres (par maximum de vraisemblance, ou par une méthode bayésienne), il est nécessaire de formuler une vraisemblance des observations, sachant les paramètres.

Nous proposons trois modèles de vraisemblance. Le problème essentiel est ici de modéliser la dépendance entre les durées.

- Vraisemblance V_1 : hypothèse d'indépendance entre les différentes durées

Nous posons d'abord l'hypothèse forte (souvent implicitement utilisée dans la construction des courbes Hauteur-Durée-Fréquence, via les régressions entre les paramètres et les durées) que les mesures sur les durées 1 h, 6 h, 12 h, 24 h, 48 h, 72 h et journalières (durées les plus couramment utilisées en hydrologie) sont indépendantes. La mesure de vraisemblance est donnée par :

$$V_1 = \prod_{d=1,6,12,24,48,72,J} \prod_{i=1}^{i_d} f_d(x_i^{(d)}; \alpha_d, \beta_d, k_d) \tag{4.16}$$

où α_d, β_d, k_d sont donnés par les équations 4.11, 4.14 et 4.15 ; $x_i^{(d)}, i = 1, \ldots, i_d$ sont les cumuls maximum annuels (ou saisonniers) mesurés sur les pas de temps d heures. L'avantage de cette vraisemblance est l'utilisation d'un nombre important d'informations. Néanmoins, la dépendance des observations de différentes durées peut se traduire par des redondances d'informations. En particulier, la variabilité des estimateurs des quantiles et des paramètres est sous-estimée.

- Vraisemblance V_2 : hypothèse d'indépendance entre quatre durées

L'hypothèse d'indépendance entre les sept durées précédentes est trop forte, on ne considère plus que quatre durées, dont la plus petite (1 h) et la plus grande (72 h). Les données journalières et de 24 h sont également importantes : d'une part les données journalières sont souvent abondamment renseignées, d'autre part elles ont un rôle important dans l'estimation des paramètres $\alpha_J, \beta_J, k_J, \Theta$.

Les maxima de pluies mesurées sur des petites durées sont dus à des phénomènes orageux ou convectifs, tandis que les maxima de pluies mesurées sur des grandes durées sont dus à des phénomènes dépressionnaires, ou frontaux. On considère donc les maxima mesurés sur des durées de 1 h indépendants de ceux mesurés sur des durées de 24 h, 72 h et journalières (Kieffer Weisse, 1998).

En général, les séries journalières sont plus longues que les séries horaires, et, pour éviter la redondance d'information et les problèmes de dépendance entre les maxima de pluies en 24 h et journalières, nous ne gardons que les pluies journalières des années (ou saisons) sans mesures horaires. La mesure de vraisemblance est donnée par :

$$V_2 = \prod_{d=1,24,72,J} \prod_{i=1}^{i'_d} f_d(x_i^{(d)}; \alpha_d, \beta_d, k_d) \tag{4.17}$$

où α_d, β_d, k_d sont donnés par les équations 4.11, 4.14 et 4.15 ; $i'_1 = i_1$, $i'_{24} = i_{24}$, $i'_{72} = i_{72}$ et i'_J est le nombre d'années de mesures journalières disjointes des années de mesures horaires.

L'inconvénient de cette dernière vraisemblance est de négliger la relation de dépendance entre les pluies de 24 h et de 72 h. Les maxima de pluie de ces deux durées sont en effet souvent dus au même phénomène pluvieux (dépression frontale). Une modélisation de la dépendance est proposée dans la troisième vraisemblance.

- Vraisemblance V_3 : quatre durées, avec modélisation de la dépendance entre les durées dépendantes.

Nous généralisons la seconde vraisemblance par une modélisation de la dépendance entre les pluies de 24 h et les pluies de 72 h, via une distribution bi-variée extrême (voir le chapitre 1 pour la théorie des lois bi-variées extrêmes). Les variables transformées $u_d(H_d) = -1/\ln(G_d(H_d))$, avec G_d la distribution GEV de la variable H_d, suivent une loi de Fréchet standard. La loi bi-variée choisie est la loi logistique, elle est couramment utilisée pour modéliser les extrêmes d'un processus bi-varié (Coles, 2001) :

$$G(x, y) = P(u_{24}(H_{24}) \leq x, u_{72}(H_{72}) \leq y) = \exp\{-(x^{-1/\Phi} + y^{-1/\Phi})^{\Phi}\}, x > 0, y > 0, \tag{4.18}$$

avec un paramètre de dépendance $\Phi \in (0, 1)$, alors on a :

$$P(H_{24} \leq x, H_{72} \leq y) = G(u_{24}(x), u_{72}(y)), x > 0, y > 0. \tag{4.19}$$

Lorsque $\Phi = 1$, les variables sont indépendantes, si $\Phi \to 0$, les variables sont parfaitement dépendantes. Pour évaluer la pertinance de ce modèle, les estimations non paramétriques de la fonction de dépendance de Pickands (Pickands, 1981), (Pickands, 1989), (Capéraà *et al.*, 1997) peuvent être comparées à l'estimation paramétrique, sous hypothèse d'un modèle logistique.

La vraisemblance est alors donnée par :

$$V_3 = \prod_{i=1}^{i_{24}} g(u_{24}(x_i^{(24)}), u_{72}(x_i^{(72)}); \alpha_{24}, \beta_{24}, k_{24}, \alpha_{72}, \beta_{72}, k_{72}, \Phi) u'_{24}(x_i^{(24)}) u'_{72}(x_i^{(72)}) \prod_{i=1}^{i'_J} f_J(x_i^{(J)}; \alpha_J, \beta_J, k_J) \prod_{i=1}^{i_1} f_1(x_i^{(1)}; \alpha_1, \beta_1, k_1) \tag{4.20}$$

où u'_d est la fonction dérivée de u_d, g est la densité de la loi bi-variée logistique. Si $k_d = 0$, $u_d(x) = \exp(-(x - \beta_d)/\alpha_d)$, et si $k_d \neq 0$, $u_d(x) = (1 - k_d(x - \beta_d)/\alpha_d)^{-1/k_d}$.

Formulation des lois a priori

Le tableau 4.1 présente les lois a priori choisies pour les paramètres du modèle. Les lois des paramètres de la GEV sont semblables à celles choisies par Coles et Pericchi (2003).

Paramètre	Loi a priori
α_J	lognormale de paramètres (0,100)
β_J	normale de moyenne 0, et variance 100
k_J	loi uniforme sur [-1,1]
Θ	loi uniforme sur [1,24]
ϑ	loi normale de moyenne 0, et variance 100, tronquée en 0
ϵ	loi lognormale de paramètres (0,100)
Φ	loi uniforme sur [0,1]

Tab. 4.1: Lois a priori des paramètres

ϵ, ϑ sont des paramètres positifs, c'est pourquoi nous leur avons choisi une loi lognormale et une loi normale tronquée en 0. On n'impose pas la condition $\vartheta < 1$ comme supposé par Koutsoyiannis *et al.* (1998). Cette généralisation que nous proposons est justifiée par le fait que $\vartheta < 1$ n'est pas imposé par des relations théoriques ou physiques. D'autre part, un algorithme MCMC a été implémenté avec une loi a priori uniforme sur [0,1] pour ϑ, mais les résultats n'ont pas été convaincants : l'algorithme n'a pas convergé.

Le paramètre Θ est relatif aux indices extrémaux des séries de pluies en 24 h et journalières, et donc à la taille des clusters de ces deux séries. Dans le cas des pluies de 24 h, les clusters sont de taille supérieure à ceux de la série journalière, du fait de la grande fréquence de mesure dans le cas des pluies de 24 h. En conséquence, $IE_{24} \leq IE_J$, donc $\Theta = 24 IE_{24}/IE_J \leq 24$. De plus, $\Theta \geq 1$. En effet, soit $N_{r_n}(u_n)$ le nombre de dépassements du seuil u_n parmi r_n mesures consécutives $(Y_1(24), Y_2(24), \ldots, Y_{r_n}(24))$ de la pluie en 24 h, avec $r_n/n \to 0$, si $n \to \infty$. On dit qu'il y a un cluster lorsque $N_{r_n}(u_n) > 0$, et l'ensemble des valeurs supérieures à u_n appartiennent au cluster. Les valeurs supérieures à u_n sont alors les valeurs du clusters. On définit, comme dans la section 1.4.1, la distribution de la taille des clusters par

$$\pi_{24,n}(j) = P(N_{r_n}(u_n) = j | N_{r_n}(u_n) > 0) \text{ pour } j = 1, \ldots, r_n. \tag{4.21}$$

La distribution limite de la taille des clusters est

$$\pi_{24}(j) = \lim_{n\to\infty} \pi_{24,n}(j), \text{ pour } j = 1, \ldots, \infty. \tag{4.22}$$

Robinson et Tawn (2000) ont montré que

$$IE_J \leq 24IE_{24}[1 - \sum_{i=1}^{23}(1 - i/24)\pi_{24}(i)], \tag{4.23}$$

donc, puisque

$$1 - \sum_{i=1}^{23}(1 - i/24)\pi_{24}(i) \geq 1 - \sum_{i=1}^{\infty}\pi_{24}(i) + \sum_{i=1}^{23} i/24\pi_{24}(i) = \sum_{i=1}^{23} i/24\pi_{24}(i) > 0, \tag{4.24}$$

cela implique que

$$\Theta = 24EI_{24}/EI_D \geq 1/\{1 - \sum_{i=1}^{23}(1 - i/24)\pi_{24}(i)\} \geq 1. \tag{4.25}$$

Ainsi $\Theta \in [1, 24]$, ce qui explique la loi a priori choisie pour Θ.

D'autre part, les paramètres du modèle doivent vérifier certaines conditions physiques, que l'on inclue dans la loi a priori. Si un paramètre ne vérifie pas ces conditions, la densité a priori en ce point est nulle.

- $d \leq d' \Rightarrow H_d \leq H_{d'} \leq \frac{d'}{d} H_d$
- $H_J < H_{24} < 2H_J$
- les quantiles de fréquence 0.9 (période de retour 10 ans) des pluies en 1 h et en 6 h doivent respectivement être compris entre [21 mm, 60 mm] et [51 mm, 95 mm], d'après une information régionale sur le département des Bouches du Rhône.

4.3 Cas d'étude de Marseille

4.3.1 Présentation des séries horaires et journalières

Nous disposons de deux séries à Marseille :

- une série pluviographique, avec 67 années de mesures horaires (1918-2002),
- une série pluviométrique, avec 122 années de mesures journalières (1882-2003).

La série pluviométrique appartient aux Séries Quotidiennes de Référence de Météo-France. Chacune des deux séries a moins de 10% de valeurs manquantes.

Pour ne pas être pénalisés par des problèmes de non homogénéité dus à la saisonnalité des processus, nous définissons deux saisons. La discrimination des saisons est réalisée à partir d'une méthode proposée par Kieffer Weisse (1998), sur la base des moyennes des maxima mensuels.

Les pluies les plus fortes en 1 h et journalières tombent entre septembre et janvier (voir fig. 4.2). Nous considérons donc que cette période est représentative des événements les plus forts. Tous les résultats de cette section ont été calculés sur cette période.

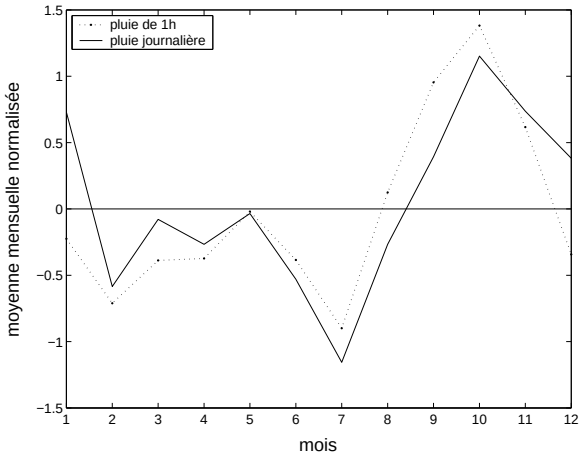


Fig. 4.2: Fluctuation saisonnière de la moyenne du maximum mensuel de pluie, normalisée et centrée, pour les pluies journalières et de 1 h.

La cohérence entre les deux séries a été vérifiée, en comparant leurs maxima annuels. Puisque la série journalière est de bonne qualité, on l'utilise comme série de référence. On compare les maxima annuels journaliers de la série horaire avec les maxima annuels journaliers de la série journalière (qui sert de référence). Si les mesures étaient exactes, les maxima annuels des deux séries devraient être égaux. Du fait des erreurs de mesures et des valeurs manquantes, en particulier lors des forts événements, les maxima des deux séries peuvent être différents. On veut donc rejeter les années de la série horaire pour lesquelles la différence entre les maxima journaliers est trop forte, et garder les années où la différence est tolérable. On doit donc fixer un seul de tolérance : si la différence entre les maxima est inférieure au seuil, l'année est conservée, et dans le cas contraire, elle est rejetée. Plus le seuil sera élevé, plus le nombre d'années conservées sera grand, mais la qualité des données horaires sera moins bonne. Le seuil cherché doit donc être assez élevé pour conserver un nombre suffisant d'années de la série horaire, et assez bas pour que les maxima annuels de la série horaire soient réellement des maxima annuels. En pratique, pour s'assurer de la qualité des données horaires observées, on compare également les quantiles de périodes de retour 2 et 10 ans des deux séries[1]. Avec des données parfaites, ceux-ci devraient être égaux. D'après la figure 4.3, on peut considérer que le seuil 19 mm est correct. Il correspond à une différence relative inférieure à 6% entre les quantiles journaliers de période de retour 2 et 10 ans, et à 45 années validées dans la série horaire. Dans la suite de cette étude, nous travaillons avec les données journalières et les 45 années de données horaires.

Une comparaison entre les valeurs extrêmes des deux séries (voir figure 4.4) montre que plusieurs valeurs fortes, présentes dans la longue série pluviométrique, n'ont pas été mesurées dans la série horaire. Par exemple, les événements extrêmes de 2000 (200 mm), 1973 (140 mm), 1932 (120 mm) sont des valeurs manquantes dans la série pluviographique.

[1]Les quantiles de la série journalière sont estimés à partir des maxima annuels de la série journalière, les quantiles de la série horaire sont estimés, pour chaque seuil de tolérance fixé, avec les maxima annuels de la série horaire, pour lesquels la différence avec les maxima annuels correspondant dans la série journalière est inférieure au seuil de tolérance.

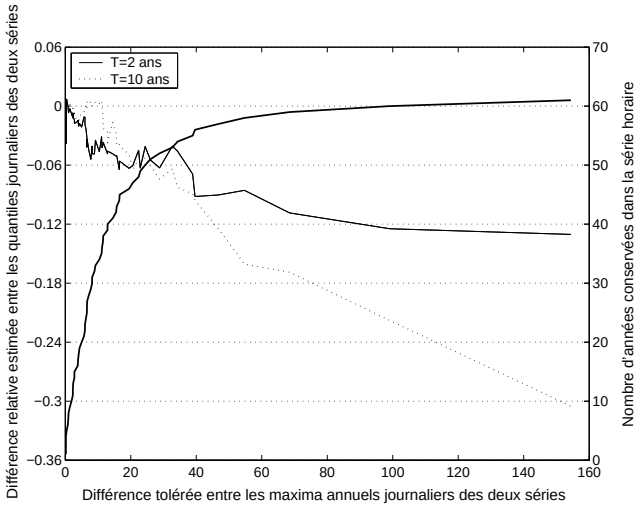


Fig. 4.3: Sélection du seuil en dessous duquel la différence entre les maxima annuels journaliers des deux séries horaires et journalières est acceptable. Axe des ordonnées de gauche : différence relative entre les quantiles journaliers des deux séries (courbes correspondantes : trait plein et pointillés pour $T = 2$ ans et 10 ans). Axe des ordonnées de droite : nombre d'années dans la série horaire, pour lesquelles la différence entre les maxima annuels journaliers de la série horaire et de la série journalière est inférieure au seuil de tolérance, pour un seuil de tolérance donné.

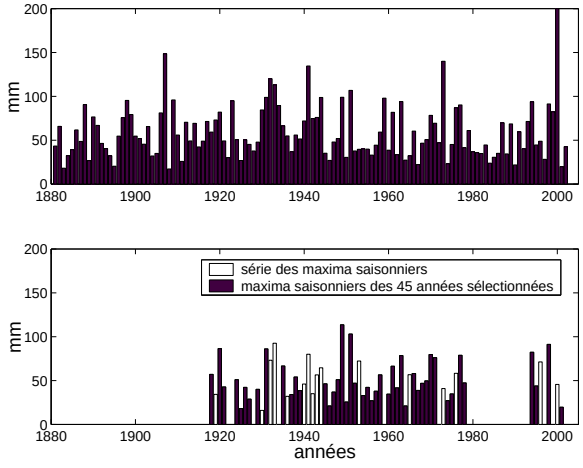


Fig. 4.4: Figure du dessus : pluie maximale journalière sur la saison des fortes pluies, à Marseille, pour la série journalière (122 années). Figure du bas : pluie maximale journalière sur la saison des fortes pluies, à Marseille, pour la série horaire (67 années). Les rectangles noirs représentent les 45 années sélectionnées de la série horaire, après vérification de la cohérence entre les séries.

La série journalière étant longue (122 années), on vérifie certains aspects de stationnarité au second ordre. Avec la même méthode qu'à la section 3.2, la série est stationnaire au sens où un modèle GEV à tendance linéaire dans les paramètres de position et d'échelle n'est pas significativement meilleur qu'un modèle GEV stationnaire. D'autre part, on vérifie la stationnarité de certaines variables : la moyenne annuelle de la quantité de pluie précipitée lors des jours pluvieux (c'est-à-dire avec plus d'1 mm de précipitation journalière), le maximum annuel, le taux annuel de jours sans pluie, le taux annuel de valeurs supérieures à un seuil élevé. Le test non paramétrique de Mann-Kendall (Mann, 1945), (Kendall, 1975) est utilisé pour détecter des tendances monotones dans des séries de données indépendantes. Nous avons d'abord vérifié l'hypothèse d'indépendance préalable pour appliquer le test de Mann-Kendall. La dépendance sérielle est analysée via le test non paramétrique de Wald-Wolfowitz (Wald et Wolfowitz, 1943). Aucune dépendance sérielle significative n'est détectée au niveau 25%. Aucune tendance significative n'est détectée par le test de Mann-Kendall, sur les quatre variables étudiées, pour un niveau 10%.

4.3.2 Analyse des dépendances entre les durées

Le coefficient de corrélation entre les maximum de pluies de 1 h et de 72 h, sur la saison des fortes pluies, est égal à 0.41. Cette valeur est assez forte, et l'hypothèse d'indépendance que l'on admet est seulement justifiée par des raisons physiques, puisque les processus pluvieux correspondants sont considérés différents (Kieffer Weisse, 1998). De même, le coefficient de corrélation est de 0.56 entre les pluies maximales de 1 h et de 24 h, et 0.90 entre les pluies maximales de 24 h et 72 h. Cette dernière corrélation est forte, et justifie l'usage d'une loi bi-variée dans la vraisemblance V_3 entre les pluies de 24 h et de 72 h.

Pour montrer l'intérêt de l'étude, on examine les estimations marginales des paramètres et des quantiles. On constate sur la figure 4.5 des quantiles centennaux de la pluie maximale sur la saison des fortes pluies, que le quantile de la pluie en 6 h est trop fort pour satisfaire la contrainte d'ordre avec le quantile de pluie en 12 h ($6 \leq 12 \Rightarrow H_6 \leq H_{12}$). Cela est dû à la valeur du paramètre $\widehat{k}_6 = -0.295$ (estimé marginalement sur la série des maxima en 6 h[2]), qui est inférieure à la valeur de $\widehat{k}_{12} = -0.166$ (estimé marginalement sur la série des maxima en 12 h). D'autre part, $\widehat{H}_{24}(100), \widehat{H}_{48}(100), \widehat{H}_{72}(100)$ ont des valeurs trop faibles pour satisfaire les contraintes d'ordre avec $\widehat{H}_J$: $H_J < H_{24} < 2H_J$ et $H_d \leq H_{d'} \leq \frac{d'}{d} H_d$ pour $d \leq d'$.

Cela s'explique par le fait que la série horaire ne contient que 45 années, tandis que la série journalière en contient 122, et les valeurs les plus fortes sont manquantes dans la série horaire.

Afin de vérifier la validité du modèle bi-varié proposé dans le modèle de vraisemblance V_3, nous avons ajusté la distribution logistique bi-variée extrême aux données de pluie maximales en 24 h et 72 h. La figure 4.6 montre que les estimations des fonctions de dépendance de Pickands paramétriques et non-paramétriques (avec les méthodes non paramétriques de Pickands (1981), Pickands (1989), Capéraà *et al.* (1997)) coïncident. L'adéquation graphique des fonctions de dépendance de Pickands montre que la distribution bi-variée logistique modélise bien la dépendance entre les pluies de 24 h et de 72 h. De plus, un test de rapport de vraisem-

[2]Nous verrons plus loin pourquoi cette valeur est si forte en valeur absolue.

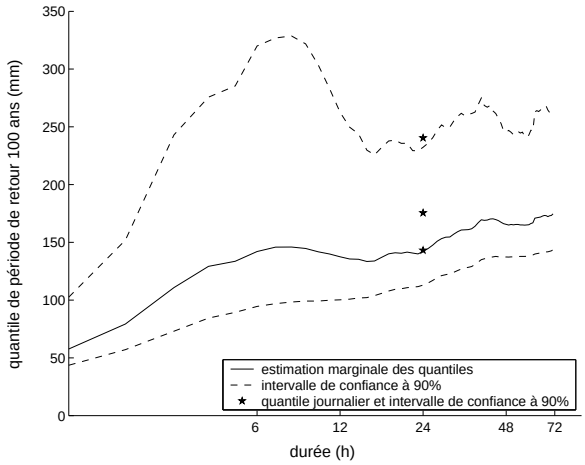


Fig. 4.5: Estimations marginales et intervalles de crédibilité à 90% (estimé par une méthode bayésienne) du quantile de période de retour 100 ans, pour les pluies de différentes durées.

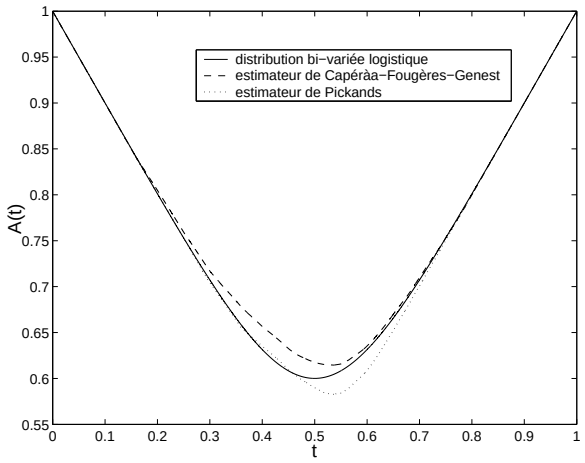


Fig. 4.6: Estimation de la fonction de dépendance de Pickands $A(t)$: comparaison entre des estimations non paramétriques de Pickands (1981), Pickands (1989), et Capéraà *et al.* (1997), et une estimation paramétrique (par maximum de vraisemblance) sous le modèle logistique.

blance a été appliqué entre des modèles emboîtés de loi bi-variée : le modèle d'indépendance, le modèle de distribution bi-variée extrême logistique, et le modèle de distribution bi-variée extrême logistique asymétrique, dont la fonction de distribution est, pour des marginales Fréchet standard :

$$G_a(x,y) = \exp(-(1-\psi_1)x^{-1} + (1-\psi_2)y^{-1} + \{(\psi_1 x^{-1})^{1/\Phi} + (\psi_2 y^{-1})^{1/\Phi}\}^{\Phi}) \qquad (4.26)$$

avec $\psi_1, \psi_2 \in [0,1]$. La distribution bi-variée extrême logistique asymétrique est une généralisation de la distribution extrême bi-variée logistique. Dans le cas de la loi logistique, les variables jouent un rôle symétrique, mais pas dans le cas de la loi logistique asymétrique. Aucune différence significative n'est détectée au niveau 10% entre les modèles logistiques symétrique et asymétrique (les estimations du maximum de vraisemblance donnent $\widehat{\Psi}_1 = 1, \widehat{\Psi}_2 = 1$), mais le modèle logistique est significativement meilleur que le modèle d'indépendance (la p-valeur est inférieure à 0.1%). Le paramètre de dépendance Φ de la distribution logistique est estimé à 0.24 par maximum de vraisemblance, affirmant une forte dépendance entre les pluies maximales en 24 h et en 72 h. L'effet de la modélisation par une loi bi-variée extrême logistique est de changer l'estimation du paramètre de forme. Marginalement, on estime $\widehat{k}_{24} = 0, \widehat{k}_{72} = 0.04$, mais avec la modélisation de la dépendance entre les maxima de pluie en 24 h et de 72 h et sans imposer la contrainte $k_{24} = k_{72}$, on obtient $\widehat{k}_{24} = -0.14, \widehat{k}_{72} = -0.12$, ce qui est proche de l'estimation marginale par maximum de vraisemblance de k_J : $\widehat{k}_J = -0.13$. Nous laissons ici une question en suspens, n'ayant pas trouvé de réponse immédiate : le fait de trouver des paramètres de forme proches de k_J est-il dû à une coïncidence ? ou bien à la structure de dépendance de la loi bi-variée logistique extrême, qui réussit à décrire marginalement les extrêmes à partir de données avec des lacunes importantes dans les extrêmes ?

4.3.3 Choix des plages de durées

Les durées sont séparées en deux plages. La durée frontière d_f entre les petites et grandes durées est considérée ici comme un paramètre supplémentaire, dans les modèles de vraisemblance V_1, V_2, V_3. La durée frontière d_f est alors estimée par maximum de vraisemblance, sous les contraintes $d \leq d' \Rightarrow H_d \leq H_{d'} \leq \frac{d'}{d} H_d$ et $H_J < H_{24} < 2H_J$. L'estimateur du maximum de vraisemblance est égal à 5.6, avec la vraisemblance V_1, tandis que les vraisemblances V_2, V_3 ne sont pas discriminantes pour d_f. Elles donnent en effet des vraisemblances égales pour $d_f = 5, 6$, ou 7 heures, tandis que les autres paramètres changent peu.

Afin de déterminer une durée de séparation entre les petites et grandes durées de pluie, on compare les estimations des paramètres α_d, β_d, k_d du modèle à deux plages, avec les estimations marginales des paramètres. La figure 4.7 montre que l'adéquation entre l'approche marginale et les approches multi-variées proposées est correcte pour $d_f = 7$ heures : les estimations des paramètres α_d, β_d restent à l'intérieur des intervalles de confiance à 95% des estimateurs marginaux de α_d, β_d. Les estimations avec d_f=5 heures sont proches des estimations marginales des paramètres α_d, mais en dehors des intervalles de confiance des estimateurs marginaux des paramètres β_d. Les estimations pour d_f=6 heures sont intermédiaires, mais extérieures aux intervalles de confiance des estimateurs marginaux de β_d.

Le paramètre k_d n'est pas utilisé pour le choix de d_f car k_d est constant dans le modèle ($k_d = k_J$) et d_f n'affecte pas le paramètre k_J des pluies journalières. La valeur choisie pour d_f est donc 7 heures.

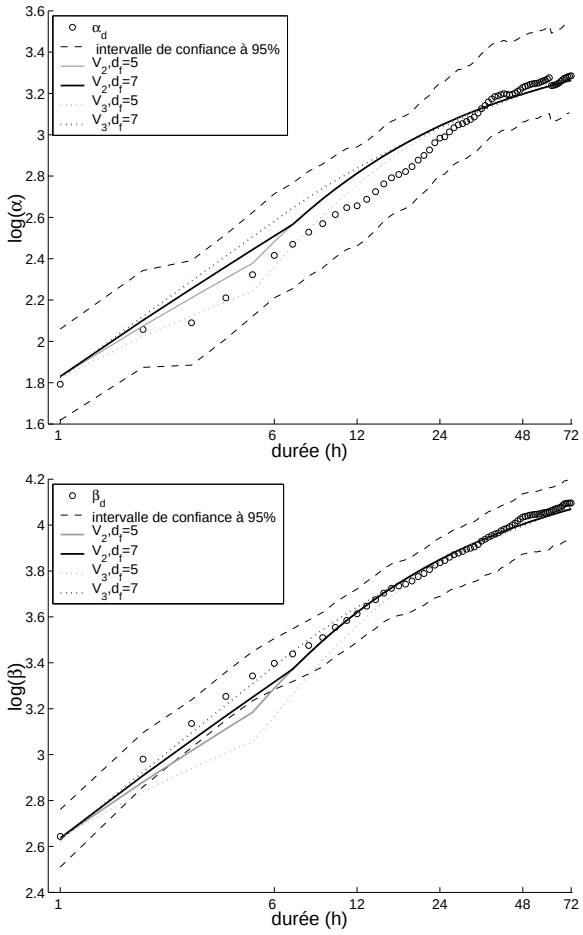


Fig. 4.7: Choix de la durée d_f de séparation entre les petites et grandes durées.

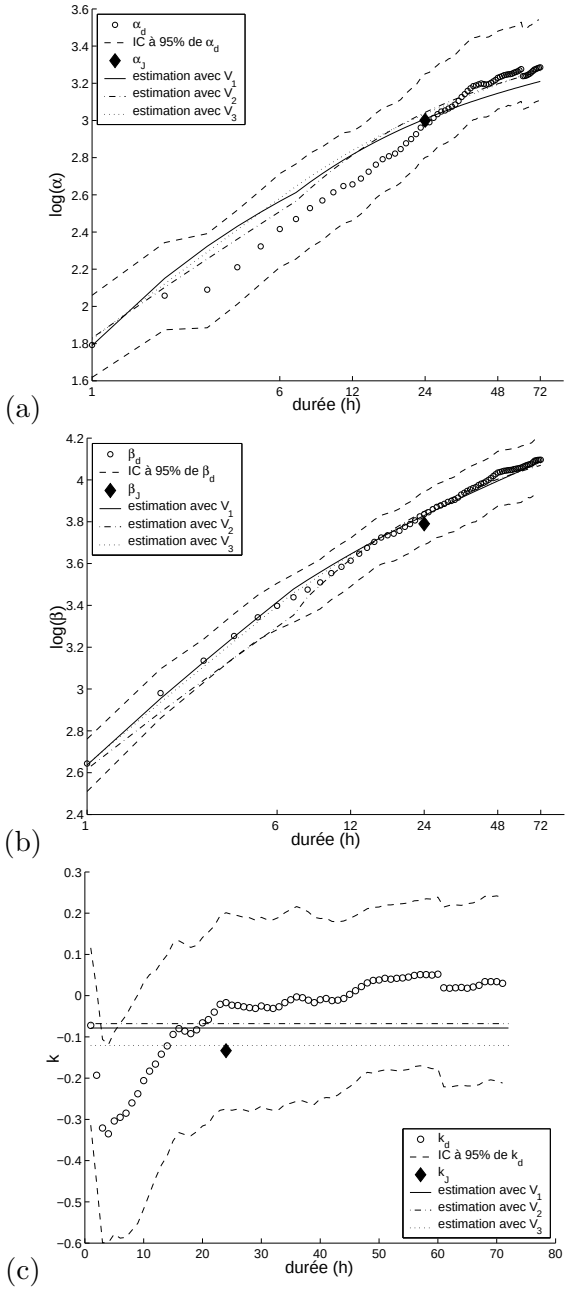


Fig. 4.8: Comparaison des estimations de (a) α_d, (b) β_d, (c) k_d par maximum de vraisemblances V_1, V_2, V_3, et sous les contraintes $d \leq d' \Rightarrow H_d \leq H_{d'} \leq \frac{d'}{d} H_d$ et $H_J < H_{24} < 2H_J$.

La figure 4.8 montre le bon ajustement des estimations par maximum de vraisemblance sous les contraintes $d \leq d' \Rightarrow H_d \leq H_{d'} \leq \frac{d'}{d} H_d$ et $H_J < H_{24} < 2H_J$, et pour les vraisemblances V_1, V_2, V_3. Les estimations de α_d, β_d avec les vraisemblances V_1, V_2, V_3 sont proches des estimations marginales des paramètres. Les estimations marginales de k_d sont approximativement constantes pour $d \geq 24$ h, et présentent un minimum pour les durées de 4 h. Cette variabilité de $\widehat{k}_d$, contraire à l'hypothèse théorique 4.13, est due à l'échantillonnage et à la sensibilité de l'estimation de k_d à l'échantillonnage. En effet, la pluie maximale en 6 heures observée sur la série horaire est de 103 mm, et est très proche de la pluie maximale en 15 heures observée sur la série horaire (104 mm).

4.3.4 Résultats, comparaison des trois modèles de vraisemblance

Les paramètres estimés sont présentés dans le tableau 4.2, par la médiane et un intervalle de crédibilité à 90% estimés sur 40000 paramètres simulés selon leur loi a posteriori, à l'aide d'un algorithme MCMC (voir les détails techniques dans l'article (Muller *et al.*, 2006) en annexe). La comparaison de la valeur médiane avec le milieu des intervalles de crédibilité à 90% des paramètres montre que la distribution a posteriori de α_J, β_J est symétrique, ce qui n'est pas vérifié pour les autres paramètres.

Le paramètre de dépendance Φ a une influence significative sur les estimations : sa médiane est estimée à 0.26, et son intervalle de crédibilité à 90% ne contient pas la valeur 1, qui correspond au cas de l'indépendance.

	V_1	V_2	V_3
α_J	20.05(18.45,21.83)	21.38(19.16,23.87)	21.23(18.89,23.89)
β_J	44.06(41.07,46.99)	44.82(40.92,48.65)	44.47(40.44,48.39)
k_J	-0.083(-0.158,-.014)	-0.076(-0.183,0.015)	-0.131(-0.236,-0.035)
Θ	1.17(1.02,1.42)	1.18(1.02,1.48)	1.19(1.02,1.54)
ϑ	4.09(-1.03,11.88)	5.61(-1.33,21.66)	9.22(-1.42,32.96)
ϵ	0.89(0.69,1.12)	0.96(0.74,1.25)	0.96(0.75,1.41)
α_p	10.15(4.41,16.89)	10.23(4.38,21.24)	9.36(4.27,18.51)
ϑ_p	0.94(-0.50,2.08)	0.88(-0.55,2.95)	0.68(-0.55,2.45)
Φ			0.265(0.212,0.335)

Tab. 4.2: Paramètres estimés et intervalles de crédibilité à 90% des paramètres.

Les estimations bayésiennes des médianes des paramètres α_d, β_d, k_d (non représentées graphiquement) sont semblables à celles obtenues par maximum de vraisemblance (présentées dans la figure 4.8). Dans les trois cas de vraisemblance, les estimations de k_J sont négatives, impliquant des quantiles non bornés supérieurement.

Les distributions a posteriori des quantiles $H_d(T)$ sont présentées dans les figures 4.9 et 4.10, pour les périodes de retour $T = 10, 100$ ans. La forme des distributions est généralement asymétrique, et le support des distributions devient de plus en plus large avec les modèles de vraisemblance V_1 à V_3. Ce dernier résultat est cohérent avec le fait que moins de données sont utilisées dans les vraisemblances V_2, V_3, et la vraisemblance V_3 inclue un paramètre supplémentaire Φ.

Puisque la série des pluies journalières contient un grand nombre de valeurs extrêmes (200 mm, 148 mm, 140 mm, 138 mm, etc.) et la série horaire ne contient pas ces valeurs

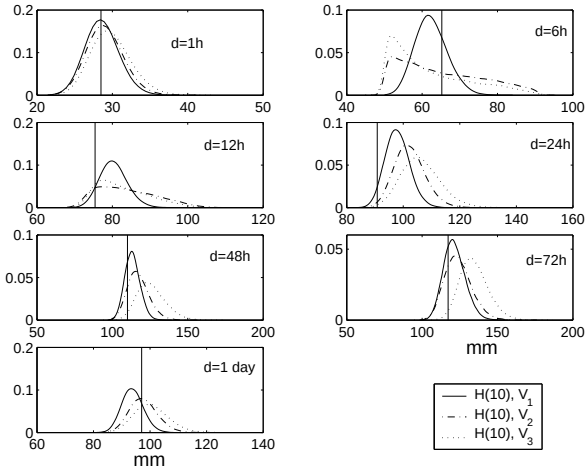


Fig. 4.9: Distributions a posteriori des quantiles $H_d(10)$ pour $d = 1$ h, 6 h, 12 h, 24 h, 48 h, 72 h.

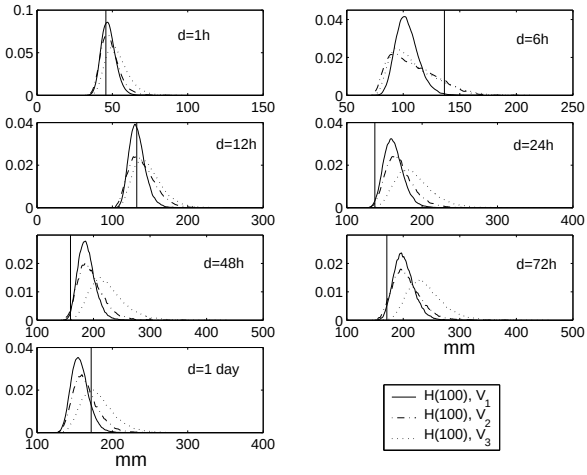


Fig. 4.10: Distributions a posteriori des quantiles $H_d(100)$ pour $d = 1$ h, 6 h, 12 h, 24 h, 48 h, 72 h.

extrêmes, les quantiles estimés des pluies de grandes durées sont significativement plus forts dans le cas de notre modèle que dans le cas marginal (traits verticaux dans les figures 4.9 et 4.10). Cette constatation est d'autant plus visible que le nombre de données horaires n'est pas trop important (cas des vraisemblances V_2, V_3) : dans le cas de V_1, le modèle est plus influencé par les données horaires que dans le cas V_2, V_3, or les données horaires ne contiennent que peu de valeurs extrêmes. L'augmentation des valeurs des quantiles de pluie horaires des grandes durées est due à la liaison entre ces pluies et les pluies journalières, en particulier via les relations 4.8, 4.9 et les relations d'ordre entre les quantiles de différentes durées. Pour la même raison, les quantiles journaliers estimés sont légèrement inférieurs à leurs estimations marginales. D'autre part, les pluies de courtes durées sont liées aux pluies journalières seulement par le paramètre k_J, par l'hypothèse de continuité des paramètres α_d, β_d à la durée frontière entre petites et grandes durées (voir les équations 4.14 et 4.15) et par les relations d'ordre entre quantiles définies dans la loi a priori. Les pluies journalières produisent donc moins d'effet sur les pluies de courtes durées. Les pluies de 6 h sont estimées moins fortes que dans leur estimation marginale, en raison de la liaison entre les pluies de 1 h et de 6 h. Les estimations marginales des pluies de 6 h sont fortes du fait de la présence dans la série des pluies de 6 h observées, du maximum de pluie précipitée en 15 h, non présent dans la série des pluies de 1 h.

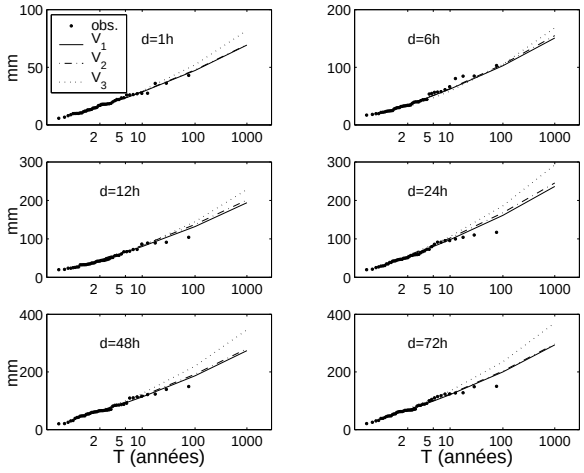


Fig. 4.11: Médiane des distributions a posteriori des quantiles $H_d(T)$ de cumuls maximum de pluies en d heures, d=1 h à 72 h (estimées par 40000 simulations par un algorithme MCMC)

Les figures 4.11 et 4.12 montrent les courbes Hauteur-Durée-Fréquence : les médianes des quantiles sont assez semblables entre les vraisemblances V_1, V_2 et sont plus fortes avec la vraisemblance V_3, du fait de la valeur de k_J dans ce cas (voir le tableau 4.2), et de la prise en compte de la dépendance entre les pluies de 24 h et de 72 h.

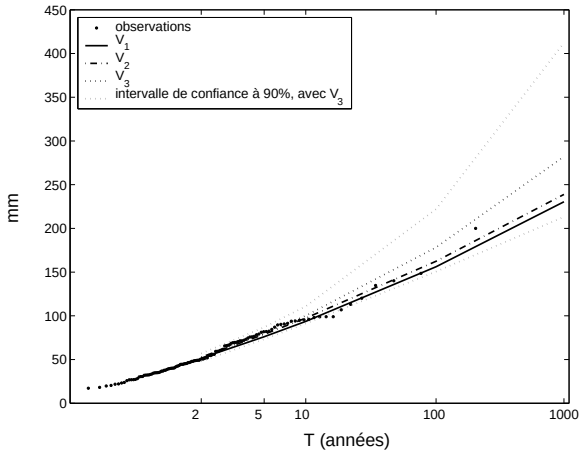


Fig. 4.12: Médiane des distributions a posteriori des quantiles H_J de cumuls maximum de pluies journalières (estimées par 40000 simulations par un algorithme MCMC)

4.4 Conclusions

L'étude présentée ici a montré plusieurs points.

- A partir d'une série horaire, avec des lacunes importantes au niveau des valeurs extrêmes, et à partir d'une série journalière de bonne qualité, les modèles proposés ont permis d'estimer des quantiles de différentes durées entre 1 heure et 72 heures.
- La dépendance entre les durées a été modélisée par trois modèles différents. Les modèles proposés possèdent huit ou neuf paramètres et sont fondés sur des distributions marginales GEV. Parmi ces modèles, plusieurs hypothèses ont été faites pour modéliser les dépendances entre durées.
  1. Les paramètres des lois marginales s'expriment en fonction de la durée et des huit ou neuf paramètres du modèle. D'autre part, les paramètres de forme des lois GEV marginales sont constants entre les différentes durées, d'après une étude théorique, vérifiée sur les données.
  2. Une étude de la relation entre la distribution des pluies de 24 heures et la distribution des pluies journalières a été le fondement de la modélisation simultanée de pluies de durées horaires et de pluies journalières. Cette modélisation est empruntée à la théorie des processus stationnaires de valeurs extrêmes et à l'indice extrémal. En ingénierie, l'intérêt d'inclure l'information journalière est important, puisque les données journalières sont en général mieux renseignées que les données horaires : les séries journalières sont plus longues, et dans le cas de Marseille, de meilleure qualité, en particulier au niveau des extrêmes.
  3. L'hypothèse d'indépendance entre pluies de petites durées, souvent dues à des orages, et pluies de grandes durées, a également été utilisée.
  4. La dépendance entre pluies de durées 24 heures et 72 heures a été modélisée par une loi bi-variée extrême : la loi logistique.
- Les paramètres ont été estimés dans un cadre bayésien. Le cadre bayésien a permis d'inclure dans les lois a priori un certain nombre d'informations sur les quantiles (relations d'ordre entre durées, informations régionales).

– Les quantiles estimés par les trois modèles de dépendance sont plus forts que dans le cas de leurs estimations marginales. Les estimations marginales sont en effet sous-estimées du fait du manque de données extrêmes dans la série horaire. Ce résultat montre que dans le cas de Marseille, l'effet de la prise en compte des données journalières dans un modèle Hauteur-Durée-Fréquence est non négligeable, puisque les estimations des quantiles des pluies de grandes durées sont en effet plus fortes que leurs estimations marginales. D'autre part, l'estimation plus forte des quantiles avec les modèles de dépendance, reste raisonnable, car les paramètres des lois GEV, estimés avec les modèles de dépendance, sont cohérents avec les intervalles de confiance des paramètres estimés marginalement sur les différentes séries.
– Les résultats ont également montré que la modélisation de la dépendance des pluies de 24 heures et de 72 heures via la loi bi-variée logistique est significativement meilleure que l'hypothèse d'indépendance. Cela a été montré de diverses manières : par un test de rapport de vraisemblance et par l'intervalle de crédibilité du paramètre de dépendance Φ. Les résultats pourraient certes être améliorés avec des lois multivariées de dimension supérieure à 2. Les lois multivariées les plus simples à manipuler sont les lois gaussiennes, mais elles sont fondées sur une hypothèse forte de vecteurs gaussiens. De plus, les modèles gaussiens impliquent l'indépendance asymptotique des co-variables, ce qui n'est peut-être pas le cas des pluies de différentes durées. La famille des lois multivariées des valeurs extrêmes impliquent en revanche la dépendance asymptotique des co-variables. Cette famille ne possède pas de paramétrisation finie, et les modèles paramétriques en dimension supérieure à trois deviennent rapidement compliqués (Barakat *et al.*, 2004). Le problème du choix d'une loi spécifique pourrait être contourné par une estimation non paramétrique de la distribution multi-variée (Beirlant *et al.*, 2004). Ces modèles plus complexes sont des perspectives pour de futures recherches.
– Enfin, dans le débat entre la loi Gumbel et la loi GEV de paramètre de forme $k < 0$, deux des intervalles de crédibilité de k sont entièrement inclus dans $\mathbb{R}^-$, tandis que le troisième intervalle de crédibilité a une borne supérieure légèrement supérieure à 0 : 0.015. Ces résultats confirment ceux de Bacro et Chaouche (2006) qui ont montré que les pluies journalières pré-hivernales (entre le 1er août et le 8 novembre) de Marseille n'appartiennent pas au domaine d'attraction de la loi Gumbel mais à celui de Fréchet : les maxima journaliers de la saison pré-hivernale sont modélisés par une loi GEV de paramètre de forme $k < 0$.

Troisième partie

Simulateurs stochastiques de pluie

Chapitre 5

Bibliographie des générateurs de pluie

Pour analyser les événements pluvieux, nous avons besoin de mesures d'intensités moyennes ou de cumuls, et ce sur différents pas de temps (de quelques minutes à environ trois jours). Pour analyser plus particulièrement les événements extrêmes, nous avons besoin de longues séries de mesures à ces différents pas de temps. En effet, d'un point de vue statistique, nous avons montré au chapitre 2 la forte incertitude des quantiles extrêmes estimés avec une série courte. D'autre part, les données de pluie peuvent servir de données d'entrée dans d'autres types de modèles : des modèles de pluie-débit, des modèles d'agronomie, des modèles climatiques, etc. Face à ces besoins, les problèmes fréquemment rencontrés sont l'insuffisance de données : faible nombre d'années de mesures, nombre important de données manquantes, ou insuffisance de la couverture spatiale par les stations de mesures. Des modèles générateurs de pluie ont donc été développés, afin de simuler des chroniques de pluie représentatives de l'ensemble des événements susceptibles de se produire. Du fait de la nature complexe du climat, la pluie est souvent modélisée comme un processus stochastique. Nous présentons dans ce chapitre les générateurs stochastiques temporels de pluie. Nous commentons plus particulièrement les modélisations des valeurs extrêmes via ces générateurs.

De nombreux modèles ont été développés, pour répondre à différents objectifs (par exemple la description des sécheresses extrêmes, et/ou la description des pluies extrêmes) tout en préservant les caractéristiques de la distribution centrale des pluies (saisonnalité, moyennes et variances annuelle, mensuelle, journalière des pluies). Nous nous intéressons ici à des modèles stochastiques capables de reproduire les valeurs extrêmes, pour différentes durées, tout en respectant les caractéristiques de la distribution centrale des observations. Nous présentons les grandes familles de modèles, en précisant leurs avantages et leurs inconvénients.

Certaines caractéristiques sont essentielles et doivent être prises en compte par les modèles temporels :

1. la présence de saisons,
2. la présence potentielle de tendances dues à un éventuel changement climatique de la pluviométrie (dans ce cas, la notion de valeur extrême varie au cours du temps),
3. la dépendance temporelle de la pluie : les intensités ou les cumuls successifs ne sont pas indépendants les uns des autres,

4. la variabilité inter-annuelle, causée par des fluctuations basse-fréquence du climat. Les fluctuations sont essentiellement dues à des interactions océan-atmosphère (Katz et Parlange, 1998). Par exemple, le phénomène de l'Oscillation du Sud El-Niño (ENSO) se caractérise par le réchauffement de la surface de l'océan qui s'étend du Pacifique central jusqu'aux côtes du Pérou et de l'Équateur. Ce réchauffement de la surface de la mer, de l'ordre de 4^oC à 6^oC, s'accompagne d'une interaction océan-atmosphère qui perturbe la circulation atmosphérique. Toute la ceinture tropicale du globe subit un bouleversement climatique qui provoque régionalement des précipitations intenses ou des périodes de sécheresse exceptionnelles. D'autres phénomènes de fluctuations climatiques basse-fréquence existent, comme l'Oscillation Nord Atlantique (NAO), l'un des plus vieux phénomènes météorologique connus. Il régit les hivers d'Europe du Nord, qui sont alors soit humides et chauds soit froids et secs.

Les méthodes les plus souvent utilisées pour modéliser ces différentes caractéristiques sont décrites ci-dessous.

1. Les effets saisonniers peuvent être modélisés par une paramétrisation spécifique pour chaque saison ou chaque mois. Les paramètres sont parfois définis par une fonction périodique du temps, nécessitant des hyper-paramètres. La fonction périodique peut être donnée par une série de Fourier :

$$\theta(t) = \theta_0 + \sum_{k=1}^{n_\theta} \theta_{2k-1} \sin(2k\pi t/365) + \theta_{2k} \cos(2k\pi t/365), \tag{5.1}$$

   avec n_θ le nombre de paires d'harmoniques utilisées dans la somme de Fourier (Coles, 1994).

   L'intérêt d'une paramétrisation continue est d'éviter le découpage arbitraire des saisons. Récemment, des approches non paramétriques ont été présentées par Rajagopalan *et al.* (1996), dans le cas des séries temporelles au pas de temps journalier. La saisonnalité est modélisée par l'intermédiaire de fenêtres mobiles des observations, centrées sur la date du jour simulé.
2. Pour un modèle paramétrique, la présence de tendance est modélisée via une fonction temporelle des paramètres $\theta(t)$. La dépendance temporelle des paramètres peut être linéaire, quadratique, ou représentée par un changement discontinu des paramètres à un instant donné (Coles, 2001).
3. L'auto-corrélation des processus peut être modélisée par des processus auto-corrélés (chaînes de Markov, processus autorégressifs dans le cas des débits, etc.). Quelques exemples sont donnés dans ce chapitre, et un modèle markovien est proposé dans la partie IV de la thèse.
4. Plusieurs méthodes ont été proposées pour tenir compte de la variabilité inter-annuelle des pluies, pour les régions concernées. Une méthode consiste à conditionner les paramètres par une co-variable (Wilks, 1999). Par exemple, Thyer et Kuczera (2000) et Thyer et Kuczera (2003) utilisent des chaînes de Markov cachées, pour modéliser les années sèches et les années humides. Wang et Nathan (2002) et Dubrovsky *et al.* (2004) proposent de forcer les simulations journalières par des données mensuelles ou annuelles. Dubrovsky *et al.* (2004) montrent qu'après avoir inséré les simulations de pluies dans un modèle pluie-débit, la variabilité mensuelle des crues, et la queue de distribution des crues sont mieux restituées avec un conditionnement aux données mensuelles que sans conditionnement.

Avant de présenter les différents générateurs de pluie dans les sections suivantes, la section 5.1 présente des critères statistiques BIC et AIC utilisés par des hydrologues pour comparer les modèles.

Les deux sections suivantes présentent les générateurs de pluie. La pluie est souvent mesurée au pas de temps journalier. À partir de ce pas de temps, des séries de pluie au pas de temps mensuel ou annuel peuvent être créées (par agrégation des pas de temps journaliers). En France, les forts événements ont une durée de quelque minutes à quelques jours, c'est pourquoi nous nous intéressons ici seulement aux générateurs stochastiques de pluie à des pas de temps journalier (section 5.2) ou inférieurs à 24 h (section 5.3).

5.1 Critères d'évaluation des modèles

On présente sommairement les critères d'Akaike (AIC) proposé par Akaike (1974) et Bayésien (BIC) proposé par Schwarz (1978). Le but de ces critères est de chercher un compromis entre une paramétrisation suffisante pour bien ajuster un modèle aux observations, et une paramétrisation la moins complexe possible. Un tel compromis permet de respecter le principe de parcimonie des modèles. Le maximum de vraisemblance ne peut donc pas être un tel critère, puisqu'il conduirait à choisir le modèle le plus complexe. Les deux critères s'expriment par

$$-2l(\widehat{\theta}) + 2\xi(\widehat{\theta}) \tag{5.2}$$

où $l(\widehat{\theta})$ est la fonction de log-vraisemblance calculée en l'estimateur du maximum de vraisemblance du paramètre : $\widehat{\theta}$, et $\xi(\widehat{\theta})$ pénalise la complexité du modèle.

5.1.1 Critère d'Akaike (AIC)

Le critère d'Akaike est fondé sur une pseudo-distance entre une 'vraie' distribution g, inconnue, et une distribution arbitraire f, paramétrée par θ de dimension K. Cette pseudo-distance, appelée nombre d'information de Kullback-Leibler, est définie par

$$I(g:f) = E_g[\log \frac{g(X)}{f(X)}] = \int_{\mathbb{R}} \log \frac{g(x)}{f(x)} g(x)dx. \tag{5.3}$$

$-I(g:f)$ est également appelée entropie. On cherche donc un modèle qui minimise la pseudo-distance ou encore maximise l'entropie.

Or $I(g:f) = E_g(\log g(X)) - E_g(\log f(X))$. Le problème revient maintenant à maximiser $E_g(\log f(X))$. Akaike (1974) a montré que l'expression suivante :

$$-\frac{1}{N}\sum_{i=1}^{N} \log f(X_i|\widehat{\theta}) + \frac{K}{N}, \tag{5.4}$$

avec N la taille de l'échantillon sur lequel est estimé $\widehat{\theta}$, est un estimateur asymptotiquement sans biais de $E_g(\log f(X|\widehat{\theta}))$.

En posant

$$AIC(f) = -2l(\widehat{\theta}) + 2K, \tag{5.5}$$

avec $l(\widehat{\theta}) = \sum_{i=1}^{N} \log f(X_i|\widehat{\theta})$, minimiser $I(g:f)$ revient donc à sélectionner f_{AIC} minimisant AIC.

On peut également montrer, par définition du critère AIC, que sélectionner un modèle via le critère AIC revient à chercher le modèle faisant le meilleur compromis biais-variance pour le nombre de données N dont on dispose (Lebarbier et Mary-Huard, 2004). Le meilleur modèle au sens de AIC dépend donc de N.

Des modifications de ce critère ont été proposées, en particulier lorsque le nombre de paramètres K est grand par rapport à N (ce qui inclue le cas des petits échantillons), voir (Burnham et Anderson, 2004) pour des références.

En pratique Cahill (2003) montre que les différences entre les valeurs des AIC sont très petites relativement aux valeurs des AIC. Cahill (2003) suggère alors d'utiliser le test du rapport de vraisemblance, lorsque les modèles appartiennent à la même famille. Si les modèles comparés ne sont pas issus de la même famille de modèles (par exemple la loi gamma et le mélange d'exponentielles), on peut utiliser une famille plus générale de modèles (par exemple les mélanges de loi gamma et d'exponentielles).

5.1.2 Critère Bayésien (BIC)

Cette fois, on se place dans un cadre bayésien : les paramètres θ des modèles M, et les modèles eux-mêmes sont aléatoires, munis d'une loi a priori $P(M)$ pour les modèles et $P(\theta|M)$ pour les paramètres des modèles. On suppose que la loi a priori des modèles est non-informative.

On recherche ici le modèle M_{BIC} qui maximise la distribution a posteriori des modèles $M_{BIC} = argmax_M P(M|X)$, c'est-à-dire le modèle le plus vraisemblable au vu des données.

Une approximation de Laplace permet (voir (Raftery, 1994), (Lebarbier et Mary-Huard, 2004)) de donner l'approximation suivante :

$$\log P(M|X) \approx l(\widehat{\theta}) - \frac{K}{2}\log(N), \tag{5.6}$$

où $\widehat{\theta}, K, N$ sont l'estimateur du maximum de vraisemblance, la dimension du paramètre θ, et la taille de l'échantillon.

Le critère BIC est défini par :

$$BIC_M = -2l(\widehat{\theta}) + K\log(N). \tag{5.7}$$

La maximisation de la distribution a posteriori des modèles revient à minimiser la valeur de BIC. Si la loi a priori des modèles est informative, on utilise le critère modifié

$$BIC_M = -2l(\widehat{\theta}) + K\log(N) - 2\log P(M). \tag{5.8}$$

On remarque que les lois a priori des paramètres des modèles n'interviennent pas dans le critère BIC. Cette remarque n'est valide qu'asymptotiquement : l'apport des informations a priori sur les paramètres est négligeable devant l'information apportée par l'échantillon.

Si on dispose d'une suite de modèles emboîtés $M_1 \subset \ldots \subset M_m$, la suite des pseudo-distances de Kullback-Leibler décroît, et il existe un modèle M_t pour lequel pour laquelle le

critère AIC ne décroît plus (ce modèle existe toujours, dans le pire des cas, $M_t = M_m$). On appelle M_t 'quasi-vrai' modèle. Le critère BIC possède la propriété d'être consistant pour le quasi-vrai modèle : asymptotiquement le critère BIC donne une valeur $+\infty$ aux modèles M_i avec $i \neq t$ (Lebarbier et Mary-Huard, 2004). Cette propriété n'est pas partagée par le critère AIC (Lebarbier et Mary-Huard, 2004).

Dans la pratique, il a été observé que le critère BIC sélectionne des modèles de dimension plus petite que le critère AIC, ce qui n'est pas étonnant puisque le critère BIC pénalise plus que le critère AIC (dès que $N > 7$). Burnham et Anderson (2004) et Lebarbier et Mary-Huard (2004) donnent quelques comparaisons des deux critères AIC et BIC. Ils concluent qu'aucun critère n'est universellement meilleur que l'autre, et donnent des indications sur le choix du critère préférable à utiliser suivant les situations (complexité des modèles, taille de l'échantillon). Des références sur l'utilisation et les limites de AIC et BIC, en hydrologie, sont données dans (Srikanthan et McMahon, 2001), et (Cahill, 2003).

5.1.3 Autres méthodes

Pour des modèles emboîtés, la méthode classique du test du rapport de vraisemblance peut être utilisée.

Dans le cadre bayésien, on peut simplement calculer les probabilités a posteriori des différents modèles, et sélectionner le modèle possédant la plus grande probabilité a posteriori. On peut également calculer les distributions prédictives de certaines variables, et les comparer avec les observations (dans le cas de l'étude des valeurs extrêmes, on peut considérer les quantiles, ou les courbes Intensité-Durée-Fréquence). Des exemples sont donnés dans (Chaouche et Parent, 1999), (Cameron *et al.*, 2000a).

Dans le cadre non paramétrique, les méthodes citées ci-dessus ne sont pas applicables. Une méthode, également valable dans le cas paramétrique, est alors de simuler plusieurs sorties du modèle, et d'en déduire des intervalles de confiance de variables comparables avec les observations.

5.2 Générateurs de pluie au pas de temps journalier

Les modèles les plus répandus en simulation journalière des pluies combinent un processus d'occurrence pour modéliser les arrivées d'évenements pluvieux, avec un modèle de cumuls de pluie pour modéliser la quantité précipitée dans l'événement. Ce type de modèle est utilisé aussi bien en version paramétrique que non paramétrique.

D'après des études de Wilks (1999) et Cahill (2003) ces générateurs de pluie sous-estiment souvent la fréquence des événements extrêmes (en particulier des sécheresses et des fortes pluies), et la variabilité inter-annuelle des précipitations (indiquée par la variabilité du cumul annuel, saisonnier ou mensuel au cours des années) pour les régions soumises à des variations inter-annuelles. Ce fait a été maintes fois constaté dans la littérature hydrologique (voir (Wilks, 1999) pour des références).

Nous présentons de manière générale ces modèles aux sections 5.2.1 et 5.2.2, puis nous en donnerons des exemples et commentaires dans la section 5.2.3, issus de la littérature

hydrologique. Enfin, nous présentons des versions non-paramétriques de ces modèles dans la section 5.2.4.

5.2.1 Processus des occurrences d'événements pluvieux

Les processus d'occurrences les plus couramment employés dans la simulation des pluies journalières sont les

- modèles markoviens : l'état pluvieux ou sec de chaque jour j dépend de l'état des jours précédents,
- modèles d'alternance entre durées sèches et pluvieuses : on simule bout à bout des durées, alternativement sèches et pluvieuses,
- processus poissonnien d'arrivées : cette modélisation est utilisée par Sharma (1996) pour modéliser des occurrences de périodes pluvieuses dans les zones semi-humides et semi-arides du Kenya, en Afrique. Néanmoins, l'utilisation de processus ponctuels reste marginale en simulation journalière des pluies, mais est plus fréquente en modélisation à temps continu. Les modèles à processus ponctuels seront donc présentés dans la section 5.3.

Dans certains modèles markoviens, il est possible de calculer théoriquement les lois des durées sèches et pluvieuses. Le modèle d'alternance ainsi défini est donc une autre formulation du modèle markovien.

Modèles markoviens

La modélisation markovienne du processus d'occurrences des pluies consiste à simuler les états (sec ou pluvieux) des jours successifs. Le plus souvent, l'ordre des chaînes de Markov est fixé à 1. Srikanthan et McMahon (2001) donnent un grand nombre de références. Cependant, il a été montré plusieurs fois que les chaînes de Markov d'ordre 1 sous-estiment souvent les durées sèches extrêmes (Buishand, 1978), (Guttorp, 1995), (Racsko *et al.*, 1991) (Wilks, 1999).

Plus récemment, les ordres 2 et 3 ont également été utilisés (Wilks, 1999), (Koutsoyiannis, 2006). Dans le cas de modèles non paramétriques, pour prendre en compte un grand nombre d'intervalles de temps passés, l'occurrence de pluie est modélisée conditionnellement au cumul des jours pluvieux précédents (Sharma et O'Neill, 2002), ou au nombre de jours pluvieux précédents (Harrold *et al.*, 2003a). Nous reviendrons sur les modèles non-paramétriques à la section 5.2.4.

Les chaînes de Markov à l'ordre k ont été généralisées par Stern et Coe (1984) par des chaînes de Markov d'ordre hybride : la mémoire de la chaîne markovienne s'étend du jour précédent au jour pluvieux le plus récent, si ce temps est inférieur à k. Ainsi, à l'intérieur des événements pluvieux, les jours sont liés par un modèle markovien d'ordre 1, et dans les périodes sèches, la chaîne de Markov peut être d'ordre supérieur. L'auto-corrélation des événements secs est donc mieux respectée, permettant ainsi de mieux reproduire les sécheresses. L'intérêt de cette méthode, par rapport à une chaîne de Markov d'ordre k est de réduire le nombre de paramètres : le modèle hybride nécessite $k+1$ paramètres, tandis que le modèle markovien nécessite 2^k paramètres.

Le nombre d'états de ces modèles est généralement de deux : sec et pluvieux. Des modèles à plus de deux états ont été également proposés. Haan *et al.* (1976), Srikanthan et McMahon

(1985) et Wang et Nathan (2002) proposent des modèles à sept états dont les états dépendent de la quantité d'eau précipitée dans la journée : cette classification permet de distinguer les jours avec des valeurs extrêmes. Dans ces modèles, la dernière classe correspond aux plus fortes pluies, et les cumuls sont modélisés par une anamorphose vers une loi normale (Srikanthan et McMahon, 1985), ou par une loi gamma d'après une modification de Wang et Nathan (2002). Le respect de la variété des types de pluies a conduit ces modèles à bien reproduire la variance et le coefficient d'asymétrie des pluies journalières, suggérant que les extrêmes peuvent être bien reproduits. De même Gregory *et al.* (1993) proposent un modèle à dix états, et la loi du cumul de l'état correspondant aux plus fortes pluies est la loi gamma. Cette modélisation a permis de mieux restituer les variations saisonnières qu'un modèle à deux états.

Le critère d'information d'Akaike (AIC) introduit par Akaike (1974), et le critère d'information Bayésien (BIC) Schwarz (1978) ont été beaucoup utilisés pour choisir l'ordre optimal de la chaîne de Markov. Ces critères sont présentés dans la section 5.1.

Processus d'alternance de temps sec et pluvieux

Dans la modélisation par un processus d'alternance, on simule alternativement des durées sèches et pluvieuses, supposées indépendantes. Les durées sont souvent modélisées par les lois exponentielle ou Weibull. Des références pour ces modèles sont données dans (Chapman, 1998), (Srikanthan et McMahon, 2001) et (Salvadori et De Michele, 2006).

Les modèles d'alternance sont parfois simplement une autre formulation des modèles markoviens. Par exemple, pour les modèles markoviens au premier ordre, la distribution des durées est une loi géométrique. En effet, on note p_{01} la probabilité de saut d'un jour sec à un jour pluvieux, et p_{11} la probabilité qu'un jour pluvieux reste pluvieux le lendemain, S la durée d'une période sèche, et H la durée d'une période humide,

$$\begin{aligned} &P(S = x|\text{état initial sec}) = p_{01}(1-p_{01})^{x-1}, x = 1, 2, \dots \\ &P(H = x|\text{état initial humide}) = (1-p_{11})p_{11}^{x-1}, x = 1, 2, \dots \end{aligned} \tag{5.9}$$

En fait, l'équation 5.9 est vraie seulement si p_{01}, p_{11} sont constants. Or p_{01}, p_{11} varient avec les variations climatiques annuelles : l'équation 5.9 est donc seulement une approximation.

Racsko *et al.* (1991) ont montré que pour des données de Hongrie, la distribution du mélange de deux lois géométriques s'ajuste mieux aux durées sèches :

$$P(S = x|\text{état initial sec}) = \alpha p_1(1-p_1)^{x-1} + (1-\alpha)p_2(1-p_2)^{x-1}, x = 1, 2, \dots, \tag{5.10}$$

avec $\alpha \in [0, 1]$, et Racsko *et al.* (1991) gardent la loi géométrique pour les durées pluvieuses.

La loi géométrique est également généralisée par la loi binomiale négative (de paramètres $p \in [0, 1], k > 1$) :

$$P(S = x|\text{état initial sec}) = \prod_{i=1}^{x-1} \frac{k+i-1}{i} p^k (1-p)^{x-1}, x = 1, 2, \dots. \tag{5.11}$$

Si $k = 1$, la loi binomiale négative est égale à la loi géométrique. Si $k > 1$, la loi binomiale négative à une queue plus légère que la loi géométrique. Si $k < 1$, la loi binomiale négative a

une queue plus lourde que la loi géométrique. Cette loi a été peu employée, la première utilisation semble être celle de Wilby *et al.* (1998). Cependant, une autre forme de loi binomiale négative, la loi binomiale négative tronquée, a été utilisée par Buishand (1978), Roldan et Woolhiser (1982).

Plus récemment, Salvadori et De Michele (2006) ont utilisé des distributions à queue lourde, la loi GPD, pour modéliser les durées sèches et pluvieuses.

Dans la plupart des modèles, les lois des durées sèches et pluvieuses sont supposées indépendantes. Mais Salvadori et De Michele (2006) ont montré sur une série du Nord-Ouest de l'Italie que les durées sèches et pluvieuses étaient significativement dépendantes. Ils ont donc modélisé la dépendance entre les durées sèches S et humides H, via une loi bi-variée, définie par une copule de Franck :

$$P(S \leq s, H \leq h) = C(P(S \leq s), P(H \leq h)), \tag{5.12}$$

$$\tag{5.13}$$

où

$$C(u,v) = \frac{1}{\ln \delta} \ln(1 + \frac{(\delta^u - 1)(\delta^v - 1)}{\delta - 1}), \tag{5.14}$$

avec $\delta \geq 0$ un paramètre de dépendance : $\delta > 1$ (respectivement $\delta < 1; \delta = 1$) modélise des variables aléatoires négativement dépendantes (respectivement positivement dépendantes ; indépendantes).

Des comparaisons de certains de ces modèles ont été menées, avec les critères d'AIC ou BIC (Chapman, 1997), (Wilks, 1999), (Srikanthan et McMahon, 2001).

5.2.2 Modèles de simulation des quantités de pluie

Après avoir identifié les jours pluvieux, il reste à simuler la quantité de pluie précipitée lors de ces jours. Dans la plupart des modèles, les cumuls journaliers sont considérés i.i.d. et indépendants des durées pluvieuses. Il a été montré que l'auto-corrélation des pluies journalières non nulles est faible, mais significative (Buishand, 1978), (Katz et Parlange, 1998), (Madden *et al.*, 1999)[1]. Cette auto-corrélation est due à des passages frontaux lents et, pour les régions sujettes à des cyclones, à des passages lents de cyclones. De nombreux modèles supposent que cette auto-corrélation n'a pas de conséquences importantes. L'hypothèse d'indépendance des pluies journalières non nulles a donc été largement utilisée (Wilks, 1999).

La distribution des cumuls de pluie présente une forte asymétrie positive (la médiane est inférieure à la moyenne). La loi gamma à deux paramètres a donc été la plus couramment employée pour modéliser les cumuls de pluie journalière (Wilks, 1999), et la loi exponentielle pour modéliser les intensités de pluie journalière (Salvadori et De Michele, 2006). Une hypothèse classique posée dans de nombreux modèles est de supposer les cumuls ou les intensités i.i.d..

L'hypothèse des pluies 'identiquement distribuées' a été assouplie par la considération de plusieurs lois gamma : les paramètres dépendent de la position du jour pluvieux dans la chronique des jours pluvieux et secs. Katz (1977), Chin et Miller (1980) et Guttorp (1995)

[1]Pour les pluies horaires, la corrélation temporelle a été considérée par Katz et Parlange (1995).

définissent deux distributions : l'une pour les jours pluvieux suivant un jour sec, l'autre pour les jours pluvieux suivant un jour pluvieux. Buishand (1978), Wilby *et al.* (1996) et Chapman (1998) définissent la distribution des quantités de pluie par trois lois : une loi pour les jours pluvieux isolés, une loi pour les jours pluvieux de début ou de fin d'événement, une loi pour les jours pluvieux à l'intérieur d'un événement. Cette définition s'appuie sur des statistiques observées ou sur des considérations physiques sur les différents types de pluie, comme expliqué ci-dessous.

- La moyenne des cumuls augmente généralement avec ces trois cas de figure. Par exemple, une étude de Chapman (1998) montre que la moyenne de la pluie des jours pluvieux isolés dans le Pacifique Ouest est inférieure à quatre fois la moyenne de la pluie journalière.
- Si l'on ajuste une loi gamma à chaque type de pluies (de jours isolés, ou de début ou fin d'événement, ou en cours d'événement), on constate que les paramètres de forme des lois gamma restent identiques, pour un poste donné et en un mois donné (Buishand, 1978).
- Les pluies d'un jour proviennent en général de systèmes convectifs, et les autres pluies de systèmes frontaux.
- Les événements pluvieux peuvent débuter ou terminer à n'importe quelle heure du jour, et l'événement peut déborder de seulement quelques heures sur le premier ou le dernier jour, qui ne contient dans ce cas qu'une faible partie des précipitations.

Ces considérations ont amené des hydrologues à définir un modèle de trois lois gamma, de même paramètre de forme mais de paramètres d'échelles différents. Ce modèle à quatre paramètres est parfois appelé 'common-α gamma' (ici α désigne implicitement le paramètre de forme des lois gamma). La validité de cette classification a été vérifiée par Chapman (1998) sur des longues séries d'Amérique du Nord, d'Afrique du Sud et d'Australie. Cette classification a été reprise dans les modèles non paramétriques, par exemple par Harrold *et al.* (2003b).

Un autre modèle, peu utilisé, mais capable de donner de bons résultats, d'après des études de Woolhiser et Roldan (1982), Foufoula-Georgiou et Lettenmeier (1987) et Wilks (1998), est donné par un modèle à trois paramètres de mélange de deux lois exponentielles.

Chapman (1998) a comparé avec le critère AIC différentes lois pour les quantités de pluie. Il a par exemple montré que le mélange d'exponentielles donnait de meilleurs résultats que la loi gamma.

5.2.3 Exemples et commentaires de ces modèles

Exemple de validation bayésienne d'un modèle simple

Chaouche et Parent (1999) ont proposé une validation bayésienne du générateur de pluies journalières simple de Stern et Coe (1984), appliqué à un régime de mousson (données d'une station soudano-sahélienne du Burkina-Faso). Ce générateur modélise le processus des occurrences par une chaîne de Markov d'ordre 1, et les cumuls journaliers des jours pluvieux par une loi gamma. Pour modéliser la saisonnalité, les paramètres sont exprimés par une somme de Fourier limitée à deux harmoniques. La méthode de validation bayésienne consiste à comparer les distributions a posteriori de certaines variables prédictives avec les observations. En particulier, Chaouche et Parent (1999) vérifient le comportement du modèle vis-à-vis des valeurs extrêmes. Ils considèrent pour chaque jour de l'année, la pluie maximale précipitée ce

jour, sur les 29 années dont ils disposent. Les intervalles de crédibilité à 90% de ces maxima contiennent les observations : le modèle restitue donc bien les maxima journaliers. Concernant le processus d'occurrences, ils valident la pertinence du modèle markovien par rapport à un modèle 'pile ou face' de Bernoulli (où la probabilité d'occurrence d'une pluie chaque jour est indépendante de l'état pluvieux ou sec du jour précédent). Ils montrent que le comportement central et la dispersion du nombre annuel de jours de pluie et les durées des épisodes pluvieux sont bien restitués par le modèle. Concernant les intensités simulées, le cumul annuel des pluies est bien reproduit. D'autre part, ils montrent que la loi exponentielle (cas particulier de la loi gamma) suffit à modéliser la hauteur de pluies journalières. Par contre, le modèle ne conserve pas la forme des épisodes pluvieux : l'hypothèse d'indépendance entre les cumuls journaliers successifs est mise en défaut. Pour résoudre ce problème, ils proposent d'introduire une modélisation de la dépendance entre les cumuls successifs, ou bien si les cumuls de pluie simulés sont corrects sur différents pas de temps, de réarranger l'ordre des jours de pluie, à l'intérieur d'un épisode pluvieux.

Comparaisons de modèles

Simulation des durées sèches par les modèles d'occurrences

Les valeurs extrêmes des durées sèches intéressent particulièrement les hydrologues. Ce point est donc plus développé dans la littérature que d'autres aspects de la pluie.

Comme déjà remarqué par d'autres auteurs, Wilks (1999) montre que les modèles markoviens au premier ordre sous-estiment la fréquence d'apparition des longues durées sèches. Cette sous-estimation est significative sur les postes de République Tchèque et de l'Ouest des États-Unis (caractérisé par un climat avec un faible cumul annuel et une forte variabilité saisonnière), mais peu importante sur l'Est (caractérisé par un climat avec un fort cumul annuel, et une faible variabilité saisonnière) et le Centre des États-Unis. En République Tchèque et sur l'Ouest des États-Unis, Wilks (1999) montre également que les modèles markoviens d'ordres supérieurs ou égaux à 2, les modèles markoviens hybrides et la loi binomiale négative améliorent les résultats du modèle markovien d'ordre 1. Dubrovsky *et al.* (2004) montrent que les modèles markoviens d'ordres supérieurs améliorent les résultats sur des données de République Tchèque.

De bons résultats sont donnés par le mélange de lois géométriques pour la République Tchèque, et par le modèle de loi binomiale négative pour l'Ouest des États-Unis. En Hongrie, Racsko *et al.* (1991) montrent de même que les chaînes de Markov d'ordre 1 sous-estiment fortement les extrêmes des durées sèches, tandis que le mélange de lois géométriques donne de bon résultats.

Par contre, à l'Est et au Centre des États-Unis, le modèle markovien au premier ordre donne des résultats corrects. Dans le même sens, Small et Morgan (1986) ont montré que la pluie journalière de certains postes des États-Unis était mieux modélisée par un processus markovien, que par un processus d'alternance supposant une loi gamma pour les durées sèches.

Concernant le biais des estimations des durées sèches, Wilks (1999) montre que les modèles markoviens plus complexes (hybrides, ordres supérieurs) diminuent le biais négatif des durées sèches maximales mensuelles, et peuvent le rendre positif. Les modèles d'alternance avec une

loi binomiale négative, ou un mélange de lois géométriques améliorent parfois les résultats, mais le biais est positif : les durées sèches maximales mensuelles simulées ont tendance à être plus longues que celles des observations.

Ces différences entre modèles markoviens et d'alternance dépendent de la capacité à respecter la saisonnalité. Dans chacun des modèles, les paramètres sont estimés pour chaque mois de l'année. Cependant, dans le modèle d'alternance, une durée est simulée dans la loi du mois auquel appartient le premier jour de la durée, or les durées simulées peuvent chevaucher plusieurs mois (les plus longues durées sèches observées sont supérieures à 100 jours dans l'Ouest des États-Unis).

Simulation des durées pluvieuses par les modèles d'occurrences

Les extrêmes des longueurs des durées pluvieuses semblent bien respectés par les différents modèles, pour les postes des États-Unis (Wilks, 1999). Racsko *et al.* (1991) montrent que la loi géométrique donne de bons résultats pour les durées pluvieuses. En République Tchèque, Wilks (1999) montre que le mélange de lois géométriques donne de bon résultats.

Simulation du nombre de jours pluvieux par les modèles d'occurrences

Katz et Parlange (1998) et Wilks (1999) montrent que le nombre mensuel de jours pluvieux est bien reproduit en moyenne, mais sa variance est sous-estimée par les différents modèles d'occurrence proposés. En particulier pour l'Est et le Centre des États-Unis, le modèle markovien au premier ordre est le modèle le plus adéquat, tandis que les modèles d'alternance donnent de meilleurs résultats pour les postes de l'Ouest des États-Unis.

Comparaison des modèles d'occurrences via les critères BIC et AIC

Wilks (1999) compare également les différents modèles markoviens via le critère BIC et montre des résultats cohérents avec les précédents. Il montre que les chaînes de Markov à l'ordre 1 sont souvent plus adéquates pour les postes de l'Est et du Centre des États-Unis. Dans l'Ouest, le modèle markovien hybride d'ordre 2 est majoritairement mieux adapté. De même, il compare les modèles d'alternance. Dans l'Ouest et le Centre, la loi binomiale négative de paramètre $k < 1$ est souvent préférée à la loi géométrique, et permet de simuler des durées sèches longues plus fréquentes. Dans l'Est, la loi binomiale négative de paramètre $k > 1$ est préférée à la loi géométrique, simulant moins fréquemment de longues durées sèches. Les résultats suivants sont fondés sur le critère AIC. Roldan et Woolhiser (1982) ont montré sur des sites des États-Unis, que la chaîne de Markov au premier ordre était préférable à un processus d'alternance avec des lois géométriques tronquées (pour les durées pluvieuses) et binomiales négatives tronquées (pour les durées sèches). Chapman (1997) a montré que les pluies journalières d'îles de l'Ouest du Pacifique étaient bien modélisées par un modèle markovien du premier ordre ou un processus d'alternance avec une distribution géométrique tronquée pour les postes situés au nord de la latitude 14 Nord. Sur une étude de 24 postes en Amérique du Nord, Chapman (1998) montre également que le critère AIC sélectionne les modèles de chaîne de Markov au second ordre (pour 12 postes), et les modèles d'alternance avec la loi binomiale négative tronquée (pour les 12 autres postes).

Chapman (1998) montre avec le critère AIC que le modèle markovien à sept états de Srikanthan et McMahon (1985) est plus performant que les modèles markoviens à deux états, même si ceux-ci traitent les données par classes de jours pluvieux isolés, début, milieu ou fin d'événements. Bien qu'il ne classe pas les jours pluvieux selon qu'ils sont isolés, en début, milieu ou fin d'événement, le modèle de Srikanthan et McMahon (1985) reproduit bien le nombre de jours de pluie dans chaque classe (isolé, début/fin ou milieu d'événement), ainsi que les variances des pluies annuelles et mensuelles.

Pour différents critères de comparaison, et pour un même type de climat, les résultats concordent. En revanche, les modèles n'ont pas les mêmes performances sur des climats différents. Le choix d'un modèle dépend donc essentiellement du climat de la région étudiée.

Comparaison des modèles des cumuls via les critères BIC et AIC

Il apparaît clairement dans les résultats de Wilks (1999) que le modèle de loi gamma i.i.d. donne de bien moins bons résultats que le modèle de mélange de lois exponentielles. Le critère BIC montre que le mélange de lois exponentielles est plus adapté pour l'Est et le Centre des États-Unis, et dans une majorité des postes de l'Ouest des Etats-Unis. Certains postes de l'Ouest des Etats-Unis sont mieux modélisés par le modèle 'common-α-gamma'.

Comme Wilks (1999), Cahill (2003) compare les modèles de mélange d'exponentielles, les modèles "common-α-gamma", et les modèles gamma. L'outil de comparaison utilisé par Cahill (2003) est le critère AIC. Il montre que dans l'ordre du meilleur au moins bon, le critère AIC choisit les modèles de mélange d'exponentielles, les modèles "common-α-gamma", et en dernier les modèles gamma. Par un test du rapport de vraisemblance, il montre également que les modèles gamma et mélange d'exponentielles sont significativement différents, ainsi que les modèles gamma et "common-α-gamma".

Comparaison des valeurs extrêmes simulées par les modèles de cumuls

Le biais entre la valeur maximale simulée en un poste et un mois donnés, et la valeur de même fréquence dans la série observée, est négatif sur les trois modèles (lois gamma, 'common-α-gamma', et mélange d'exponentielles) comparés par Wilks (1999) : les modèles ne reproduisent pas des valeurs maximales suffisamment élevées. Le modèle de mélange d'exponentielles et le modèle de lois gamma ont respectivement le biais le moins important, et le biais le plus important.

Comparaison des variances mensuelles des pluies journalières par les modèles de cumuls

Les résultats précédents restent vrais lorsque le critère de comparaison est la restitution de la variance mensuelle des pluies journalières. La variance reste biaisée, du fait d'une estimation par maximum de vraisemblance, et non par une méthode des moments (dans les trois modèles comparés, la moyenne n'est pas biaisée).

La sous-estimation de la variance des pluies journalières, associée à la sous-estimation de la variance du nombre de jours pluvieux ont pour conséquence la sous-estimation de la variabilité saisonnière des cumuls pluvieux (Wilks, 1999).

Conclusions

La plupart des modèles présentés ici sont ajustés sur l'ensemble des valeurs journalières. Or, les valeurs extrêmes observées proviennent souvent d'événements météorologiques inhabituels (ouragans, phénomènes convectifs méso-échelle complexes), et donc d'une population différente de celle de la majorité des observations journalières. Si la particularité des phénomènes extrêmes est prise en compte, via un mélange de modèles, ou une classification, les résultats concernant les extrêmes sont meilleurs.

Le degré de complexité du modèle dépend des caractéristiques climatiques, et du résultat souhaité. Sur certains postes, les modèles d'alternance des durées (loi binomiale négative et mélange de lois géométriques) montrent de meilleurs résultats au niveau des extrêmes et de la variabilité inter-annuelle des pluies (Wilks, 1999). Mais les modèles markoviens (hybrides ou non) sont intéressants pour la modélisation multi-site, grâce à leur capacité à prendre en compte les corrélations spatiales (Wilks, 1998).

Enfin, pour les postes où la variabilité inter-annuelle est importante (par exemple en Australie), Katz et Parlange (1998) et Wilks (1999) concluent qu'aucune combinaison de modèles n'a permis de respecter la variance inter-annuelle des cumuls mensuels, même si les résultats sont meilleurs avec des modèles plus complexes.

Évolutions récentes : introduction de lois à queue lourde, et de la dépendance entre variables

Salvadori et De Michele (2006) considèrent la dépendance entre la durée pluvieuse H d'un événement, la durée sèche S suivante, et l'intensité moyenne I de pluie précipitée pendant la période pluvieuse. Ils montrent que la durée pluvieuse et l'intensité de pluie précipitée en cette durée sont négativement associées (à partir de données du Nord Ouest de l'Italie). Cette dépendance a été également relevée par Arnaud (1997) sur des postes français, ou Heneker *et al.* (2001) sur des données australiennes. Salvadori et De Michele (2006) montrent de plus une corrélation négative entre une durée pluvieuse et la durée sèche suivante.

Salvadori et De Michele (2006) montrent que si les distributions de I, H ne sont pas à queue lourde, alors la distribution du volume $V = I \cdot H$ de pluie précipitée pendant l'événement n'est pas à queue lourde. Les trois variables I, H, S sont modélisées marginalement par des lois GPD, et conjointement par une loi multivariée de dimension trois, à l'aide d'une copule particulière. Dans le cas d'étude italien proposé, la loi de V s'ajuste visuellement bien aux données. Par ailleurs, ils développent théoriquement le calcul de la distribution du nombre d'événements dans une fenêtre temporelle donnée, et donnent un résultat asymptotique pour la convergence de cette loi lorsque la fenêtre temporelle devient infinie. Cependant, le résultat théorique ne s'applique pas dans leur cas d'étude, car la fenêtre temporelle utilisée est trop petite (120 h). Ils introduisent également un algorithme original pour approcher la distribution du volume de pluie dans une fenêtre temporelle de taille arbitraire, précédant un événement fort. La méthode est fondée sur un mélange de lois et des convolutions multi-dimensionnelles.

Un exemple : la version journalière de Shypre

Le principe du modèle journalier Shypre a été repris de Buishand (1977). Les durées sèches et pluvieuses sont générées alternativement. Les durées sèches sont simulées par une

loi binomiale négative tronquée, les durées pluvieuses sont générées par une loi géométrique. Lorsqu'on génère une durée pluvieuse, on génère en même temps les hauteurs d'eau correspondant à chaque jour de l'épisode pluvieux. Ce modèle différencie les hauteurs de pluie d'un jour pluvieux isolé, d'un jour aux extrémités d'un événement pluvieux, et d'un jour au milieu d'un événement pluvieux. Ces hauteurs sont calées sur une loi de Weibull à trois paramètres. Les cumuls de pluies des jours successifs sont supposés indépendants (faible coefficient d'auto-corrélation de la série des cumuls non nuls), de même que la durée sèche avec la durée pluvieuse qui suit. Ces lois sont calées aux variables mois par mois, pour 52 postes du Sud de la France (Arnaud, 1997).

La validation du modèle Shypre journalier repose sur la restitution des distributions de la pluie journalière maximale du mois, et de la pluie totale du mois, par une chronique simulée de taille 1000 années. Pour les pluies journalières maximales dont la moyenne et l'écart-type sont bien restitués, la distribution des valeurs simulées est très proche de la distribution des valeurs observées. Cependant, même si l'erreur relative sur l'écart-type est supérieure à ±30%, les distributions simulées et observées sont assez proches (Arnaud, 1997).

Arnaud (1997) montre que, tous mois confondus :
- sur la moyenne des pluies maximales journalières du mois, 85% des postes étudiés ont une erreur relative comprise entre ±20%,
- sur la moyenne des pluies mensuelles, 80% des postes étudiés ont une erreur relative comprise entre ±20%,
- sur l'écart-type des pluies maximales journalières du mois, 90% des postes étudiés ont une erreur relative comprise entre ±30%,
- sur l'écart-type des pluies mensuelles, 75% des postes étudiés ont une erreur relative comprise entre ±30%.

De ces points de vue, les performances du générateur Shypre journalier sont satisfaisantes.

Le modèle horaire de Shypre est décrit plus loin dans la section 5.3 : au lieu de simuler une chronologie continue d'alternance entre périodes sèches et pluvieuses, il simule une succession d'événements composés d'averses.

5.2.4 Modèles non-paramétriques

Les modèles non paramétriques ont été proposés comme alternative aux modèles paramétriques multivariés, devenus trop complexes (Young, 1994). On rencontre les modèles non paramétriques principalement dans la modélisation multi-site des précipitations, ou dans la modélisation conjointe de variables climatiques (pluie, température, évapo-transpiration, pression atmosphérique, couverture neigeuse, etc.). Ces derniers modèles sont utilisés par exemple dans des modèles hydrologiques pluie-débit ou de ressources en eau, pour des générations de scénarios de changement climatique, ou en agronomie.

La méthode non paramétrique la plus répandue est appelée en anglais 'nearest-neighbour resampling', que nous traduirons ici par méthode du voisinage le plus proche. Cette méthode est présentée par Rajagopalan et Lall (1995) et Lall et Sharma (1996). Comme son nom l'indique, cette méthode consiste à effectuer des tirages avec remise dans une fenêtre temporelle centrée sur la date pour laquelle on veut simuler des variables climatiques. L'utilisation de la fenêtre temporelle permet de reproduire les variations saisonnières des variables.

La méthode des analogues, utilisée en météorologie, est un cas particulier de la méthode du voisinage le plus proche (Zorita *et al.*, 1995). La méthode du voisinage le plus proche a été

appliquée par Lall et Sharma (1996) pour générer des séries temporelles hydrologiques, par Rajagopalan *et al.* (1996) pour simuler la pluie journalière avec un modèle de Markov non-homogène, par Lall *et al.* (1996) pour simuler des occurrences de pluie avec une méthode de retirage homogène par saison, par Rajagopalan *et al.* (1997) pour simuler plusieurs variables climatiques journalières, par Tarboton *et al.* (1998) avec une procédure de désagrégation pour simuler des débits, par Wojcik et Buishand (2003) pour simuler la pluie et la température sur des pas de temps de 6 heures, à l'aide d'une méthode de désagrégation (les méthodes de désagrégation sont présentées à la section 5.3.3). Le modèle de Wojcik et Buishand (2003) reproduit d'ailleurs bien les valeurs extrêmes de pluie.

Principe de la méthode du voisinage le plus proche

La méthode permet de simuler une série temporelle constituée de $m \geq 1$ variable(s) climatique(s). On note $\mathbf{X}_u = (x_1(u), \ldots, x_m(u))$ le vecteur des m variables observées le jour u, et on note $\mathbf{X}_t^* = (x_1^*(t), \ldots, x_m^*(t))$ le vecteur des m variables du jour t de la série simulée.

Supposons que les variables ont été simulées du jour 1 au jour t. La simulation des variables du jour $t+1$ repose sur les données $\mathbf{X}_t^*$: on recherche dans une fenêtre centrée en t de la chronique observée, les situations climatiques les plus 'proches' de la situation $\mathbf{X}_t^*$.

- `A propos de la fenêtre temporelle :`
  Si la date en cours de simulation est le jour t, la fenêtre temporelle des observations autour de cette date est définie par l'ensemble des jours autour du jour Julien correspondant à la date t, et ceci pour les différentes années de la série d'observation. Donnons un exemple pour une fenêtre temporelle de 61 jours : si l'on souhaite simuler le jour t numéro 380 de la série simulée, on considère le jour $u = 380$ modulo 365, et la fenêtre temporelle est définie par les intervalles de jours $[u-30$ modulo 365,$u+30$ modulo $365]_i$ pour chaque année i de la série observée.
- `A propos de la distance :`
  La notion de voisinage nécessite la définition d'une distance, et au préalable la définition des variables prises en compte dans le calcul de la distance. Ces variables sont appelées variables de conditionnement. Par exemple, les variables de conditionnement peuvent être le cumul de pluie et la température du jour t simulé. Les plus 'proches' situations correspondent alors aux situations de la fenêtre temporelle, avec une pluie et une température 'proches' de la situation en X_t^*. La notion de proximité est objectivée par une distance. Les distances les plus couramment employées sont la distance Euclidienne pondérée (voir par exemple (Brandsma et Buishand, 1998)), et la distance de Mahalanobis (voir par exemple (Kendall *et al.*, 1983) pour la définition de cette distance, et (Wojcik et Buishand, 2003) pour un exemple d'utilisation).

Une des situations voisines est ensuite tirée aléatoirement, avec remise. Soit $\mathbf{X}_v$ cette situation observée. La situation $\mathbf{X}_{t+1}^*$ est alors égale à la situation $\mathbf{X}_{v+1}$, c'est-à-dire la situation du jour suivant dans la chronique observée.

La méthode ci-dessus nécessite de préciser :
- la définition d'une distance entre situations climatiques.
- la largeur de la fenêtre temporelle autour du jour t. La fenêtre temporelle permet de respecter le cycle annuel des variables climatiques, et ne pose pas le problème du découpage de l'année en saisons. Brandsma et Buishand (1998), Sharma et Lall (1999), Wojcik et Buishand (2003) utilisent une fenêtre de 61 jours.

- l'entier $k \in \mathbb{N}^*$ égal au nombre de voisins le plus proches. k est de l'ordre de $\sqrt{n}$, où n est le nombre d'observations utilisées dans la fenêtre temporelle (Lall et Sharma, 1996). n est donc égal au nombre d'années de la série observée multiplié par la largeur de la fenêtre temporelle. Buishand et Brandsma (2001) donnent quelques conseils sur le choix de k.
- la loi de probabilité discrète pour le tirage du voisinage $\mathbf{X}_v$. Souvent, la probabilité de tirer le $j$ème plus proche voisin, $1 \leq j \leq k$ est donnée par $p_j = \frac{1/j}{\sum_{j=1}^{k} 1/j}$.

Exemples et résultats de ces méthodes

Dans un modèle simulant les pluies et les températures, Brandsma et Buishand (1998) et Wojcik et Buishand (2003) montrent que la méthode du voisinage le plus proche est capable de :

- bien reproduire la distribution des valeurs extrêmes des maxima journaliers (et en 6 h avec la procédure de désagrégation de Wojcik et Buishand (2003)). Les maxima en un, quatre et dix jours sont bien reproduits, et leurs distributions dans un diagramme de Gumbel présentent une allure alignée, suggérant une loi Gumbel. Wojcik et Buishand (2003) utilisent une série de 42 ans de données. A ce propos, nous rappelons ici que l'estimateur du paramètre de forme d'une loi GEV est fortement variable pour une taille d'échantillon égale à 42. D'autre part, Koutsoyiannis et Baloutsos (2000) ont montré que, sur une portion d'une série pluviométrique longue, la distribution des maxima annuels peut présenter une allure alignée sur un diagramme de Gumbel, alors que la distribution des maxima de la série entière présente une allure convexe.
  La procédure de désagrégation de Wojcik et Buishand (2003) permet également d'en déduire la distribution des maxima saisonniers en 6 h. Wojcik et Buishand (2003) présentent le résultat dans un diagramme de Gumbel. La distribution a une courbure convexe et un plafond. La courbure suggère une distribution GEV pour les maxima. Le plafonnement de la distribution est dû à la méthode, qui ne peut simuler des valeurs supérieures aux observations.
- bien reproduire les coefficients d'auto-corrélation de la série temporelle des précipitations,
- bien reproduire les moments d'ordre 1 et 2 des cumuls mensuels.

Les modèles paramétriques de pluie journalière ont été adaptés en modélisation non paramétrique. On donne ci-dessous trois exemples.

1. Le modèle présenté par Sharma et Lall (1999) est une version non paramétrique des modèles combinant un processus d'occurrences avec un modèle de cumuls de pluie journalières. Dans le processus d'occurrences, la variable de conditionnement pour simuler les durées est :
   - la durée de l'événement pluvieux précédent, si l'on simule une durée sèche,
   - la durée sèche précédente, si l'on simule une durée pluvieuse.

   Dans le modèle des cumuls de pluie, les variables de conditionnement sont :
   - la date du début de l'événement : pour respecter la fonction d'auto-corrélation des pluies journalières d'un événement,
   - la durée de l'événement : pour tenir compte de la dépendance entre les cumuls de pluie et la durée des événements pluvieux (les événements de type orage sont souvent courts et de volumes élevés, les événements de type dépression sont souvent longs et de volumes faibles).

– le cumul de pluie du jour précédent.

Sharma et Lall (1999) montrent sur une longue série de 123 années à Sydney (Australie) que le modèle reproduit bien la variation saisonnière de la moyenne, la variance, et l'asymétrie de la pluie journalière. De même, la variation saisonnière du nombre de jours pluvieux dans les fenêtres mobiles de 60 jours est bien reproduite.

Sharma et Lall (1999) ne précisent pas si la distribution des valeurs extrêmes est préservée, ni le type de distribution des valeurs extrêmes qui ressort de cette modélisation.

D'autre part, le modèle ne respecte pas la variabilité inter-annuelle, causée par les fluctuations climatiques basse-fréquence. L'inclusion d'une variable de conditionnement de périodicité basse-fréquence (relative par exemple aux interactions océan-atmosphère, telles que l'Oscillation du Sud El-Niño dans le cas de Sydney) permettrait peut-être d'améliorer la variabilité inter-annuelle. Par exemple, Harrold *et al.* (2003a) introduisent des variables de conditionnement caractérisant l'état très sec, sec, moyen, humide, ou très humide des périodes précédentes (trois mois, un an, cinq ans). Ces variables ne sont pas exogènes aux données pluviométriques, ce qui peut être utile lorsqu'on ne dispose pas des autres données climatologiques. Harrold *et al.* (2003a) montrent que cette méthode donne de bons résultats sur la série de Sydney.

2. Le modèle proposé par Harrold *et al.* (2003a,b) combine également un processus d'occurrences et un modèle de cumuls des pluies. La particularité de ce modèle, par rapport à celui de Sharma et Lall (1999), est d'introduire un classement des jours pluvieux (jour pluvieux isolé, jour de début, fin ou milieu d'événement), comme déjà introduit dans certains modèles paramétriques. D'autre part, il génère les cumuls avec une densité estimée par une méthode de noyaux. La méthode, appliquée aux données de Sydney, montre que la moyenne, l'écart-type et le coefficient d'asymétrie des pluies journalières sont bien reproduits, ainsi que le coefficient d'auto-corrélation d'ordre 1.

3. Rajagopalan *et al.* (1996) présentent un modèle de Markov au premier ordre pour la génération de la pluie journalière sur un site. La matrice de probabilités de transitions de la chaîne de Markov varie jour après jour tout au long de l'année, elle est estimée via une méthode de noyau. Les cumuls de pluie sont également simulés à partir d'une distribution estimée par une méthode de noyaux. Une application du modèle à la pluie de Salt Lake City montre que les statistiques de pluie sont bien reproduites aux échelles saisonnières et annuelles.

Limites des approches non paramétriques

Les simulations sont dépendantes des observations : si des erreurs existent dans les données, elles seront propagées dans les simulations. Plus la série observée est longue (et de bonne qualité), plus la qualité de la simulation sera meilleure. Si la série observée est trop courte pour contenir des événements forts, la simulation n'en contiendra pas. Ce type de modèle ne permet pas l'extrapolation. Pour éviter de ne retirer que des valeurs observées, Harrold *et al.* (2003b) proposent de modéliser les quantités de pluie journalière à l'aide d'une fonction de densité estimée par la méthode non-paramétrique des noyaux.

5.3 Générateurs de pluie à des pas de temps inférieurs à 24 h

Les séries de mesure de la pluie à un pas de temps inférieur à 24 h sont bien moins renseignées que les séries journalières. Or, la connaissance des processus pluvieux à des faibles pas de temps est importante en hydrologie urbaine (en particulier pour l'estimation des crues). À des pas de temps fins, la pluviométrie est beaucoup plus variable qu'au pas de temps journalier, les modèles sont donc plus compliqués. On distingue différentes approches conceptuelles de simulation de la pluie au pas de temps horaire. Nous présentons d'abord le modèle Shypre, une approche directe originale pour construire des hyétogrammes en simulant des averses successives. Nous présentons ensuite deux autres types de modèles. Les modèles d'agrégation simulent des arrivées d'entités pluvieuses, qui peuvent se chevaucher dans le temps, et définissent la pluie à chaque instant par la somme des entités pluvieuses présentes à cet instant. Enfin, les modèles de désagrégation simulent la pluie à un pas de temps donné, puis la désagrègent en un pas de temps inférieur.

5.3.1 Modèle de construction de hyétogrammes : Shypre

Inspiré des travaux de Tourasse (1981), puis Lebel (1984), Shypre est un modèle de simulation pluie-débit développé au Cemagref d'Aix-en-Provence avec les thèses de Cernesson (1993) et Arnaud (1997). Ce modèle est fondé sur une description des hyétogrammes, au moyen de variables aléatoires. Le modèle reconstruit des événements pluvieux constitués d'averses, par simulation de ces variables aléatoires (Arnaud, 1997).

Description des événements et des averses

- Les événements pluvieux sont d'abord extraits de la série journalière : un événement pluvieux est une successions de jours pluvieux, pour lesquels la pluie journalière est supérieure ou égale à 4 mm, et l'un au moins des jours a une pluie journalière supérieure à 20 mm.
- Une fois identifié au pas de temps journalier, l'événement est décrit au pas de temps horaire par des averses et des périodes sèches. Une averse est définie par une pluie continue avec un seul maximum. Si une pluie continue, au pas de temps horaire, présente deux pics, on dit qu'il y a deux averses. Par convention, le minimum relatif appartient à la première averse. À l'intérieur d'un événement, un regroupement d'averses, sans période sèche, forme une période pluvieuse. Un événement contient une ou plusieurs périodes pluvieuses, séparées d'une ou plusieurs heures sans pluie.

La figure 5.1 illustre la définition d'événement pluvieux, de périodes pluvieuse et sèche, et d'averses.

Les variables descriptives des hyétogrammes sont :

- NE le nombre d'événements par an ou par saison,
- NG le nombre de périodes pluvieuses par événement,
- DIA la durée sèche entre deux périodes pluvieuses (en heures),
- NA le nombre d'averses par périodes pluvieuses,
- DA la durée d'une averse (en heures),

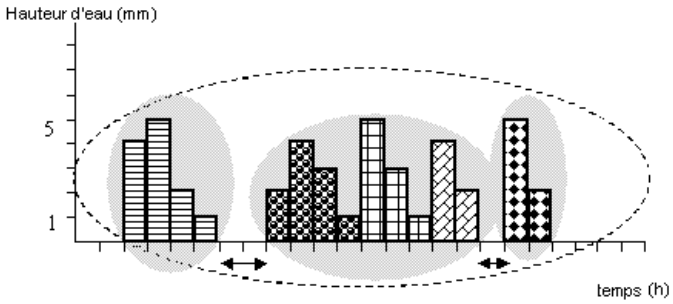


Fig. 5.1: Description d'un événement pluvieux, au pas de temps horaire. L'événement est constitué de trois périodes pluvieuses. La première période pluvieuse contient une averse, la deuxième en contient trois, et la dernière en contient une. Figure issue de (Arnaud, 2004).

- VOL le volume d'une averse (en dixièmes de mm) [2]
- RX le rapport entre la pluie maximale en 1 heure de l'averse et le volume de l'averse (VOL), RX n'a de sens que pour les averses de plus d'une heure et $1/DA \leq RX \leq 1$.

Arnaud (1997) distingue deux populations d'averses : les averses apportant les plus grandes quantités d'eau sont appelées **averses principales**, les autres averses sont appelées **averses ordinaires**. Une autre variable doit définir le nombre d'averses principales par événement. Arnaud *et al.* (2003) fait l'hypothèse que ce nombre est égal au nombre de jours consécutifs avec une pluie journalière supérieure à 20 mm. L'intérêt d'expliquer ce nombre par l'information journalière est de permettre, pour d'autres études, la régionalisation du générateur Shypre. Une nouvelle variable est donc introduite :

- $DUR20C$ le nombre de jours consécutifs avec plus de 20 mm de pluie journalière.

Arnaud (2004) a analysé en détail les événements extrêmes :

- les événements extrêmes ne sont pas nécessairement dus aux plus fortes averses, surtout en climat méditerranéen.
- Mais un événement présentant la plus forte pluie journalière dure souvent dans le temps, et présente plusieurs journées successives avec plus de 20 mm. Il est composé d'averses de volumes forts, mais pas des plus forts. La pluie maximale journalière n'est pas constituée d'une seule averse principale mais de plusieurs fortes averses (principales et ordinaires). Le nombre d'averses principales d'un événement a également un rôle important sur le caractère extrême de l'événement.
- D'autre part, les plus fortes averses d'une série (principales ou ordinaires) ont tendance à être regroupées dans un même événement. Ce phénomène est appelé '**persistance des averses**'. Par exemple, sur un poste de Saint-Michel du Touvet dans le massif de la Chartreuse, Arnaud (2004) montre que le plus fort événement pluvieux observé a une pluie totale de 234 mm, dont 152 mm sont tombés en 24 heures. Cet événement possède trois averses principales. Or ces trois averses principales sont aussi les trois plus fortes averses principales de la saison hiver. De plus, les deux plus fortes averses ordinaires de la saison sont dans cet événement, et placées à côté de la plus forte averse principale.

[2] La caractérisation des averses par leur volume semble préférable à la caractérisation par leur intensité. En effet, si le mode de dépouillement des données est de mauvaise qualité, la durée et l'intensité des averses risquent d'en être influencées. De plus, la corrélation entre le volume et la durée est moins importante qu'entre l'intensité et la durée.

Arnaud (2004) a montré que le phénomène de persistance se traduit par une dépendance entre les averses fortes :

- lorsqu'il y a plusieurs averses principales dans un événement pluvieux, ces averses ne sont pas indépendantes les unes des autres.
- Lorsque le nombre d'averses principales augmente, on a logiquement une plus grande probabilité d'avoir une averse principale forte, mais on constate aussi que les autres averses principales et les averses ordinaires de l'événement sont fortes.
- Le degré de la persistance des averses principales est croissant avec la variable $DUR20C$.
- Plus l'averse principale la plus forte de l'événement est forte, plus les autres averses principales de l'événement seront fortes, et ceci même pour les postes du climat tempéré, et d'autant plus pour les postes méditerranéens.
- La persistance des averses principales n'est pas marquée en climat tempéré.
- La persistance des averses principales est marquée en climat méditerranéen.

Enfin, Arnaud (1997) montre un phénomène d'agglomération des averses les plus fortes autour de la plus forte averse de l'événement. Il est important de prendre en compte cet aspect dans la simulation car il est responsable de pluies extrêmes. Il est surtout visible dans les postes d'altitude ou de climat méditerranéen.

Simulation des événements

Le tableau 5.1 liste les variables descriptives et les lois de probabilité utilisées pour le modèle, classées dans l'ordre de leur tirage dans le modèle. Ces lois de probabilités ont été déterminées par l'étude d'une cinquantaine de postes pluviographiques situés dans les départements du pourtour méditerranéen français.

Les incertitudes de modélisation sont dues à la fois au choix des lois, à l'estimation des paramètres, mais également à la typologie des averses. En effet, la typologie des averses est une étape délicate qui peut influencer les estimations des paramètres des lois des variables.

La modélisation du nombre d'averses principales par événement et la prise en compte de la dépendance entre les variables RXP et $VOLP$ a été formulée dans le modèle Shypre à l'aide de lois empiriques. De même, la persistance des averses, qui se traduit par une dépendance des fréquences des volumes des averses, a été modélisée de manière empirique. Enfin, l'agglomération des fortes averses est modélisée en regroupant les averses les plus fortes d'un événement autour de la plus forte averse principale de l'événement.

Le schéma 5.2 suivant représente le déroulement de la simulation des événements dans le modèle Shypre.

Validation du modèle

Le modèle Shypre parvient à reproduire les quantiles de pluie sur différents climats (méditerranéen, tempéré, alpin). En particulier, l'inclusion de la dépendance entre RXP et $VOLP$ a conduit le modèle à bien restituer les pluies maximales en 1 h. La modélisation de l'agglomération des fortes averses permet de bien restituer les quantiles de durée 6 h, et la modélisation de la persistance associée à celle de l'agglomération des averses permet de bien reproduire les quantiles de durées 24 h.

Variable	*Définition*	*Loi de probabilité*
NE	Nombre d'événements pluvieux par an	Poisson
NG	Nombre de périodes pluvieuses par événement	géométrique
NA	Nombre d'averses par périodes pluvieuses	géométrique
$DUR20C$	Nombre d'averses principales par événement	empirique, estimée régionalement
$VOLP$	Volume des averses principales	exponentielle
$VOLO$	Volume des averses ordinaires	exponentielle
DAP	Durée des averses principales (14 à 20 heures maximum selon la saison)	**Hiver** : $DAP \leq 10$: Poisson ; $10 \leq DAP \leq 20$: uniforme **Eté** : $DAP \leq 8$: Poisson ; $8 \leq DAP \leq 14$: uniforme
DAO	Durée des averses ordinaires (12 h maximum)	$DAO \leq 6$: Poisson ; $6 \leq DAO \leq 12$: uniforme
RXP	Rapport de la pluie maximale en 1 heure de l'averse sur le volume de l'averse (cas averse principale)	empirique, dépendante de $VOLP$
RXO	Rapport de la pluie maximale en 1 heure de l'averse sur le volume de l'averse (cas averse ordinaire)	empirique
DIA	Durée sèche entre les périodes pluvieuses	$DIA \leq 12$: géométrique ; $DIA \geq 12$: uniforme

Tab. 5.1: Liste des variables descriptives et des lois utilisées par le modèle. Sauf RXP et $VOLP$ et si cela est mentionné dans la description des techniques de simulation ci-dessous, les différentes variables descriptives sont supposées indépendantes entre elles. L'hiver correspond aux mois décembre-mai, et l'été correspond aux mois juin-novembre.

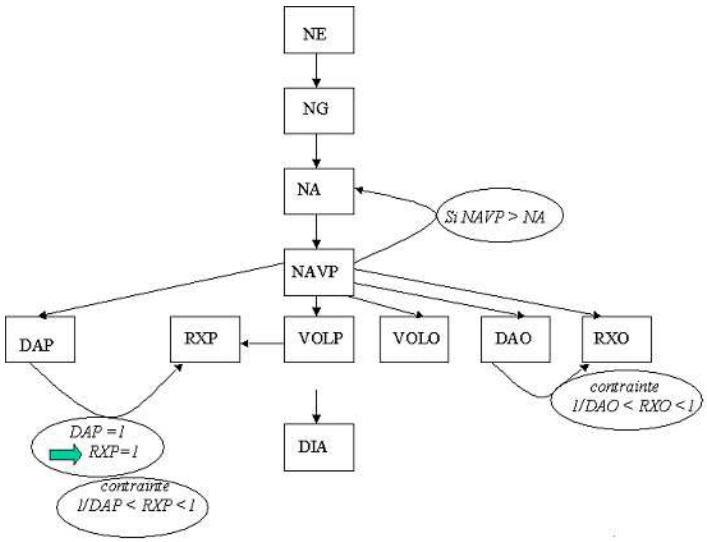


Fig. 5.2: Organigramme du programme de simulation des événements au pas de temps horaire, par Shypre

Enfin, une version journalisée de ce modèle a été proposée, dans laquelle la plupart des paramètres est estimée régionalement, tandis que d'autres sont expliqués par l'information

journalière locale. Nous analyserons l'incertitude des quantiles de pluies de 24 h simulés par cette version journalisée dans le prochain chapitre.

5.3.2 Modèles fondés sur des regroupements ou agrégation d'entités pluvieuses : Bartlett-Lewis et Neyman-Scott

Parmi les modèles de génération de pluie, les modèles fondés sur des agrégations d'entités pluvieuses font l'objet de très nombreuses publications. Ces modèles sont construits à partir de processus ponctuels : la propriété de regroupement des événements pluvieux a été modélisée par deux processus de Poisson, correspondant à un processus d'arrivée d'événements, et un processus d'arrivée des averses à l'intérieur des événements. A l'origine, ces modèles ont été empruntés à un modèle de distribution spatiale des galaxies de Neyman et Scott (1952) pour décrire la propriété de regroupement des galaxies.

Deux modèles, assez proches, ont été proposés et décrivent un processus continu de la pluie : ce sont les modèles de Neyman-Scott et de Bartlett-Lewis Rodriguez-Iturbe (1987, 1988). Ils sont utilisés pour décrire la pluie sur des pas de temps allant de quelques minutes à plusieurs jours. Deux processus de Poisson modélisent les occurrences des événements et des entités pluvieuses qui les composent. À chaque unité de temps, le cumul est égal à la somme des contributions de pluie apportées par la ou les entités pluvieuses présentes pendant cette unité de temps. Les entités pluvieuses sont décrites par leurs origines, leurs durées de vie et leurs intensités moyennes. Le processus de génération de la pluie dans ces modèles est décrit ci-dessous :

- l'origine des événement suit une loi de Poisson,
- chaque événement contient une ou plusieurs entités pluvieuses (ce point est modélisé différemment dans les modèles de Neyman-Scott et de Bartlett-Lewis),
- la durée de vie des entités suit une loi exponentielle,
- les intensités moyennes des entités suivent une loi exponentielle,
- à chaque pas de temps, l'intensité est obtenue par la somme des intensités des entités présentes sur le pas de temps.

Les origines temporelles des entités au sein des événements sont positionnées différemment suivant le processus de Neyman-Scott ou de Bartlett-Lewis utilisé :

- dans l'approche de Neyman-Scott, le nombre d'entités dans un événement pluvieux suit une loi de Poisson ou une loi géométrique et leur positionnement à partir de l'origine de l'épisode suit une loi exponentielle. La durée totale de l'événement se déduit du nombre, de la position et de la durée des entités composant l'événement. Le cas de Neyman-Scott est illustré dans la figure 5.3,
- dans l'approche de Bartlett-Lewis, la durée de l'événement est simulée en premier, suivant une loi exponentielle. Les dates d'origine des entités de l'événement sont alors simulées suivant un processus de Poisson, jusqu'à ce que cette date dépasse la durée de l'événement. Dans ce cas, le nombre d'entités par événements découle de la durée de l'événement et du résultat du processus poissonnien. On peut montrer qu'il suit une loi géométrique (Rodriguez-Iturbe, 1987).

Les deux modèles possèdent chacun cinq paramètres. Cowpertwait (1998) a montré que les deux modèles avaient les mêmes propriétés au second ordre (c'est-à-dire variance, covariance

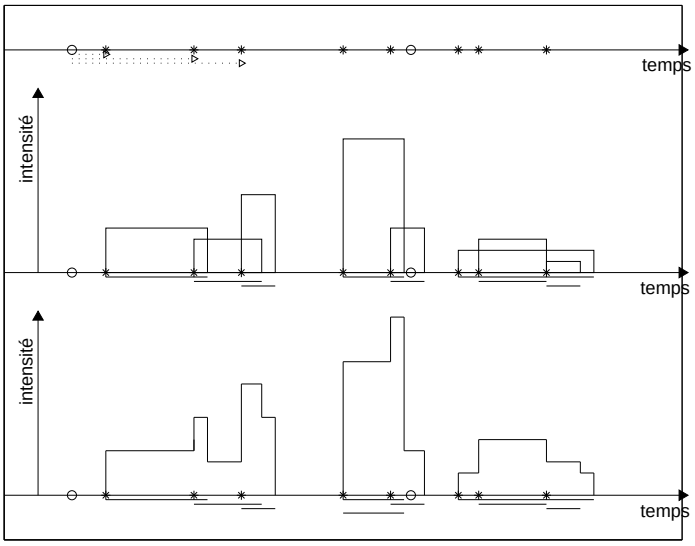


Fig. 5.3: Graphique du dessus : les origines des événements arrivent suivant un processus de Poisson. Chaque événement génère ensuite un nombre aléatoire d'entités pluvieuses, et le temps d'attente entre l'origine de l'événement et l'arrivée d'une entité est distribué selon une loi exponentielle. Graphique du milieu : les entités pluvieuses ont une durée et une intensité moyenne distribuées selon des lois exponentielles. Graphique du bas : l'intensité totale en chaque instant est égale à la somme de toutes les intensités présentes à cet instant.

et moyenne des cumuls de pluie sur différentes durées). Les modèles originaux sous-estiment les quantiles extrêmes des pluies maximales entre 1 heure et 24 heures, sur-estiment le taux de périodes sèches, et reproduisent mal l'auto-corrélation des pluies journalières. Rodriguez-Iturbe (1988) a donc modifié le modèle Bartlett-Lewis en faisant varier le paramètre de la loi exponentielle des durées des entités, suivant une loi gamma. Cela a permis de mieux reproduire les périodes sèches, l'auto-corrélation des intensités et les valeurs extrêmes (Velghe *et al.*, 1994).

Pour améliorer l'auto-corrélation des pluies journalières et les événements extrêmes, Onof et Wheater (1994) modélisent les intensités par une loi avec une queue plus lourde que la loi exponentielle : la loi gamma. Le modèle ainsi modifié améliore la restitution des distributions des maxima annuels, mais cette amélioration n'est pas encore satisfaisante selon différents auteurs (Onof et Wheater, 1994), (Khaliq et Cunnane, 1996), (Verhoest *et al.*, 1997), (Cameron *et al.*, 2000a).

Plus particulièrement, des études sur des postes de Belgique et du Royaume Uni, par Verhoest *et al.* (1997) et Cameron *et al.* (2000a), ont montré que le modèle de Bartlett-Lewis modifié par Onof et Wheater (1994) reproduit bien les extrêmes des pluies de 24 h : les intervalles de confiance des quantiles simulés contiennent les intervalles de confiance des observations, et la médiane des quantiles simulés de pluie maximale en 24 heures est convexe sur le diagramme de Gumbel (Cameron *et al.*, 2000a). Cependant, le modèle sous-estime encore les valeurs extrêmes sur les petites durées : pour 120 simulations, la médiane des quantiles (de périodes de retour de 1 an à 100 ans) simulés de pluie maximale en 1 heure a une allure légèrement concave sur un diagramme de Gumbel, et les observations sont supérieures à l'intervalle de confiance à 95% des quantiles simulés, obtenu par les 120 simulations.

Or, la restitution des pluies maximales en 1 h est un sujet important en hydrologie, notamment en hydrologie urbaine. Pour mieux estimer les pluies de 1 h, Cameron *et al.* (2000a) suggèrent d'utiliser une loi plus flexible que la loi gamma pour les intensités, par exemple une loi GPD. Une autre idée, empruntée à Cowpertwait et O'Connell (1997), consiste à distinguer la population des entités pluvieuses de type convectif, de celles de types stratiforme. Cette méthode a permis de reproduire avec succès les valeurs extrêmes. Dans la même idée, Cameron *et al.* (2000c) distinguent les entités de fortes intensités et les autres. Ils simulent les intensités des entités de fortes intensité par une loi de Pareto généralisée, les autres entités par une loi exponentielle. Cette fois, les valeurs extrêmes des hauteurs de pluie sont bien reproduites au pas de temps horaire, mais légèrement surestimées au pas de temps journalier. Notons que Cameron *et al.* (2000b) utilisent une procédure GLUE sur les paramètres de la loi GPD, pour estimer l'incertitude autour des quantiles simulés. La médiane et les intervalles d'incertitudes autour des quantiles de pluie maximale en 1 et 24 heures, déterminés par la méthode GLUE aux paramètres de la loi de Pareto généralisée sont satisfaisants.

Concernant les statistiques de la distribution centrale des pluies simulées (moyenne et variance des pluies en 1 h et 24 h, coefficients d'auto-corrélation, proportion de périodes sèches, moyenne et variance des durées sèches et pluvieuses, nombre moyen d'événements par saison), le modèle de Bartlett-Lewis modifié par Onof et Wheater (1994) donne d'assez bons résultats sur des données du Royaume-Uni, selon Cameron *et al.* (2000a). Ils notent cependant une sur-estimation de la moyenne de la pluie en 1 h, et une sous-estimation des durées moyennes des événements en 1 h et 24 h.

Enfin, le modèle de Neyman-Scott a été généralisé au cas spatio-temporel par Cowpertwait *et al.* (2002). Frost *et al.* (2005) ont montré que les courbes Hauteur-Durée-Fréquence estimées à partir des simulations du modèle Neyman-Scott spatio-temporel restent significativement proches des observations.

Estimation des paramètres

L'estimation des paramètres de ces modèles est délicate : plusieurs procédures d'estimation ont été proposées. La méthode du maximum de vraisemblance n'est pas réalisable, en particulier avec les données agrégées (Rodriguez-Iturbe, 1988) : les pluies horaires simulées par le modèle ont une structure de dépendance et une distribution marginale compliquées.

La méthode la plus fréquemment utilisée, parce qu'elle est réalisable et possède un sens physique, généralise la méthode des moments. Elle consiste à établir des équations entre des caractéristiques simulées et observées. Une méthode d'optimisation permet de résoudre les équations. Le choix des caractéristiques est discuté par Cowpertwait *et al.* (1996), Khaliq et Cunnane (1996), Smithers *et al.* (2002). Parmi les caractéristiques importantes, on distingue celles relatives aux valeurs extrêmes et au taux de périodes sèches.

Favre *et al.* (2004) donnent une nouvelle technique d'estimation des paramètres. Ils calculent les paramètres en deux étapes de façon à diminuer le nombre de paramètres à estimer dans la procédure de minimisation de l'erreur entre le modèle et les observations. Par ailleurs, ils proposent une méthode d'évaluation des incertitudes avec des formules de Taylor pour estimer la variance des paramètres.

5.3.3 Modèles de désagrégation

Bien que le nombre de séries de mesures journalières est assez important dans de nombreuses régions du monde, les données à des pas de temps inférieurs à la journée sont plus rares. Les modèles de désagrégation ont été proposés pour tenter de désagréger l'information de pluie contenue dans une série, afin de créer une nouvelle série avec un pas de temps plus fin. Ces approches reposent souvent sur des relations entre les moments et les pas de temps de mesure de la pluie : ces relations sont établies via des analyses fréquentielles régionales, ou des modèles d'échelle.

Utilisation des modèles de Neyman-Scott et Bartlett-Lewis avec des données journalières

Certains auteurs Bo *et al.* (1994), Koutsoyiannis et Onof (2001), Cowpertwait *et al.* (1996), Smithers *et al.* (2002), Gyasi-Agyei (2005) ont tenté d'appliquer les modèles Neyman-Scott ou Bartlett-Lewis avec des données journalières. La méthode générale dans ce cas est d'expliquer les paramètres du modèle Neyman-Scott ou Bartlett-Lewis par des statistiques journalières, ou par une information régionale sur les pluies horaires.

Cowpertwait *et al.* (1996), Smithers *et al.* (2002) montrent que les performances des modèles Neyman-Scott ou Bartlett-Lewis, dont les paramètres sont estimés avec uniquement l'information journalière sont mauvaises. Compte tenu des différents processus de pluie générant par exemple les pluies maximales en 1 h et les pluies maximales journalières (orages et dépressions), les niveaux d'agrégation supérieurs ne contiennent pas suffisamment d'informations sur le comportement de la pluie à des pas de temps plus fins. Cowpertwait *et al.* (1996), Gyasi-Agyei (2005) ont alors suggéré de considérer des relations empiriques régionales entre les variances horaires et journalières, ce qui permet d'estimer la variance des pluies horaires à partir d'une estimation de la variance des pluies journalières. Gyasi-Agyei (2005) montre de bons résultats au niveau des extrêmes.

Selon Smithers *et al.* (2002), la méthode régionale n'est pas applicable en Afrique du Sud pour cause de manque de données : ils montrent alors une relation approximativement linéaire entre le logarithme de la variance des pluies et le logarithme de la durée, pour des durées supérieures à 1 h et lors des saisons pluvieuses.

Smithers *et al.* (2002) utilisent le modèle de Bartlett-Lewis modifié par Onof et Wheater (1994) et montrent que les quantiles simulés, pour des pas de temps inférieurs à 24 h, sont corrects pour l'Afrique du Sud.

Processus de cascades multiplicatives

Les processus de cascades multiplicatives distribuent, de manière aléatoire, une quantité mesurée sur un intervalle initial, sur b subdivisions de cet intervalle. Ce processus de distribution est ensuite répété sur les subdivisions, et ainsi de suite. L'idée de ce modèle est empruntée à un modèle de turbulence de l'énergie cinétique. On présente ce processus de manière succincte. Pour plus de détails, voir (Gupta et Waymire, 1993), (Over et Gupta, 1994), (Over et Gupta, 1996).

On note $I(T_0)$ (ou $H(T_0)$) l'intensité (ou le cumul) de pluie mesurée sur le pas de temps initial T_0, et qu'il s'agit de répartir sur des pas de temps inférieurs à T_0. La figure 5.4

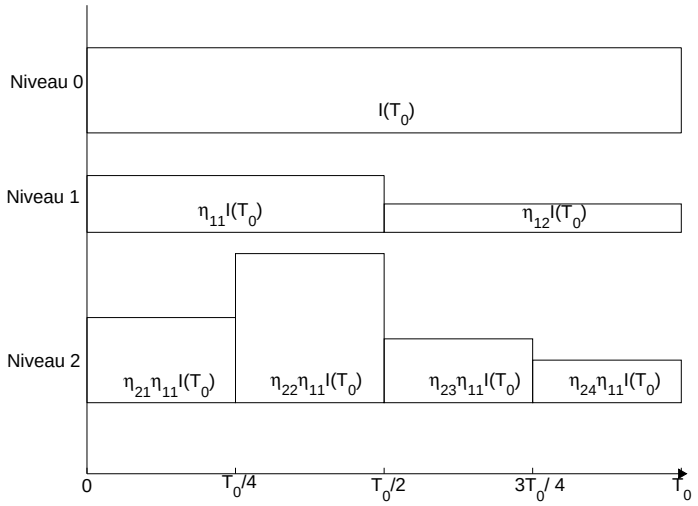


Fig. 5.4: Schéma descriptif de la construction d'une cascade multiplicative

présente un exemple simple, où le pas de temps est successivement divisé par $b = 2$. Le niveau correspondant à l'intensité initiale est le niveau 0. Au niveau 1 on a deux pas de temps de longueur $\Delta t_1 = \frac{T_0}{2}$ et deux intensités dont les valeurs sont respectivement :

$$I_{11} = \eta_{11} I(T_0) \tag{5.15}$$
$$I_{12} = \eta_{12} I(T_0) \tag{5.16}$$

$\eta_{12}, \eta_{11} \geq 0$ et $\frac{\eta_{11}+\eta_{12}}{2} = 1$ pour conserver les quantités de pluie.

De la même manière, au niveau 2, les intensités I_{11}, I_{12} sont réparties en deux pour obtenir quatre pas de temps de longueur $\Delta t_2 = \frac{T_0}{4}$. Les intensités correspondantes sont alors :

$$I_{21} = \eta_{21}\eta_{11} I(T_0) \tag{5.17}$$
$$I_{22} = \eta_{22}\eta_{11} I(T_0) \tag{5.18}$$
$$I_{23} = \eta_{23}\eta_{12} I(T_0) \tag{5.19}$$
$$I_{24} = \eta_{24}\eta_{12} I(T_0) \tag{5.20}$$
$$\tag{5.21}$$

De même et avec des notations évidentes, $\eta_{i,j}, \eta_{i,j+1} \geq 0$ et $\frac{\eta_{i,j}+\eta_{i,j+1}}{2} = 1$ pour conserver les quantités de pluie. Dans le modèle de cascade multiplicative, $\eta_{i,j}$ est une réalisation d'une variable aléatoire η, appelée générateur de la cascade.

Si on continue la désagrégation jusqu'au niveau n, nous obtenons 2^n pas de temps de longueur $\Delta t_n = 2^{-n} T_0$ avec des intensités correspondantes de :

$$I_{ij}(\Delta t_n) = I(T_0) \prod_{k=1}^{n} \eta_k \tag{5.22}$$

– i est le niveau de la cascade,

– j est la position du pas de temps dans le niveau considéré,
– η_k est une réalisation de la variable aléatoire η.

Si on se place dans un cas plus général où les intensités ne sont plus réparties en deux à chaque niveau mais en b pas de temps, l'expression du pas de temps au niveau n est donnée par

$$\Delta t_n = b^{-n} T_0. \tag{5.23}$$

Pour simplifier, on reste dans le cas $b = 2$. Dans certains modèles de cascade multiplicative, on impose la contrainte de conservation de la masse : $\frac{\eta_{i,j}+\eta_{i,j+1}}{2} = 1$ (la cascade est dite 'micro-canonique'). Dans ce cas, la distribution de η est celle d'une partition. Dans les autres modèles, on impose seulement $E(\eta) = 1$ avec des variables η i.i.d. (la cascade est dite 'canonique'). Les modèles canoniques sont les modèles de cascades multiplicatives les plus utilisés.

Le réalisme des séries ainsi générées dépend de la pertinence du générateur η. Les lois utilisées pour η peuvent être discrètes ou continues. Mouhous (2003) donne quelques exemples et considérations sur le choix de la loi de η. Par exemple, l'intermittence de la pluie est caractérisée par une alternance de durées sèches et pluvieuses. Cette intermittence est respectée par le modèle, si η a une probabilité non nulle de prendre la valeur 0. La variable aléatoire η peut donc suivre une loi semi-continue : elle possède une mesure non nulle en 0 (parfois aussi en 1), et une distribution continue entre 0 et 1.

Les modèles de cascade multiplicative vérifient certaines propriétés statistiques, que nous présentons très succinctement. Ces résultats sont issus de la théorie des fractales et multifractales présentée avec plus de détails dans (Hubert *et al.*, 1993), (de Lima et Grasman, 1999), (Schertzer et Lovejoy, 1987), (Tessier *et al.*, 1993), (Pandey *et al.*, 1998), (Lavallee *et al.*, 1991), (Tessier *et al.*, 1996).

Les moments d'ordre q des séries générées par cascade multiplicative s'écrivent comme une fonction puissance du pas de temps de désagrégation Δt_n :

$$E[I(\Delta t_n)^q] = I(T_0)^q \Delta t_n^{-K(q)} T_0^{\log_b E(\eta^q)} \tag{5.24}$$

$K(q) = \log_b E(\eta^q)$ est appelée fonction exposant d'échelle, c'est une fonction linéaire et convexe.

Cette propriété s'écrit plus simplement, avec $\lambda = \frac{T_0}{\Delta t_n}$, et $I_\lambda = \frac{I(\Delta t_n)}{I(T_0)}$:

$$E[I_\lambda^q] \propto \lambda^{K(q)} \tag{5.25}$$

Les processus dont les séries ont des moments d'ordre q qui s'écrivent comme une fonction puissance du pas de temps sont dits invariants d'échelle. Les propriétés d'invariance d'échelle et la validité des modèles de cascade pour la pluie ont été vérifiées par Olsson *et al.* (1993), Hubert *et al.* (1993), Olsson (1995), Svensson *et al.* (1996), Tessier *et al.* (1996), Venugopal et Foufoula-Geoorgiou (1996), Veneziano *et al.* (1996), Olsson et Burlando (2002), Pavlopoulos et Gupta (2003), Bernardara (2004). La propriété d'invariance d'échelle a également été vérifiée par Olsson et Burlando (2002) sur des séries simulées par le modèle de Neyman-Scott.

On peut également montrer que la distribution des extrêmes est à queue lourde :

$$P(I_\lambda \geq \lambda^\gamma) \approx \lambda^{-C(\gamma)}, \tag{5.26}$$

pour $\gamma > 0$ et $C(\gamma)$, appelée fonction de codimension fractale, est une fonction convexe et non linéaire.

Guntner *et al.* (2001) présentent un modèle de cascade multiplicative micro-canonique. Pour désagréger une intensité (ou un cumul) $I(T)$, ils utilisent une loi η conditionnelle à la position de T (en début, fin, ou milieu d'événement pluvieux ou encore si T est un intervalle pluvieux isolé). L'utilité de ce conditionnement a déjà été démontré dans le cas des pluies par Buishand (1978), et appliqué dans le cas des modèles de cascade multiplicative par Olsson (1998). Dans chaque classe de conditionnement, la loi η choisie par Guntner *et al.* (2001) est la loi empirique estimée sur les données observées (à un pas de temps de 1 h). Pour cela, Guntner *et al.* (2001) agrègent des données mesurées à des pas de temps 1 h, en des données de pas de temps 32 h.

Guntner *et al.* (2001) montrent que les hypothèses d'invariance d'échelle sont vérifiées sur les climats semi-aride du Brésil et tempéré du Royaume-Uni, entre les pas de temps 1 h et 32 h. De plus, le modèle reproduit les caractéristiques de la pluie en 1 h avec une bonne précision dans les deux climats (moyennes et variances du volume de la pluie en 1 h, du volume des événements, de la durée des événements, et de la durée sèche entre événements). Cependant, en climat tempéré, les événements extrêmes sont sur-estimés. Du fait des précipitations frontales sur le Royaume-Uni, les forts volumes sont accumulés sur de longues durées : à une fine résolution, les volumes sont bien moins forts. Le modèle ne paraît pas capable de reproduire cette désagrégation de volumes forts sur un pas de temps grossier à des volumes beaucoup plus atténués sur des pas de temps fins. En revanche, le modèle reproduit bien les maxima et le nombre de dépassements d'un seuil élevé en climat semi-aride : du fait de la forte activité convective des précipitations au Brésil, les forts volumes sont principalement constitués d'événements pluvieux courts et de fortes intensités.

Molnar et Burlando (2005) comparent un modèle canonique et un modèle micro-canonique pour désagréger des pluies de 1280 minutes (soit 21.3 h) en pluies de 10 minutes. Ils montrent que les deux modèles reproduisent bien l'intermittence de la pluie au pas de temps 10 minutes. Le modèle micro-canonique préserve bien la moyenne des maxima annuels, mais sous-estime leur variance. En revanche, la variance est bien reproduite par le modèle canonique pour le pas de temps 10 minutes. Ces différences entre le modèle micro-canonique et le modèle canonique montrent l'influence de la loi du générateur η de la cascade : l'hypothèse de conservation de la masse en moyenne (dans le modèle canonique), et non pas de manière exacte (dans le modèle micro-canonique) laisse un degré de liberté supplémentaire dans la génération des pluies. Enfin, en terme de pluies extrêmes, les modèles micro-canonique et canonique préservent la moyenne des maxima annuels observés, mais le modèle micro-canonique sous-estime sa variabilité. En revanche, la variabilité des observations est respectée par le modèle canonique, mais les résultats du modèle canonique se détériorent sur les longues durées : les extrêmes sont sur-estimés. Ce résultat est peut-être à rapprocher du résultat de Guntner *et al.* (2001) sur le climat tempéré car le cas d'étude de Molnar et Burlando (2005) est situé à Zürich, en Suisse.

Ces modèles de cascades sont également utilisés pour désagréger des événements. Menabde et Sivapalan (2000) simulent une succession d'événements, via une loi empirique pour les durées sèches, et des distributions Levy-stables à queue lourde (indépendantes) pour les durées et les intensités moyennes des événements[3]. Ils utilisent ensuite un modèle de cascade

[3]La distribution Levy-stable possède quatre paramètres ($\alpha \in (0, 2)$ décrit le comportement de la queue

pour passer de cette chronologie à une série au pas de temps 5 minutes. Ils trouvent que ces distributions s'ajustent mieux à la queue de distribution des durées et des quantités de pluies que les modèles classiques avec une loi gamma pour l'intensité des événements.

Modèles de désagrégation événementiels par profils d'averses

De même que certains modèles journaliers, les modèles de désagrégation à profil d'averses utilisent des distributions indépendantes ou jointes pour caractériser un événement en terme de son intensité moyenne, sa durée, et la durée entre deux événements. A la différence des modèles journaliers, le cumul total apporté par un événement est ensuite désagrégé en un hyétogramme à un pas de temps donné, selon un profil ou une courbe de répartition de la masse aléatoire. La loi des intensités est souvent conditionnelle à la durée de l'événement. La figure 5.5 illustre la méthode de désagrégation. Des exemples de ces modèles sont donnés par Eagleson (1972), Acreman (1990), Heneker *et al.* (2001).

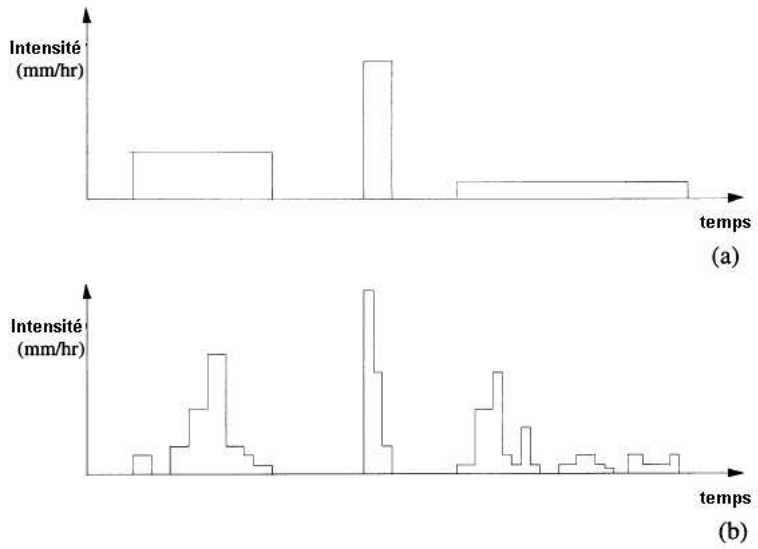


Fig. 5.5: Modèle de désagrégation. Graphique du dessus : dans une première étape, des événements sont simulés (par leurs durées, et leurs intensités moyennes). Graphique du bas : les événements sont désagrégés en hyéogrammes, selon un profil aléatoire.

Le modèle d'Eagleson (1972), modifié ensuite par Hebson et Wood (1982), Diaz-Granados *et al.* (1984), Moughamian *et al.* (1987), Beven (1987), Blazkova et Beven (1997), génère l'intensité moyenne, la durée des événements, et la durée entre événements par trois lois exponentielles indépendantes. Pour une saison donnée, le modèle ne nécessite donc que trois paramètres, estimés par la moyenne des trois variables. Durant la simulation d'un événement, un profil d'averse est tiré au hasard parmi tous les profils d'averse de la série horaire observée[4],

de distribution, $\sigma \in \mathbb{R}+^*$ le paramètre d'échelle, $\beta \in [-1, 1]$ le paramètre de forme, $\mu \in \mathbb{R}$ le paramètre de position). Il n'existe pas de forme analytique explicite pour la densité de la distribution Levy-stable, cependant le comportement asymptotique est décrit par $f(x) \propto x^{-\alpha-1}$ pour $x \to \infty, \alpha < 2$. Le cas $\alpha = 2$ est dégénéré, et la loi Levy-stable coïncide avec une loi gaussienne.

[4]Le profil d'une averse observée est défini par la courbe de l'intensité en fonction du temps, pendant la

et dans la saison donnée. Acreman (1990) et Blazkova et Beven (1997) montrent sur des cas d'études que ces modèles reproduisent bien les caractéristiques des durées entre événements, et ce à différents pas de temps. En particulier, ils montrent que ces modèles reproduisent bien les extrêmes.

Contrairement aux cas d'études de Acreman (1990) et Blazkova et Beven (1997), Cameron *et al.* (2000a) montrent sur un autre cas d'étude que le modèle d'Eagleson (1972) sous-estime les quantiles des pluies maximales en 1 heure et en 24 heures (pour les périodes de retour de 1 à 100 ans). Pour 120 simulations, les intervalles de confiance à 95% des quantiles simulés ne contiennent pas les observations. Ce mauvais résultat peut être dû à l'hypothèse d'indépendance entre la durée et l'intensité d'un événement, et/ou au choix de la loi des intensités.

Cameron *et al.* (1999) proposent un autre modèle, fondé sur des lois empiriques. Comme le modèle précédent, ce modèle simule des événements par leurs durées, leurs intensités moyennes, les durées entre événements, et les profils des averses, et le modèle est défini sur deux saisons. Les temps d'inter-arrivée entre les événements sont supposés indépendants de l'intensité et de la durée des événements et sont simulés à partir de leur loi empirique, évaluée sur les données disponibles. La durée est également simulée à partir de la distribution empirique des durées. Elle est cependant extrapolée par une loi de Pareto généralisée, ajustée à 5% des valeurs les plus fortes de l'échantillon des durées, tout en étant bornée supérieurement par des records. La loi de Pareto généralisée permet de simuler des événements avec des durées plus longues que les durées observées.

Contrairement au modèle précédent, les intensités moyennes des événements dépendent des durées. Cette dépendance est modélisée grâce à la classification des événements en sept classes de durées (1 h, 2-3 h, 4-10 h, 11-16 h, 17-27 h, 28-62 h, $\geq$ 63 h). Dans chaque classe, les intensités sont simulées suivant leur loi empirique. Si le nombre d'événements de la classe est suffisant, la queue de distribution des intensités est extrapolée par une loi GPD, bornée supérieurement si nécessaire.

Enfin, les profils d'averses observés sont triés par saisons, classes de durées, et type d'événement (intensité extrême ou non). Ils sont normalisés et tirés au hasard, conditionnellement à la saison, la durée, et le caractère extrême ou non de l'intensité.

Cameron *et al.* (2000a) annalysent les quantiles simulés par ce modèle. Selon Cameron *et al.* (2000a), les maxima annuels observés sont à l'intérieur des limites de l'intervalle de confiance à 95% des quantiles des pluies maximales en 1 heure et en 24 heures (de périodes de retour 1 à 100 ans), calculés avec 120 simulations et représentés sur un diagramme de Gumbel. L'allure de la médiane des quantiles simulés est légèrement convexe pour les pluies maximales en 1 heure, et convexe pour les pluies maximales en 24 heures. Cependant, la médiane des quantiles en 1 h simulés est inférieure à la borne inférieure de l'intervalle de confiance des quantiles en 1 h observés. Selon Cameron *et al.* (2000a), cette sous-estimation vient du fait qu'environ 25% des pluies maximales en 1 h sont associées à des évenements de durées longues (supérieures ou égales à 5 h). Pour reproduire des valeurs extrêmes de 1 h, dans un événement de longue durée, il faut un profil d'averse particulier. Or ces profils d'averses forment une très petite partie des profils d'averse des événements de longues durées. En revanche, le modèle reproduit bien les valeurs extrêmes en 1 h, pour des événements de petites durées (1 h-3 h).

durée de l'averse. Puisque les averses ont des durées différentes, les profils d'averses sont normalisés par un terme d'échelle en intensité et en durée, avant d'être utilisés dans le modèle.

Pour améliorer le modèle, Cameron *et al.* (2000a) proposent différentes méthodes : utiliser d'autres classes de durées, et générer des profils d'averse capables d'augmenter la fréquence des profils générateurs d'averses extrêmes de 1 h dans des événements de longue durée.

En conclusion, les lois empiriques extrapolées en queue de distribution par une loi appropriée (GPD), et la prise en compte de différents types d'événements, par une classification suivant la durée des événements et l'intensité extrême ou non de l'événement, ont amélioré les résultats du modèle d'Eagleson (1972). En effet, les événements orage et dépression sont de nature différentes, et les pluies qu'ils génèrent appartiennent à des populations différentes. La modélisation de Cameron *et al.* (1999) permet de le prendre en compte.

Concernant les statistiques de la distribution centrale des pluies simulées (moyenne et variance des pluies en 1 h et 24 h, coefficient d'auto-corrélation, proportion de périodes sèches, moyenne et variance des durées sèches et pluvieuses, nombre moyen d'événements par saison), les deux modèles d'Eagleson (1972) et Cameron *et al.* (1999) donnent d'assez bons résultats sur des données du Royaume-Uni, selon Cameron *et al.* (2000a). On constate une légère sous-estimation de la variance des pluies en 1 h. Le modèle d'Eagleson (1972) sous-estime les durées sèches (moyenne, variance et proportion).

Nous présentons enfin un modèle largement utilisé en Australie : le modèle de Heneker *et al.* (2001). Les données d'entrée de ce modèle sont les événements, calculés sur des données de 6 minutes. Ces données d'événements permettent au modèle d'être peu sensible aux valeurs manquantes dans la série de pas de temps 6 minutes.

Le modèle fonctionne en deux étapes :

- il simule d'abord une série d'événements, par leurs durées, leurs intensités moyennes, et les durées sèches entre événements. Les durées entre événements et les durées d'événements sont simulées par une loi exponentielle généralisée de paramètres $\theta_1, \theta_2, \theta_3, \theta_4$ et de fonction de répartition :

$$F(x) = 1 - \exp(-\theta_1 x^{\theta_2})(1 - \theta_3 x/\theta_4)^{1/\theta_3} \tag{5.27}$$

  avec $\theta_3 < 0, \theta_1, \theta_2, \theta_4 > 0$. L'avantage de cette loi par rapport à la loi exponentielle est sa plus grande flexibilité et sa capacité à reproduire les extrêmes. On remarque qu'elle généralise la loi GPD.
  Les intensités moyennes des événements sont simulées par une loi GPD. Les paramètres de la GPD tiennent compte de la relation entre l'intensité et la durée de l'événement.
- La deuxième étape est l'étape de désagrégation des événements, elle répartit la quantité de pluie dans l'événement. La méthode de désagrégation est fondée sur une marche aléatoire qui génère une courbe adimensionnelle du profil du cumul (rendu adimensionnel par normalisation) pendant la durée d'un événement (rendue adimensionnelle par normalisation). La méthode, reprise de Woolhiser et Osbortn (1985), inclue des périodes sèches dans les événements.

Ce modèle utilise 54 paramètres, estimés par maximum de vraisemblance.

Heneker *et al.* (2001) ont montré que le modèle reproduit bien les caractéristiques temporelles internes aux événements. Frost *et al.* (2005) ont montré que ce modèle reproduit bien les courbes Intensité-Durée-Fréquence de postes d'Australie. Cependant, le modèle de Heneker *et al.* (2001) sur-estime fortement la queue de distribution des pluies de durée 10 minutes. Frost *et al.* (2005) ont également montré que le modèle de Heneker *et al.* (2001) donnait de meilleurs résultats que le modèle de Neyman-Scott de Cowpertwait *et al.* (2002)

sur les petits pas de temps (1 h-6 h). Cette meilleure performance du modèle de Heneker *et al.* (2001) semble due à la modélisation explicite des durées d'événements et des durées entre événements, selon Frost *et al.* (2005).

Ce dernier exemple montre encore que l'introduction de distributions à queues lourdes, et la prise en compte de la dépendance entre durées et intensités donnent de bons résultats.

Néanmoins, comme déjà constaté pour de nombreux modèles appliqués sur des postes australiens, la variabilité inter-annuelle des pluie est sous-estimée. En effet, le modèle ne prend pas en compte la variabilité inter-annuelle, mais seulement la variabilité saisonnière.

5.4 Conclusions

Ce chapitre a cherché à présenter les différentes classes de modèles stochastiques de simulation temporelle de la pluie. On distingue deux grandes classes de modèles : les générateurs de pluie journalière, et les générateurs de pluie à des pas de temps plus fins que 24 heures.

Les générateurs de pluie journalière sont essentiellement composés de deux générateurs. Le premier générateur simule les occurrences, et peut être un processus markovien ou un processus d'alternance entre durées sèches et pluvieuses. Le second générateur simule les quantités de pluie précipitées dans les événements simulés par le premier générateur. Les lois utilisées pour les générations des cumuls ou des intensités des événements sont souvent de type gamma ou exponentielle. Récemment, des distributions à queues lourdes ont été utilisées.

Des références bibliographiques montrent que ces modèles sous-estiment souvent la fréquence des événements extrêmes et la variabilité inter-annuelles des précipitations.

Ces modèles journaliers existent en versions paramétriques et non-paramétriques.

Pour la modélisation des pluies à des pas de temps inférieurs à la journée, de nombreux générateurs différents existent. En particulier, nous avons présenté le modèle Shypre de génération de hyétogrammes par simulation d'événements, les modèles de type Bartlett-Lewis et Neyman-Scott, fondés sur des processus poissonniens simulant les occurrences des événements pluvieux, et enfin les modèles de désagrégation. Les modèles de désagrégation sont eux-même de différents types : on retrouve des modèles Bartlett-Lewis et Neyman-Scott, des modèles de cascades multiplicatives empruntées à la théorie des fractales, et enfin des modèles fondés sur les profils d'averses.

Différents critères statistiques ou visuels de comparaison de distributions ou de moments entre les quantiles observés et simulés permettent de juger des aptitudes des modèles. Dans l'ensemble, les analyses référencées dans la bibliographie montrent que la validité des modèles dépend du climat de la région étudiée. Les modèles prenant en compte les différents types de pluie donnent de meilleurs résultats. Différentes classifications de la pluie existent :

- dans le cas des pluies journalières, les pluies peuvent être classées en fonction de leur place dans l'événement pluvieux (événement de durée un jour, ou pluie en début ou fin d'événement, ou pluie en milieu d'événement),
- en fonction de l'intensité de pluie de l'événement pluvieux,
- en fonction de la durée de l'événement pluvieux.

Enfin, dans l'exemple de Shypre, la modélisation empirique de la dépendance des valeurs extrêmes d'un même événement et la modélisation de la dépendance entre certaines variables du modèle ont permis à Shypre de respecter les valeurs extrêmes.

Dans le chapitre suivant, nous nous intéressons plus particulièrement au modèle Shypre, dont nous analysons les incertitudes et la capacité à respecter les valeurs extrêmes.

Chapitre 6

Calculs d'incertitudes du modèle Shypre journalisé et de la loi GPD

Après avoir présenté les différents générateurs de pluie, nous analysons plus particulièrement les incertitudes et la capacité à reproduire les valeurs extrêmes du modèle Shypre journalisé. Le modèle Shypre journalisé simule des chroniques de pluies horaires. A partir de ces chroniques horaires, on extrait la distribution des pluies maximales ou supérieures à un seuil donné, pour différents pas de temps (journalier et durées supérieures ou égales à 1 heure). Ainsi, lorsque nous parlons de quantiles simulés ou estimés par Shypre, il s'agit de quantiles déduits de la simulation des chroniques de pluies horaires.

Bien qu'une simulation par le modèle de pluies conduise à disposer de l'ensemble des pluies de différentes durées, on ne s'intéressera qu'à la restitution des pluies journalières, pour simplifier l'étude. Dans ce chapitre, nous comparons les incertitudes des quantiles estimés par Shypre avec l'incertitude des quantiles estimés par une méthode classique issue de la théorie des valeurs extrêmes. Plus particulièrement, nous analysons les incertitudes dues aux données d'entrée et aux estimateurs des paramètres des modèles.

Nous présentons d'abord le modèle Shypre journalisé et les données du cas d'étude. Nous étudions ensuite l'incertitude des données d'entrée seules, à travers une analyse de la sensibilité à l'échantillonnage. L'évaluation de la sensibilité à l'échantillonnage est particulièrement importante pour évaluer la robustesse des estimations vis à vis de la taille et des valeurs extrêmes de l'échantillon. Après cette première analyse, nous évaluons enfin l'effet de l'incertitude des paramètres sur les quantiles estimés.

6.1 Présentation du modèle Shypre journalisé et des données

6.1.1 Le modèle Shypre horaire journalisé

Le modèle Shypre journalisé est identique au modèle Shypre horaire. La différence est dans l'estimation des paramètres. Dans le modèle journalisé, la paramétrisation du modèle de pluies horaires est réalisée par la seule information disponible en journalier. Cette information journalière permet de déterminer, par des relations linéaires, certains paramètres principaux du modèle de pluie. Les autres paramètres sont fixés régionalement, car ils varient peu ou car le modèle est peu sensible à leurs variations.

Les paramètres utilisés dans la version journalisée de Shypre sont les valeurs moyennes $ne, pjx, dtot$ des variables suivantes :

- NE : le nombre d'événements par an, ou par saison,
- PJX : la pluie journalière maximale de l'événement (en mm),
- $DTOT$: la durée totale de l'événement (en jours).

Le modèle génère de longues séries d'événements indépendants (par exemple 1000 séries de 1000 ans), et calcule ensuite la moyenne des distributions empiriques simulées[1]. En sortie, le modèle donne les distributions des cumuls annuels maximum et des dépassements du seuil 20 mm[2], pour différents durées entre 1 h et 72 h. Le modèle est paramétré sur deux saisons : l'été entre Juin et Novembre, et l'hiver entre Décembre et Mai.

Nous analyserons ici l'incertitude des quantiles simulés de pluie journalière supérieurs au seuil 20 mm.

D'après la théorie des valeurs extrêmes, et puisque nous considérons des pluies supérieures à un seuil, nous comparons les résultats de Shypre avec ceux d'une loi GPD, ajustée sur les dépassements du seuil. La fonction de répartition de la loi GPD est notée $F(x) = 1 - (1 - k(x - 20)/\alpha)^{1/k}$, avec x la pluie journalière supérieure à 20 mm, exprimée en mm. En fait, x est la pluie journalière maximale d'un événement, la moyenne des valeurs x observées est donc égale à pjx.

6.1.2 Les données journalières de Marseille

Notre étude est réalisée sur la longue série de Marseille, présentée au chapitre 4 (122 années de mesures, entre 1882 et 2003). À Marseille, la saison des plus fortes pluie est l'été. Nous basons donc notre étude sur cette saison, et pour se placer dans les mêmes conditions que Shypre, nous considérons que l'été est défini par les mois de Juin à Novembre.

On reprend la définition des événements, selon le modèle Shypre. La définition des événements est donnée à la section 5.3.1. Puisque les événements sont séparés par un minimum d'un jour avec moins de 4 mm, les événements sont supposés indépendants. La série de Marseille contient 444 événements. Le seuil au dessus duquel les valeurs sont ajustées à une loi GPD est choisi à partir de la méthode de Coles (2001), expliquée au premier chapitre. D'après la figure 6.1, on constate que le seuil 20 mm est acceptable. Ainsi, on sélectionne le même seuil que le modèle Shypre dans sa définition des événements. Les données de Marseille supérieures à 20 mm sont présentées sur la figure 6.2, ainsi que l'estimation des quantiles par Shypre et par ajustement d'une loi GPD. On constate la plus forte valeur de la série (200 mm) observée en 2000. Là encore, nous avons vérifié la stationnarité au second ordre (ou plus précisément l'absence de tendance linéaire du paramètre d'échelle) de la série des dépassements de seuil à l'aide d'un test de rapport de vraisemblance. Le test a comparé une loi GPD 'stationnaire' (c'est-à-dire à paramètres d'échelle et de forme fixes), avec une loi GPD à tendance linéaire sur le paramètre d'échelle. Le modèle à tendance linéaire n'est pas significativement meilleur que le modèle 'stationnaire', à un niveau de risque supérieur à 25%. Cela confirme les précédents résultats de stationnarité de la série de Marseille, présentés au chapitre 4.

[1] pour chaque fréquence, le quantile correspondant est estimé par la moyenne des quantiles empiriques des 1000 simulations.

[2] le seuil 20 mm correspond à la valeur minimale de pluie journalière pour identifier un événement.

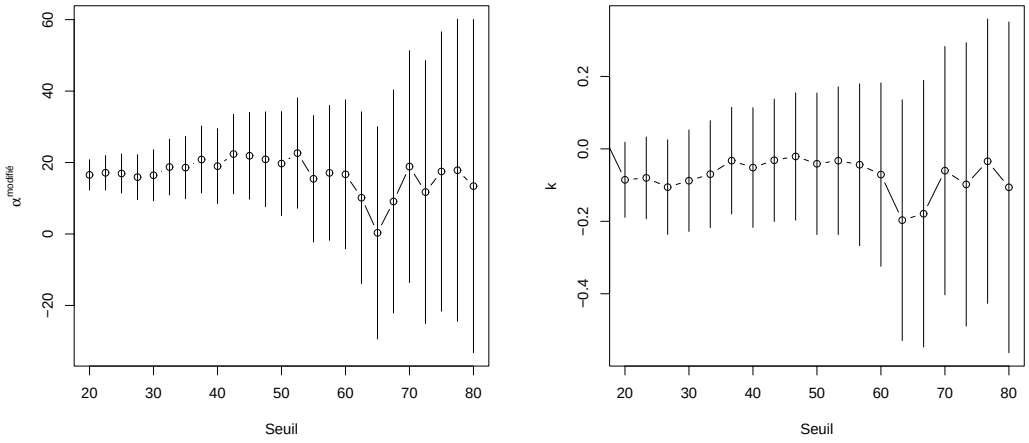


Fig. 6.1: Estimation des paramètres en fonction du seuil pour les pluies journalières de Marseille, avec intervalle de confiance à 95%. Les graphes ont été tracés avec le package POT proposé par Ribatet (2006) sous le logiciel libre R. $\alpha^{modifié}(u) = \alpha(u) + k(u)u$ où $\alpha(u), k(u)$ sont les paramètres d'échelle et de forme de la loi GPD au dessus du seuil u.

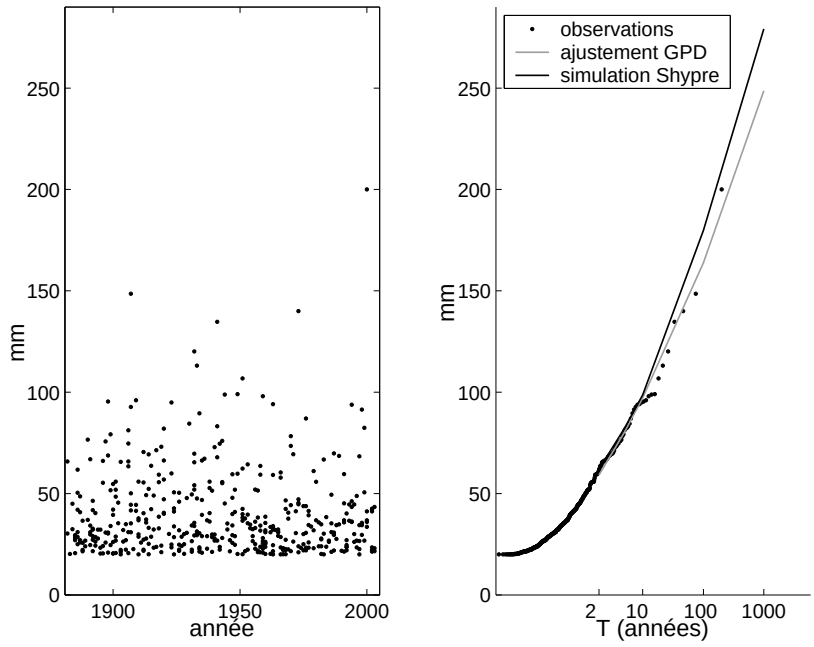


Fig. 6.2: Série des pluies journalières dépassant le seuil 20 mm à Marseille, pour la saison Juin-Novembre. À gauche : série temporelle entre 1882 et 2003 ; à droite : ajustement des valeurs journalières supérieures au seuil 20 mm par une loi GPD, et distribution simulée par Shypre.

6.2 Analyse des incertitudes d'échantillonnage

Nous nous intéressons à la sensibilité des modèles (Shypre et GPD) à l'échantillonnage. Plus particulièrement, nous étudions la sensibilité des modèles à la présence plus ou moins importante de valeurs extrêmes dans l'échantillon. Pour cela, nous considérons l'incertitude des estimateurs des paramètres et des quantiles. Les quantiles considérés ici sont les quantiles de pluie journalière supérieure à 20 mm.

6.2.1 Méthodologie

Cas du générateur Shypre

La paramétrisation de Shypre est fondée sur des événements. Si le nombre d'événements est grand, l'estimation des paramètres sera précise. Par exemple, dans le cas de Marseille, la série est longue de 122 années, et possède 444 événements (donc 444 dépassements de seuils supposés indépendants). Les moyennes ne, pjx, $dtot$ sont respectivement calculées sur 122, 444 et 444 valeurs. Nous considérons que les distributions d'échantillonnage marginales des paramètres sont des lois gaussiennes. Sur la série de Marseille, la corrélation entre les paramètres pjx et $dtot$ est 0.22, et elle est inférieure à 0.05 entre ne et les autres paramètres. La distribution jointe des trois paramètres est donc modélisée par une loi gaussienne en trois dimensions, de marginales indépendantes, centrées sur les estimateurs des paramètres et de variances égales aux variances d'échantillonnage des paramètres. Un grand nombre de paramètres est ensuite simulé dans cette distribution gaussienne, ainsi que les quantiles correspondants. La médiane et les intervalles de confiance des quantiles sont ensuite déduits.

Le point de vue adopté ici pour analyser les incertitudes d'échantillonnage est fréquentiel.

Cas de la loi GPD

Par souci de cohérence avec l'analyse des incertitudes de Shypre, nous adoptons ici une analyse fréquentielle des incertitudes des estimations de loi GPD. Nous utilisons la méthode des moments pondérés, d'après une analyse comparative des estimateurs des moments pondérés, des moments et du maximum de vraisemblance, réalisée par Hosking et Wallis (1987) et mentionnée au chapitre 2 de la thèse. Hosking et Wallis (1987) montrent en effet que la méthode des moments pondérés donne de meilleurs résultats lorsque le paramètre de forme est négatif, et proche de -0.2. Or dans notre cas, différentes estimations de k (par maximum de vraisemblance et moments pondérés) montrent que k est négatif et de l'ordre de -0.1. D'après Hosking et Wallis (1987), la distribution asymptotique des estimateurs des moments pondérés des paramètres est normale de matrice de variance-covariance

$$var \begin{pmatrix} \widehat{\alpha} \\ \widehat{k} \end{pmatrix} \tag{6.1}$$

approchée par :

$$\frac{1}{n(1+2k)(3+2k)} \begin{pmatrix} \alpha^2(7+18k+11k^2+2k^3) & \alpha(2+k)(2+6k+7k^2+2k^3) \\ \alpha(2+k)(2+6k+7k^2+2k^3) & (1+k)(2+k)^2(1+k+2k^2) \end{pmatrix} \tag{6.2}$$

où n est le nombre de dépassements utilisés pour l'estimation. La simulation de la loi des estimateurs $\widehat{\alpha}, \widehat{k}$ permet d'en déduire la distribution des estimateurs des quantiles. Nous utilisons la procédure suivante de simulation des estimateurs des paramètres : soient $\widehat{\alpha}_0, \widehat{k}_0$ les valeurs des estimateurs de α, k par les moments pondérés, alors $\widehat{\alpha}, \widehat{k}$ sont simulés via la relation suivante :

$$\begin{aligned} \widehat{\alpha} &= \widehat{\alpha}_0 + \sqrt{var(\widehat{\alpha})} Z_1 \\ \widehat{k} &= \widehat{k}_0 + \sqrt{var(\widehat{k})} (\rho Z_1 + (1-\rho^2)^{1/2} Z_2) \end{aligned} \tag{6.3}$$

avec Z_1, Z_2 des lois normales centrées réduites indépendantes et $\rho = \frac{Cov(\widehat{\alpha},\widehat{k})}{\sqrt{var(\widehat{k})}\sqrt{var(\widehat{\alpha})}}$. L'intérêt de cette méthode de simulation est d'éviter le calcul parfois délicat de la racine carrée de la matrice de variance-covariance des estimateurs des paramètres.

Avant de poursuivre l'analyse des incertitudes d'échantillonnage de la loi GPD, nous présentons une courte comparaison des intervalles de confiance obtenus par la méthode du maximum de vraisemblance (la matrice de variance-covariance des estimateurs s'exprime à l'aide de la matrice d'information de Fisher) et par la méthode des moments pondérés (de matrice de variance-covariance des paramètres donnée par 6.2). Cette étude a déjà été menée par Hosking et Wallis (1987) dans le cas d'estimation des paramètres par maximum de vraisemblance via une méthode d'optimisation fondée sur l'algorithme de Newton-Raphson. D'autres méthodes existent, en particulier la méthode du simplexe (Press *et al.*, 1987). Nous avons vérifié le résultat de Hosking et Wallis (1987) en considérant l'algorithme d'optimisation du simplexe au lieu de celui de Newton-Raphson, dans l'estimation par maximum de vraisemblance. Nous avons donc comparé les deux méthodes d'estimation (maximum de vraisemblance via la méthode d'optimisation du simplexe, et moments pondérés) sur un échantillon de taille 78 (la valeur 78 est la taille d'un échantillon d'une série de 20 ans étudiée à la section 6.2.2) de loi GPD de paramètres d'échelle $\alpha = 1$ et de forme $k = -0.2, -0.1, 0, 0.1$ (ces valeurs de k sont celles rencontrées en pratique lors d'estimations des échantillons observés à la section 6.2.2). Les méthodes d'estimation étant invariantes par changement d'échelle des données, le choix de α n'affecte pas les résultats. Le tableau 6.1 montre que même avec la méthode du simplexe, les intervalles de confiance à 90% des quantiles de fréquences 0.9, 0.99 et 0.999, calculés par la méthode du maximum de vraisemblance, ne recouvrent pas mieux les vraies valeurs des quantiles que la méthode des moments pondérés.

	ML			PWM		
	F=0.9	F=0.99	F=0.999	F=0.9	F=0.99	F=0.999
k=-0.2	87.7	84.5	84.5	86.7	86.7	87.4
$k = -0.1$	87.5	84.1	84.2	88.3	90	92.1
$k = 0$	87.6	82.9	83	88.2	91.1	92.7
$k = 0.1$	87.6	82.3	82.4	87.8	91.5	92.5

Tab. 6.1: Taux de recouvrement (en %) des vraies valeurs des quantiles par les intervalles de confiance à 90% estimés avec la méthode de maximum de vraisemblance (ML) et avec la méthode des moments pondérés (PWM), pour un échantillon de loi GPD de taille 78.

6.2.2 Application à la série de Marseille

Analyse des incertitudes d'échantillonnage avec la série entière

Les paramètres de Shypre sont estimés par des moyennes d'un grand nombre de données. Les variances des estimateurs des paramètres sont donc très faibles, comme le montre le tableau 6.2. On constate d'autre part que le paramètre de forme k est significativement différent de 0 au niveau 90% (l'intervalle de confiance est [-0.21, -0.03]). Cela confirme les précédentes études du paramètre de forme des maxima saisonniers de la pluie journalière de la série de Marseille (voir le tableau 4.2 du chapitre 4).

Paramètre	Moyenne (Ecart-type)
ne	3.7 (0.17)
pjx (mm)	39.64 (1.03)
$dtot$ (jours)	1.72 (0.05)
α	17.3 (1.3)
k	-0.119(0.22)

Tab. 6.2: Estimations des paramètres : moyennes et écarts-types d'échantillonnage.

L'intervalle de confiance des quantiles simulés par Shypre est plus étroit que celui des estimateurs des quantiles de la distribution GPD, comme le montre la figure 6.3. En effet, les paramètres de Shypre sont estimés par des moyennes, tandis que les estimateurs des paramètres de la GPD dépendent de moments pondérés d'ordre 0 et 1 (le moment pondéré d'ordre 0 est égal à la moyenne de l'échantillon et le moment pondéré d'ordre 1 est estimé par la moyenne pondérée $\frac{1}{n}\sum_{i=1}^{n} p_{i,n}x_{i,n}$ avec $x_{i,n}$ les statistiques d'ordre de l'échantillon et $p_{i,n}$ la probabilité empirique associée à $x_{i,n}$). Ainsi, même si le modèle Shypre a trois paramètres, tandis que la loi GPD en a seulement deux, les quantiles simulés par Shypre ont une variabilité moins forte.

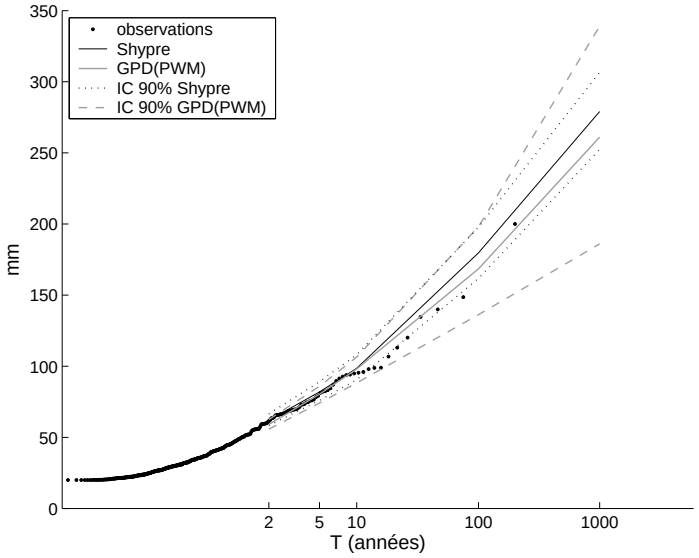


Fig. 6.3: Intervalle de confiance à 90% des quantiles de pluie journalière supérieure à 20 mm simulés par Shypre et estimés avec les moments pondérés (PWM) pour la loi GPD.

Analyse des incertitudes d'échantillonnage avec des sous-séries de 20 ans

Le plus souvent, les séries pluviométriques sont bien plus courtes que la série de Marseille, et ont une taille d'environ 20 à 30 années. En raison de l'échantillonnage, une série de 20 ans peut contenir aucune, une ou plusieurs valeurs extrêmes. Les conséquences de cette répartition aléatoire des valeurs extrêmes sont explorées ici. La longue série de 122 années est décomposée en six sous-séries de 20 ans chacune. Quelques caractéristiques des valeurs fortes des sous-séries sont présentées dans le tableau 6.3. Si les nombres moyens d'événements par

	1882-1901	1902-1921	1922-1941	1942-1961	1962-1981	1982-2001
ne	3.79(0.38)	3.74(0.41)	3.63(0.47)	3.88(0.41)	3.58(0.43)	3.53(0.33)
pjx (mm)	36.74(1.86)	44.17(2.74)	42.39(2.85)	39.03(2.33)	37.47(2.49)	39.48(3.19)
$dtot$ (jours)	1.65(0.09)	1.69(0.11)	1.78(0.11)	1.69(0.11)	1.84(0.15)	1.74(0.13)
$PJ > 100$ (mm)	aucune	148	113,120,135	107	140	200

Tab. 6.3: Estimation des moyennes et écarts-types des paramètres des six sous-séries de Marseille (le nombre de valeurs utilisées pour l'estimation des paramètres est respectivement de 20 pour ne, et environ 74 pour pjx et $dtot$). Valeurs fortes (supérieures à 100 mm) relevées dans chaque sous-série.

été (ne) sont à peu près équivalents parmi les sous-séries, les moyennes de pluies journalières maximales par événement pjx, le nombre et les valeurs des forts événements varient fortement d'une série à l'autre.

On reproduit l'étude précédente sur chacune des six sous-séries. La figure 6.4 montre les intervalles de confiance des estimateurs des paramètres de Shypre et de la GPD. Les différences les plus importantes sont observées entre les séries 1 et 2 pour Shypre et entre les séries 2 et 5 pour la GPD. On constate dans notre cas d'étude une relative stabilité des

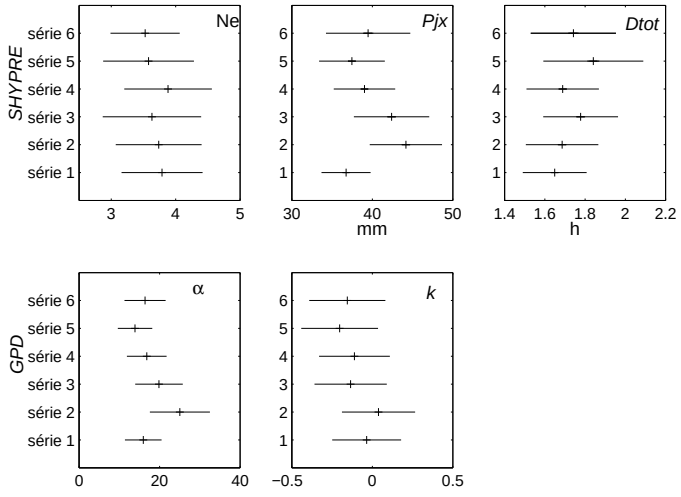


Fig. 6.4: Intervalles de confiance à 90% et médianes des estimateurs des paramètres de Shypre et de la loi GPD, pour chaque sous-série.

estimateurs du paramètre k, qui est pourtant délicat à estimer de manière générale. Avec seulement 20 ans de données, l'incertitude des estimateurs est plus grande, et la valeur 0 est cette fois incluse dans les intervalles de confiance de k : le rejet de la loi exponentielle n'est donc plus évident au niveau 10%.

La figure 6.5 montre les intervalles de confiance ainsi que les médianes des quantiles estimés par le modèle Shypre et la loi GPD.

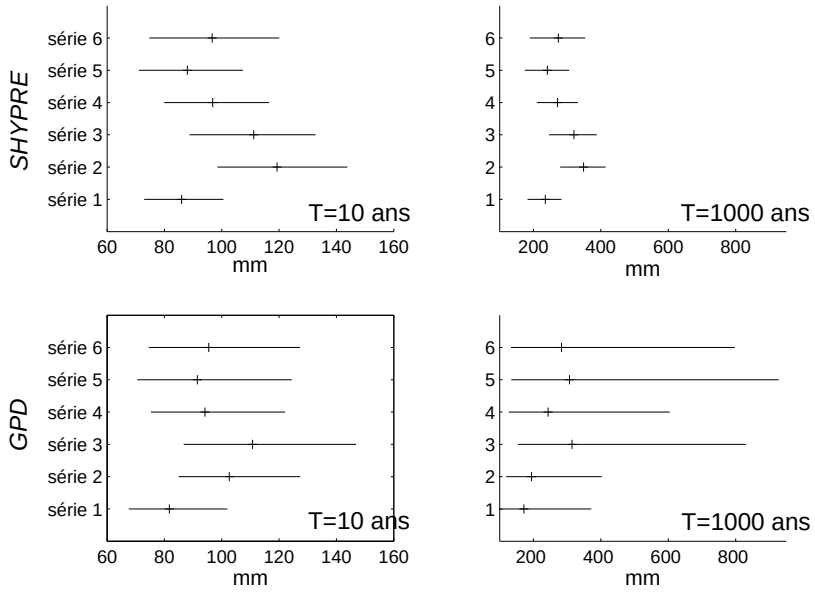


Fig. 6.5: Intervalles de confiance à 90% et médianes des estimations des quantiles de pluie journalière supérieure à 20 mm simulés par Shypre et estimés par la loi GPD, pour chaque sous-série. Les périodes de retour considérées sont 10 et 1000 ans.

Analyse des médianes des quantiles

La variabilité des quantiles, due à l'échantillonnage, est peu marquée pour la période de retour 10 ans : les différences relatives entre les médianes des quantiles estimés sur les six sous-séries sont faibles, et similaires entre les deux approches (voir la figure 6.5). Les quantiles des deux modèles se situent autour de 80 et 120 mm. Par contre, en extrapolation, les résultats des deux modèles diffèrent de manière significative. Pour la période de retour 1000 ans, les médianes des quantiles varient entre 170 et 315 mm dans le cas de la loi GPD, et entre 235 et 350 mm dans le cas du modèle Shypre. En terme de médianes des quantiles, Shypre est donc plus stable que le modèle GPD. Cette stabilité est due à la paramétrisation : Shypre est paramétré par des moyennes, tandis que les paramètres de la loi GPD dépendent des deux premiers moments pondérés.

Analyse des intervalles de confiance des quantiles

Les distributions d'échantillonnage des quantiles semblent symétriques dans le cas de Shypre, tandis qu'elles sont asymétriques dans le cas de la loi GPD. Dans le cas de Shypre, la symétrie des distributions des quantiles est due à la distribution symétrique des estimateurs des paramètres (de loi normale) et à la réponse de Shypre à une variation de ses paramètres. Dans le cas de la loi GPD, les quantiles sont très sensibles à la valeur du paramètre de forme k. Les largeurs des intervalles de confiance des quantiles millennaux de Shypre et de la loi GPD sont très différentes. La forte incertitude relevée sur les quantiles millennaux des séries

3, 5 et 6 est due à des valeurs assez fortement négatives des estimateurs du paramètre de forme k.

Les intervalles de confiance obtenus montrent d'autre part que le taux de recouvrement des vraies valeurs des quantiles de pluie par les intervalles de confiance estimés n'est pas nécessairement de 90%. Dans le cas de la loi GPD, nous remarquons que l'intervalle de confiance à 90% du quantile millennal estimé avec la troisième série est inclus dans l'intervalle de confiance à 90% du quantile millennal estimé avec la cinquième série : cela prouve que l'intervalle de confiance du quantile millennal de la cinquième série a un taux de recouvrement supérieur à celui de la troisième série. De même, sur le quantile décennal, l'intervalle de confiance estimé avec la série 4 est inclus dans l'intervalle de confiance estimé avec la série 5. Ces résultats sont à rapprocher de l'analyse préliminaire sur le taux de recouvrement des vraies valeurs de quantiles par les intervalles de confiance à 90% estimés à partir des moments pondérés (voir le tableau 6.1) : pour des paramètres de forme k entre -0.1 et 0.1, le taux de recouvrement des quantiles élevés est supérieur à 90%.

En revanche, l'étroitesse des intervalles de confiance des estimateurs des quantiles de Shypre implique qu'ils peuvent être quasiment disjoints (séries 1 et 2). Cette disjonction entre les intervalles de confiance montre que le taux de recouvrement de la vraie valeur du quantile par les intervalles de confiance obtenus est inférieur à 90%. En effet, il est impossible que l'intersection de deux ensembles de probabilité 0.9 soit presque vide. Néanmoins, nous devons nuancer ce propos. En effet, les intervalles de confiance obtenus sont des intervalles aléatoires, et chaque intervalle contient la vraie valeur du quantile avec une probabilité 90%. Les six intervalles étant indépendants, puisque construits sur des données disjointes en temps, on peut montrer à l'aide d'un raisonnement avec une loi binomiale, que la probabilité qu'au moins un intervalle ne contienne pas la vraie valeur du quantile est égale à 0.47. Il est donc probable, sur six intervalles, d'observer un intervalle ne contenant pas la vraie valeur avec la probabilité 90%. En résumé, les intervalles de confiance obtenus pour Shypre indiquent :

- la plage de variation des valeurs des quantiles simulés par le modèle (cette plage de variation semble cohérente avec les données observées, d'après la figure 6.3),
- et la stabilité de Shypre par rapport à l'échantillonnage.

6.3 Analyse des incertitudes de modélisation

6.3.1 Méthodologie

La section précédente a présenté une analyse des incertitudes d'échantillonnage des quantiles estimés par les modèles Shypre et GPD. À présent, nous analysons l'incertitude de modélisation du modèle Shypre et de la loi GPD. Nous définissons l'incertitude de modélisation par les faits suivants :

- la structure du modèle n'est pas parfaite,
- le choix des paramètres est soumis aux incertitudes sur les données d'entrée du modèle,
- des jeux de paramètres différents peuvent donner des résultats semblables (d'après la notion d'équifinalité de Beven et Binley (1992)).

Notons que Shypre n'est pas un modèle dont les paramètres sont estimés par une calibration des sorties sur les observations. Il est donc important de vérifier que le modèle reste fidèle aux observations.

Afin de quantifier l'incertitude de modélisation, nous proposons une analyse bayésienne. A la différence du point de vue fréquentiel précédent, on considère les paramètres eux-même aléatoires. Cette analyse permet d'associer une probabilité a posteriori à chaque jeu de paramètre, décrivant la capacité du modèle à être fidèle aux observations. Les observations prises en compte dans cette comparaison entre simulations et observations sont les pluies journalières supérieures à 20 mm.

Notons qu'avec un échantillon de taille 444, nous pourrions heuristiquement penser que la distribution a posteriori est beaucoup plus influencée par les données que par la loi a priori des paramètres. Dans ce cas, la distribution a posteriori des paramètres approcherait la distribution d'échantillonnage de l'estimateur du maximum de vraisemblance des paramètres. Cette hypothèse heuristique est en fait un théorème, valable sous certaines conditions de régularité, et démontré par Le Cam (1953).

En pratique, la distribution a posteriori des paramètres est simulée à l'aide d'un algorithme MCMC proposé par Renard *et al.* (2006) et présenté en annexe dans l'article de Muller *et al.* (2006).

Cas du générateur Shypre

La définition de la loi a posteriori des paramètres de Shypre nécessite une fonction de vraisemblance, et donc une loi de probabilité pour les pluies journalières supérieures à 20 mm, de paramètres $ne, pjx, dtot$. Or Shypre n'est pas une loi de probabilité. Cependant, puisque la distribution étudiée ici et simulée par Shypre est la distribution des pluies journalières supérieures à 20 mm, nous supposons que cette distribution est proche de celle d'une loi GPD. Nous proposons alors une procédure d'ajustement d'une distribution GPD sur la distribution simulée par Shypre. D'un point de vue pratique, cet ajustement a été réalisé avec une méthode des moindres carrés pondérés, par minimisation de

$$\sum_i \{F_{th}(y_i)[F_{th}(y_i) - \widehat{F}(y_i)]/[1 - F_{th}(y_i)]\}^2, \tag{6.4}$$

où F_{th} est la distribution de probabilité d'une loi GPD de paramètre α_s, k_s, $\widehat{F}(x)$ est la fréquence empirique associée à la valeur x, et $(y_i)_i$ sont les quantiles simulés[3], donnés pour différentes périodes de retour. Les paramètres à optimiser sont les paramètres α_s, k_s. Le facteur $F_{th}/(1-F_{th})$ donne un poids plus important aux valeurs de fortes périodes de retour.

L'étude des estimations des paramètres de Shypre dans un rayon de 50 km autour de Marseille a permis de définir des lois a priori des paramètres par des lois uniformes sur [3,6] pour ne, [33,45] pour pjx, [1.4,2.1] pour $dtot$.

La méthode des moindres carrés a été testée avec 10000 jeux de paramètres, échantillonnés dans leur loi a priori. Les erreurs relatives entre les quantiles simulés et les quantiles déduits de la distribution GPD ajustée sur la distribution simulée sont inférieures à 10% pour les périodes de retour 2 à 100 ans, et inférieures à 4% pour la période de retour 1000 ans. Ces résultats confirment que la distribution simulée est proche d'une loi GPD. Un exemple d'ajustement est montré sur la figure 6.6.

[3]En sortie de Shypre, on dispose de 35 quantiles dont 9 ont des périodes de retour supérieures à 100 ans et 21 ont des périodes de retour inférieures à 10 ans. La somme 6.4 est calculée sur ces 35 quantiles.

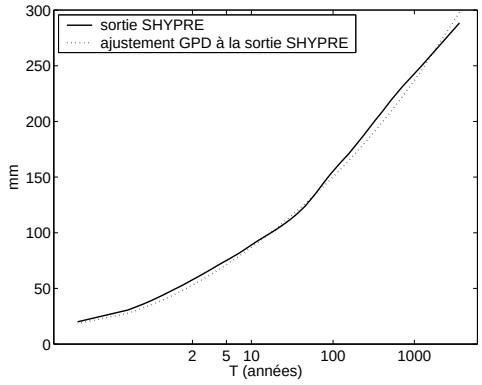


Fig. 6.6: Exemple d'ajustement d'une loi GPD sur une distribution simulée par Shypre.

Cas de la loi GPD

Les lois a priori des paramètres sont choisies larges : la loi lognormale de paramètres 0 et 100 pour le paramètre d'échelle α et uniforme sur [-1,1] pour le paramètre de forme k. Les paramètres sont simulés selon leur loi a posteriori, via un algorithme Monte Carlo de chaîne de Markov.

6.3.2 Analyse avec la série entière des observations

Dans les deux cas, Shypre et GPD, nous avons vérifié la convergence de l'algorithme MCMC par le calcul de la statistique R de Gelman *et al.* (1997) sur six séries simulées par l'algorithme de Metropolis (de 80 000 itérations), simulées en parallèle, et possédant des points de départ tirés dans la loi a priori des paramètres. Dans le cas de Shypre, R est égal à 1.000 après 40 000 itérations de l'algorithme de Metropolis. Les 40 000 itérations suivantes sont considérées comme des réalisations des paramètres selon leur loi a posteriori. Dans le cas de la loi GPD, la valeur de R est proche de 1 (1.000) après 50 000 itérations de l'algorithme de Metropolis. Les 30 000 simulations suivantes sont considérées comme simulations des paramètres selon leur loi a posteriori.

Dans le cas de la loi GPD, les intervalles de crédibilité du cas bayésien sont assez semblables aux estimations du cas fréquentiel (avec les moments pondérés), comme le montre le tableau 6.4, confirmant l'hypothèse heuristique sur le comportement asymptotique de la distribution a posteriori des quantiles de la loi GPD.

T (ans)	PWM	Méthode bayésienne
2	59.1(55.8,62.5)	59.4(56.1,63.1)
5	80.3(74.1,85.9)	80.6(74.9,88.0)
10	98.0(88.3,106.4)	98.1(89.6,110.2)
100	168.5(136.3,198.1)	167.1(140.2,213.6)
1000	261.1(186.0,339.1)	256.6(194.0,381.5)

Tab. 6.4: Estimations des quantiles et des intervalles de confiance et de crédibilité à 90%, avec les moments pondérés (PWM) et une méthode bayésienne. La valeur estimée du cas bayésien est la valeur médiane de la distribution a posteriori des quantiles.

Dans le cas de Shypre, les résultats montrent que les distributions a posteriori marginales

des paramètres *ne* et *dtot* de Shypre sont très semblables à leur loi a priori uniforme. En revanche, la distribution a posteriori de pjx est proche d'une distribution normale, de moyenne 40.3 et d'écart-type 1.01. La distribution a posteriori de pjx est donc proche de la distribution d'échantillonnage de pjx (voir tableau 6.2). Ces résultats montrent que la calibration de Shypre n'est pas sensible aux deux paramètres $ne, dtot$, et ne dépend que du paramètre pjx, qui est lui-même peu variable.

L'intervalle de crédibilité des quantiles obtenu est étroit (voir la figure 6.7), du fait de l'étroitesse de la loi a posteriori de pjx.

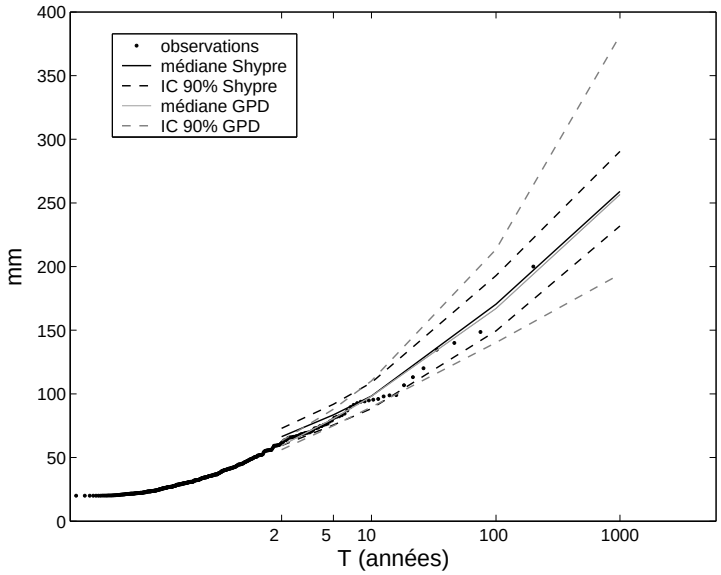


Fig. 6.7: Intervalles de crédibilité à 90% des quantiles de Shypre et de la GPD.

Malgré l'étroitesse de l'intervalle de crédibilité des quantiles du modèle Shypre, on observe que les simulations de Shypre sont fidèles aux observations :

- la médiane estimée est proche des observations, et proche de la loi GPD estimée sur les données,
- l'intervalle de crédibilité contient 443 des 444 observations,
- à partir de la période de retour 10 ans, l'intervalle de crédibilité des quantiles simulés par Shypre est inclus dans l'intervalle de crédibilité des quantiles de la loi GPD.

Enfin, la différence de taille entre les intervalles de crédibilité des quantiles des deux modèles s'explique en partie par la différence de précision des lois a priori utilisées dans les deux cas de figure. Dans le cas de la loi GPD, les lois a priori sont larges, et peu informatives, tandis que dans le cas de Shypre, les lois a priori sont plus informatives, du fait de l'interprétation physique des paramètres de Shypre.

6.3.3 Analyse avec seulement les 20 dernières années

A présent, nous nous intéressons à l'incertitude de modélisation de Shypre, si seules les 20 dernières années de la série sont considérées (1883-2003)[4]. Nous reproduisons la même

[4]Il y a objectivement 21 années, mais l'année 2001 ne contient aucun événement avec une pluie journalière supérieure à 20 mm.

analyse que précédemment, avec la modification suivante. Avec 20 années de données, la distribution des pluies journalières simulées par Shypre peut être fortement incertaine en queue de distribution. Dans la procédure d'ajustement d'une loi GPD à la distribution simulée par Shypre, la méthode des moindres carrés pondérés de la relation 6.4 est réutilisée, mais la somme ne porte que sur les quantiles y_i de périodes de retour inférieures à 10 ans. On calcule ainsi les paramètres α_s, k_s de la GPD qui s'ajuste le mieux à la distribution simulée par Shypre, sur sa partie inférieure (c'est-à-dire pour des périodes de retour inférieures à 10 ans). La méthode d'ajustement a été vérifiée : sur 10 000 paramètres de Shypre tirés dans leur loi a priori, la différence relative entre les quantiles donnés par Shypre et les quantiles de la GPD est inférieure à 4% pour les périodes de retour 2, 5 et 10 ans. Comme la procédure des moindres carrés n'a pas pris en compte la partie supérieure de la distribution de Shypre, ces différences relatives entre les quantiles de périodes de retour 100 et 1000 ans peuvent atteindre 50%.

De même que précédemment, six algorithmes MCMC sont simulés en parallèle pour estimer la loi a posteriori des paramètres de Shypre. La loi a priori est la même que dans la section précédente avec l'ensemble de la série. La vraisemblance est celle d'une loi GPD de paramètres α_s, k_s calculée sur les pluies journalières supérieures à 20 mm des 20 dernières années, avec α_s, k_s les paramètres de la GPD ajustée à la distribution de Shypre (sur sa partie inférieure).

Après vérification de la convergence des algorithmes, avec la statistique R de Gelman, on estime les intervalles de crédibilité à 90% des quantiles de Shypre. Comme attendu, ceux-ci sont plus larges que dans le cas de la série entière puisque les données sont moins nombreuses. On observe qu'ils contiennent les valeurs observées de la série des 20 dernières années. La figure 6.8 illustre le résultat, et compare le nouvel intervalle de crédibilité obtenu avec celui obtenu avec la série entière. On remarque que le nouvel intervalle contient l'ancien, or celui-ci contenait 443 des 444 observations de la série entière.

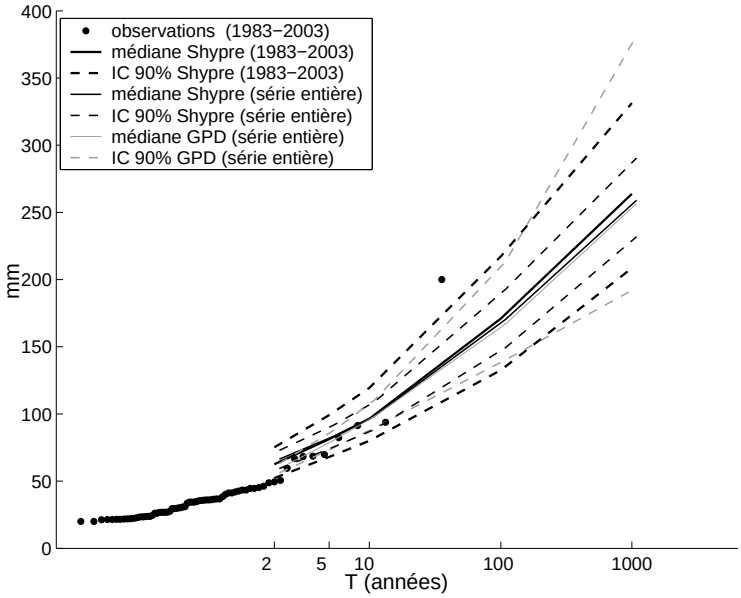


Fig. 6.8: Analyse des incertitudes de modélisation de Shypre dans le cas où on ne considère que les 20 dernières années de mesures de Marseille. Comparaison avec le cas où toutes les années sont considérées.

Enfin, nous avons comparé ce dernier intervalle de Shypre avec l'intervalle de crédibilité donné par une modélisation par une loi GPD. Les intervalles de crédibilité des quantiles de la loi GPD sont calculés avec un algorithme MCMC. Les données utilisées dans la vraisemblance sont les pluies journalières supérieures à 20 mm des 20 dernières années de mesures de Marseille, la vraisemblance est celle de la loi GPD, et la loi a priori des paramètres de la loi GPD est encore la loi a priori utilisée dans les sections précédentes pour la loi GPD. Cette fois encore, l'intervalle obtenu pour les quantiles de la loi GPD est bien plus large que celui obtenu pour les quantiles de Shypre, donnant des quantiles irréalistes[5], comme le montre la figure 6.9.

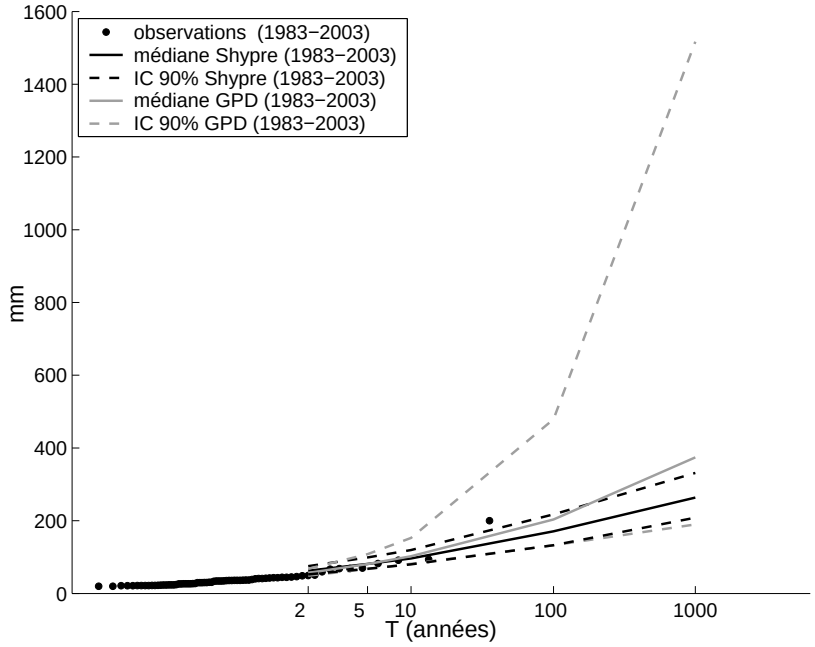


Fig. 6.9: Intervalles de crédibilité à 90% des quantiles de Shypre et de la loi GPD, calculés avec les 20 dernières années de mesures de la série de Marseille.

La figure 6.9 montre que les quantiles estimés par la loi GPD sont à l'intérieur de l'intervalle de crédibilité des quantiles de Shypre, pour les périodes de retour inférieures à 100 ans, mais le quantile millennal estimé par la loi GPD prend une valeur de 374 mm, dépassant la limite supérieure de l'intervalle de Shypre. Une fois encore, ces résultats montrent

- la forte variabilité des estimations des quantiles par une loi GPD, avec les lois a priori considérées dans cette étude,
- la faible variabilité de Shypre.

Pour terminer, la redéfinition des paramètres de la GPD par des quantiles (à la manière de Coles et Tawn (1996) et Seong et Lee (2003)), à l'aide d'un changement de variables, pourrait permettre de mieux appréhender la loi a priori des paramètres. En effet, une telle redéfinition mettrait à profit les connaissances sur les quantiles des experts en hydrologie. En particulier, les records observés en France ou dans le monde peuvent déjà donner une borne supérieure à la loi a priori des quantiles. Une autre idée serait de conserver la paramétrisation

[5]Toutes les valeurs de pluies journalières observées dans la région sont inférieures à 500 mm, le record pluviométrique français est de 1000 mm, et le record mondial est de 1800 mm, sur l'île de la Réunion.

par α, k telle qu'elle a été considérée dans cette étude, mais de rajouter une condition a priori sur les quantiles, qui ne sont en fait qu'une fonction des deux paramètres α et k. Cette idée a été mise en pratique en imposant la condition suivante : le quantile millennal doit être inférieur à 1000 mm, record pluviométrique observé en France. Cette condition est certes discutable, mais permet de donner un premier aperçu des conséquences d'une loi a priori plus informative. Cette méthode a été implémentée avec un algorithme MCMC, et le résultat (voir figure 6.10) montre que la borne supérieure de l'intervalle de crédibilité du quantile millennal a presque été divisée par deux. L'estimation médiane en revanche reste quasiment identique.

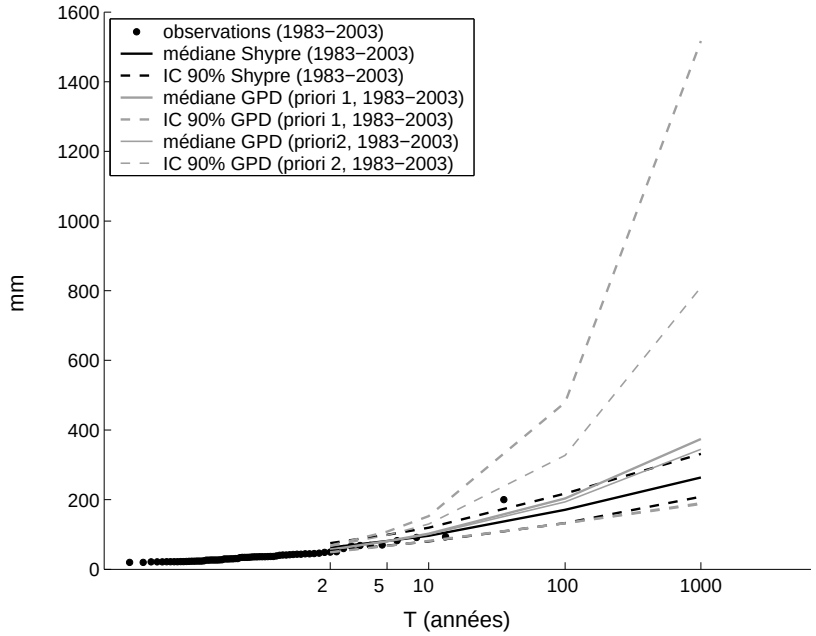


Fig. 6.10: Comparaison des incertitudes des deux modèles Shypre et GPD sur les 20 dernières années, effet d'une loi a priori plus informative sur les paramètres de la loi GPD. Priori 1 désigne la loi a priori large sur les paramètres de la loi GPD, utilisée dans les analyses du chapitre. Priori 2 désigne la loi a priori 1 tronquée par la condition suivante : le quantile millennal doit être inférieur à 1000 mm, record pluviométrique observé en France.

6.4 Conclusions

Cette étude a permis de montrer plusieurs points.

- Le modèle Shypre horaire, en version journalisée, est capable de bien reproduire les valeurs extrêmes de la pluie journalière. Ce résultat a été montré via les analyses d'incertitude d'échantillonnage et de modélisation.
- Nous avons proposé une analyse des incertitudes de Shypre, comparée avec celle de la loi GPD. Dans le cas de Shypre, les différentes sources d'incertitudes sont difficiles à évaluer en même temps. Les incertitudes d'échantillonnage et de modélisation ont été évaluées dans deux analyses séparées. Les outils utilisés sont également différents : l'analyse fréquentielle et l'analyse bayésienne ne sont certes pas fondées sur les mêmes hypothèses, mais elles ont permis de quantifier certaines incertitudes.

- Shypre estime des quantiles en étant peu sensible à l'échantillonnage, ce qui n'est pas le cas de la loi GPD.
  Les intervalles des quantiles de Shypre sont très étroits, et le fait que deux intervalles de confiance à 90% de quantiles soient disjoints pourrait laisser penser que le taux de recouvrement de la vraie valeur des quantiles par l'intervalle estimé dans le cas de Shypre est en réalité inférieur à 90%. Cependant, en statistique, la probabilité que la vraie valeur d'un quantile de fréquence donnée soit contenue dans l'intervalle de confiance à 90% de l'estimateur du quantile est de 90%. La vraie valeur du quantile peut donc être en dehors de l'intervalle avec une probabilité de 10%. Il n'est donc pas impossible que, sur six intervalles de confiance de la même variable, deux intervalles de confiance soient disjoints.
  A l'inverse, les intervalles de confiance des quantiles de la loi GPD peuvent être très larges, et un intervalle de confiance estimé sur une série de 20 ans peut être inclus dans l'intervalle de confiance estimé sur une autre série de 20 ans. L'analyse préliminaire des taux de recouvrement de la vraie valeur des quantiles élevés laisse penser que le taux de recouvrement des quantiles élevés par les intervalles de confiance à 90% estimés par les moments pondérés peuvent être plus grands que 90%. Notons que l'analyse des incertitudes sur la série entière, avec la loi GPD, a montré que le paramètre de forme est significativement négatif (impliquant une loi GPD non bornée supérieurement et non exponentielle), tandis que l'analyse sur des périodes de 20 ans ne permet pas de donner cette conclusion (l'incertitude étant plus forte, les intervalles de confiance sont plus larges et contiennent 0).
- Les incertitudes de modélisation, analysées comme des incertitudes de paramétrisation, ont été analysées dans un cadre bayésien. Les résultats sont similaires aux incertitudes d'échantillonnage, lorsque la série entière des données est exploitée. Avec une série de 20 ans, les incertitudes de modélisation augmentent, avec une incertitude excessivement large pour la loi GPD : les bornes supérieures des intervalles de confiance obtenus sont irréalistes. Néanmoins, les deux modèles GPD et Shypre sont comparés avec des lois a priori de différents types : la loi a priori des paramètres de la loi GPD est large, par manque d'information sur ses paramètres, tandis que la loi a priori des paramètres de Shypre est plus informative, grâce à une analyse régionale de ses paramètres. Dans le but de restreindre l'incertitude de modélisation dans le cas de la loi GPD, nous proposons de donner une loi a priori plus informative sur les paramètres de la GPD. Un exemple a été donné en imposant une borne supérieure aux quantiles millennaux, donnée par le record pluviométrique observé en France : la borne supérieure de l'intervalle de crédibilité du quantile millennal a presque été divisée par deux. Une loi a priori fondée sur une analyse régionale de la variabilité des paramètres de la loi GPD permettrait également de restreindre l'incertitude de modélisation de la loi GPD.

Nous avons considéré l'incertitude de modélisation via l'incertitude du choix des paramètres du modèle. Cependant l'incertitude de modélisation d'un modèle tel que Shypre peut regrouper d'autres sources d'incertitude que sa paramétrisation. Par exemple nous n'avons pas considéré l'incertitude relative à certains choix dans la construction du modèle (le choix des lois, la définition des averses en deux classes : averses principales et ordinaires, etc.). Ces incertitudes sont plus difficiles à évaluer. Une première ébauche d'analyse pourrait être donnée par une comparaison des résultats de différentes variantes du modèle, via par exemple les outils de choix de modèle donnés par l'analyse bayésienne.

Cette étude a été réalisée sur les données de pluie journalière supérieure à 20 mm. Cependant, le modèle simule également les pluies de durées 1 h à 72 h, et l'analyse des incertitudes des simulations de Shypre dans ces cas là reste encore à explorer. La méthode proposée ici peut être réutilisée à cette fin.

Quatrième partie

Modélisation de la dépendance des extrêmes

Chapitre 7

Un modèle de dépendance des extrêmes

On cherche ici à modéliser la dépendance des valeurs extrêmes d'un processus $X_1, X_2, \ldots$ stationnaire, de loi marginale notée $F(.)$. Pour un processus stationnaire, le comportement des valeurs supérieures à un seuil élevé est décrit par :

- la probabilité de dépasser le seuil,
- la distribution des dépassements de seuil,
- la structure de dépendance ou plutôt de non-dépendance à long terme entre des dépassements,
- et la structure de dépendance à l'intérieur d'un regroupement de valeurs fortes dépendantes (cluster).

Les deux premiers items sont déterminés par la loi marginale du processus des dépassements. Pour le troisième item, on peut supposer qu'à long terme, les dépassements sont approximativement indépendants, si le processus vérifie la condition 1.43 de non-dépendance à long terme $\mathcal{D}(u_n)$ de Leadbetter *et al.* (1983). Cette condition est vérifiée par toute chaîne de Markov à espace d'état continu, avec des fonctions de transition non dégénérées (O'Brien, 1987). Nous nous intéressons ici au quatrième item, déterminé par la structure de dépendance temporelle à court terme, ou encore par la structure de dépendance entre valeurs d'un même cluster. En particulier, nous reprenons la notion de persistance des valeurs fortes, introduite par Arnaud *et al.* (1998) pour les averses, définie par la propagation dans le temps de fortes valeurs de la variable X_t du processus. Nous parlerons de persistance pour désigner la persistance des valeurs fortes de la série.

La modélisation de la dépendance des valeurs extrêmes passe par les lois multivariées. Nous nous limitons ici à la modélisation bi-variée. La prise en compte de plus de deux variables est un domaine encore en perspectives. Au premier chapitre, nous avons présenté des modèles de lois bi-variées : les lois bi-variées extrêmes, et les lois bi-variées utilisant les fonctions à variation lente et introduites par Ledford et Tawn (1996, 1997). En particulier, nous avons vu que la notion d'indépendance asymptotique n'existe pas dans les lois bi-variées extrêmes, excepté le cas trivial de deux lois indépendantes. En revanche, les lois bi-variées introduites par Ledford et Tawn (1996, 1997) permettent de modéliser des variables asymptotiquement indépendantes, ou asymptotiquement dépendantes.

Dans la suite, nous avons repris l'idée de Smith *et al.* (1997), Bortot et Tawn (1998), Ledford et Tawn (2003) et Sisson et Coles (2003) : à l'intérieur des clusters, la dépendance entre

valeurs consécutives fortes est modélisée via un processus markovien. Nous présentons dans la section 7.1 un modèle de dépendance des extrêmes (noté M_L). Avant d'utiliser ce modèle pour représenter des données réelles de pluie (*chapitre 8*), nous proposons dans ce chapitre de l'appliquer à des données simulées. Nous présentons donc à la section 7.2 un modèle théorique de Morgenstern, duquel seront tirées les données simulées. Puis nous vérifions à la section 7.3 la capacité du modèle de dépendance M_L à restituer le phénomène de persistance des valeurs extrêmes des données simulées. Les conclusions de cette étude sont tirées dans la section 7.4.

7.1 Modèle de dépendance M_L

7.1.1 Présentation du modèle

Soit $X_1, X_2, \ldots$ un processus stationnaire i.d.. Nous proposons le modèle de dépendance M_L suivant (le L de M_L désigne Ledford et Tawn (1996, 1997), dont le modèle est inspiré) :

– la loi marginale est modélisée par

$$F(x) = \begin{cases} (1-\lambda)F_{v;u}(x) & \text{si } v < x < u; \\ 1-\lambda(1-k\frac{x-u}{\alpha})^{1/k} & \text{si } x \geq u \text{ et } \quad 1-k\frac{x-u}{\alpha} > 0 \end{cases} \tag{7.1}$$

où v est la valeur minimale des réalisations de X_t et u est choisi de telle sorte que les variables aléatoires X_t supérieures à u suivent une loi GPD. Le choix de la distribution de type GPD pour $X_t \geq u$ est justifié par la théorie des valeurs extrêmes. A partir d'une série d'observations de $X_1, X_2, \ldots$, u est choisi avec la seconde méthode proposée par Coles (2001), et présentée au premier chapitre de la thèse. $F_{v;u}$ est la loi des variables X_t, conditionnellement à $X_t < u$. Puisque l'objectif pratique de cette étude est de proposer un modèle de dépendance temporelle pour les séries d'averses, le choix de la loi exponentielle pour $v < x < u$ est repris de la modélisation des volumes d'averses dans Shypre (Arnaud *et al.*, 1998) :

$$F_{v;u}(x) = \frac{1-\exp\{-\alpha'(x-v)\}}{1-\exp\{-\alpha'(u-v)\}}, \tag{7.2}$$

pour $v < x < u$ et $\alpha' > 0$.

– la dépendance temporelle entre les variables successives du processus est modélisée par une chaîne de Markov. Nous empruntons la formalisation de Ledford et Tawn (1996, 1997) pour modéliser la loi jointe de (X_t, X_{t+1}) :

$$\bar{F}(x,y) = P(X_t > x, X_{t+1} > y) = a\{-\log F(x)\}^{1/2\eta}\{-\log F(y)\}^{1/2\eta} \tag{7.3}$$

avec $0 < \eta \leq 1$, $a > 0$. Rappelons que le cas $0 < \eta < 1$ correspond au cas de l'indépendance asymptotique de (X_t, X_{t+1}), avec association positive si $1/2 < \eta < 1$ et association négative si $0 < \eta < 1/2$.

7.1.2 Calcul analytique de la loi jointe du modèle M_L

Nous donnons ci-dessous les expressions analytiques de la loi jointe du modèle M_L. Ces expressions sont valables pour toute fonction de répartition $F(.)$ marginale. On note $f(.)$

la fonction de densité marginale, et $F(x,y) = P(X_t \leq x, X_{t+1} \leq y)$, $\bar{F}(x,y) = P(X_t > x, X_{t+1} > y)$ les fonctions de répartition et de survie jointes. Notons que

$$\begin{aligned} F(x,y) &= F(x) + F(y) - P(X_t \leq x \text{ ou } X_{t+1} \leq y) \\ &= F(x) + F(y) - \{1 - \bar{F}(x,y)\}. \end{aligned} \tag{7.4}$$

Pour le calcul de la vraisemblance, nous aurons besoin de connaître l'expression de $f(x,y) = \frac{\partial^2 F}{\partial x \partial y}(x,y)$. Nous donnons aussi l'expression de

$$P(X_2 \leq y | X_1 = x) = \frac{\frac{\partial F}{\partial x}(x,y)}{f(x)}, \tag{7.5}$$

qui sera utile dans la section 7.1.3 pour la simulation de la série $\{X_t\}$.

Avant d'établir ces équations, nous posons les hypothèses suivantes :

- $H1$: conditionnellement à $X_2 \leq u$, X_2 est indépendant de X_1. On ne modélise la dépendance qu'au dessus du seuil.
- $H2$: conditionnellement à $X_1 = x$, avec $x < u$, la loi de X_2 ne dépend que du fait que $X_1 < u$. Dans le cas où $X_2 \leq u$, on a X_1 et X_2 indépendants d'après H_1.

On distingue quatre cas.

Cas $x \geq u, y \geq u$

$\bar{F}(x,y) = a(-\log F(x))^{1/2\eta}(-\log F(y))^{1/2\eta}$, avec $a > 0, 0 < \eta \leq 1$.

$\bar{F}$ vérifie $0 \leq \bar{F}(x,y) \leq 1$ si la condition suivante est vérifiée :

$$0 \leq \bar{F}(u,u) = a(-\log(1-\lambda))^{1/\eta} \leq 1. \tag{7.6}$$

Des équations 7.3 et 7.4, on déduit :

$$\begin{aligned} \frac{\partial F}{\partial x}(x,y) =& f(x) + \frac{\partial \bar{F}}{\partial x}(x,y) \\ =& f(x) - \frac{f(x)}{F(x)} \frac{a}{2\eta}(-\log F(y))^{1/2\eta}(-\log F(x))^{1/2\eta - 1} \end{aligned} \tag{7.7}$$

d'où

$$P(X_2 \leq y | X_1 = x) = 1 - \frac{a}{2\eta} \frac{(-\log F(y))^{1/2\eta}(-\log F(x))^{1/2\eta-1}}{F(x)}. \tag{7.8}$$

Cette fonction de probabilité est croissante en y. On a $P(X_2 \leq y | X_1 = x) \leq 1$, et il faut imposer la condition suivante :

$$1 - \frac{a}{2\eta} \frac{(-\log(1-\lambda))^{1/\eta - 1}}{1-\lambda} \geq 0 \tag{7.9}$$

pour que la fonction de probabilité reste positive.

La densité $f(x,y)$ est alors donnée par :

$$f(x,y) = (\frac{1}{2\eta})^2 a(-\log F(x))^{1/2\eta-1}(-\log F(y))^{1/2\eta-1} \frac{f(x)f(y)}{F(x)F(y)}. \tag{7.10}$$

Cas $x \geq u, y < u$

Par conditionnement, on a :

$$\begin{aligned} P(X_2 \leq y|X_1 = x) &= P(X_2 \leq u|X_1 = x)P(X_2 \leq y|X_1 = x, X_2 \leq u) \\ &= (1 - \frac{a}{2\eta}\frac{(-\log(1-\lambda))^{1/2\eta}(-\log F(x))^{1/2\eta-1}}{F(x)})F_{v;u}(y). \end{aligned} \tag{7.11}$$

Pour établir cette équation, nous avons utilisé l'équation 7.8 et l'hypothèse $H1$. Si $y \to u$, les deux équations 7.8 et 7.11 donnent des résultats identiques. Cette fonction appartient à [0,1], quelles que soient les valeurs des paramètres, et la densité $f(x, y)$ est donnée par :

$$f(x,y) = f(x)\{1 - \frac{a}{2\eta}\frac{(-\log(1-\lambda))^{1/2\eta}(-\log F(x))^{1/2\eta-1}}{F(x)}\}\frac{\alpha' \exp\{-\alpha'(y-v)\}}{1-\exp\{-\alpha'(u-v)\}}. \tag{7.12}$$

Cas $x < u, y \geq u$

D'après l'hypothèse $H2$:

$$\begin{aligned} P(X_2 \leq y|X_1 = x) &= P(X_2 \leq y|X_1 < u) \\ &= \frac{F(y) - \lambda + \bar{F}(u,y)}{1-\lambda}. \end{aligned} \tag{7.13}$$

Cette fonction appartient à [0,1] puisque en dérivant la fonction $g(y) = F(y) + \bar{F}(u, y)$, on peut montrer que g est croissante à condition que la relation 7.9 soit vérifiée. La densité $f(x, y)$ est donnée par :

$$f(x,y) = \frac{f(x)f(y)}{1-\lambda}(1 - \frac{a}{2\eta}\frac{(-\log F(y))^{1/2\eta-1}(-\log(1-\lambda))^{1/2\eta}}{F(y)}). \tag{7.14}$$

Cas $x < u, y < u$

En utilisant les hypothèses $H1, H2$, on a :

$$\begin{aligned} P(X_2 \leq y|X_1 = x) &= P(X_2 \leq y|X_1 < u) \\ &= P(X_2 \leq y, X_2 < u|X_1 < u) \\ &= P(X_2 \leq y|X_1 < u, X_2 < u)P(X_2 < u|X_1 < u) \\ &= F_{v;u}(y)\frac{1 - 2\lambda + \bar{F}(u,u)}{1-\lambda}. \end{aligned} \tag{7.15}$$

Les deux équations 7.13 et 7.15 ont la même limite lorsque $y \to u$.

La fonction de probabilité est comprise entre 0 et 1 si la condition suivante est vérifiée

$$0 \leq \frac{1 - 2\lambda + a(-\log(1-\lambda))^{1/\eta}}{1-\lambda} \leq 1. \tag{7.16}$$

La densité $f(x, y)$ est donnée par :

$$f(x,y) = f(x)\frac{1 - 2\lambda + a(-\log(1-\lambda))^{1/\eta}}{1-\lambda}\frac{\alpha' \exp(-\alpha'(y-v))}{1-\exp(-\alpha'(u-v))}. \tag{7.17}$$

Les conditions 7.6, 7.9 et 7.16 seront prises en compte dans l'estimation des paramètres.

7.1.3 Une procédure de simulation du modèle M_L

Nous donnons ci-dessous une méthode de simulation du modèle M_L. Simuler des données selon le modèle M_L peut être utile si l'on souhaite par exemple compléter des données observées par des données simulées, ou, dans le cadre bayésien, inférer sur des variables prédictives.

On définit p_1, p_2 :

$$\begin{aligned} p_1 &= P(X_{i+1} < u | X_i > u) \\ &= 1 - \bar{F}(u,u)/\lambda \\ &= 1 - \frac{a(-\log(1-\lambda))^{1/\eta}}{\lambda}, \end{aligned} \tag{7.18}$$

et

$$\begin{aligned} p_2 &= P(X_{i+1} < u | X_i < u) \\ &= \frac{P(X_{i+1} < u) + P(X_i < u) - (1 - \bar{F}(u,u))}{1-\lambda} \\ &= \frac{1 - 2\lambda + \bar{F}(u,u)}{1-\lambda} \\ &= 1 - \frac{\lambda - a(-\log(1-\lambda))^{1/\eta}}{1-\lambda}. \end{aligned} \tag{7.19}$$

Puisque $P(X_t < u) = 1 - \lambda$, on simule x_1 une réalisation de X_1 de la manière suivante : on tire un nombre *alea* dans une loi uniforme sur [0,1].

– Si $alea < 1 - \lambda$, alors la valeur $x_1 \leq u$, et

$$x_1 = F^{-1}(alea) = v - \frac{1}{\alpha'} \log[1 - alea \frac{1 - \exp\{-\alpha'(u-v)\}}{1-\lambda}]. \tag{7.20}$$

– Si $alea > 1 - \lambda$, alors $x_1 \geq u$ et

$$x_1 = F^{-1}(alea) = u + \frac{\alpha}{k}(1 - (\frac{1 - alea}{\lambda})^k). \tag{7.21}$$

Ensuite, à partir de la réalisation x_i de la variable X_i, X_{i+1} est simulée suivant le schéma présenté ci-dessous. On reprend la notation de 7.19. Les valeurs de x_{i+1} se calculent en inversant $P(X_{i+1} \leq x_{i+1} | X_i = x_i) = alea$.

1. Si $X_i \geq u$, soient x_i sa valeur, et *alea* un nombre tiré dans une loi uniforme sur [0,1]. La probabilité critique séparant les cas $X_{i+1} \geq$ et $\leq u$ est :

$$p_3 = P(X_{i+1} \leq u | X_i = x_i) = 1 - \frac{a}{2\eta} \frac{\{-\log(1-\lambda)\}^{1/2\eta} \{-\log F(x_i)\}^{1/2\eta - 1}}{F(x_i)}. \tag{7.22}$$

– si $alea < p_3$: $x_{i+1} < u$ et

$$x_{i+1} = v - \frac{1}{\alpha'} \log[1 - alea(1 - \exp\{-\alpha'(u-v)\})/p_3]. \tag{7.23}$$

– si $alea \geq p_3$: $x_{i+1} \geq u$, et

$$x_{i+1} = F^{-1}(-alea'), \tag{7.24}$$

où $alea' = \exp\{-[\frac{(1-alea)F(x)2\eta}{a(-\log F(x))^{1/2\eta-1}}]^{2\eta}\}$, et F est la loi marginale de X_i.

2. Si $X_i < u$, soient x_i sa valeur et *alea* un nombre tiré dans une loi uniforme sur [0,1].
   – si $alea < p_2 : x_{i+1} < u$, et

$$x_{i+1} = v - \frac{1}{\alpha'} \log[1 - \frac{alea(1-\lambda)(1-\exp\{-\alpha'(u-v)\})}{1-2\lambda+\bar{F}(u,u)}]. \tag{7.25}$$

   – si $alea \geq p_2 : x_{i+1} \geq u$, et x_{i+1} s'obtient en résolvant

$$\frac{F(x_{i+1}) - \lambda + \bar{F}(u, x_{i+1})}{1-\lambda} = alea, \tag{7.26}$$

   par exemple, par une méthode de dichotomie.

7.2 Validation du modèle M_L sur simulations : le modèle de Morgenstern

7.2.1 Le modèle de Morgenstern

Le processus théorique

Soit une suite $\{X_t\}$ de variables aléatoires. La loi marginale de la série est encore donnée par :

$$F(x) = \begin{cases} (1-\lambda)F_{v;u}(x) & \text{si } v < x < u; \\ 1-\lambda(1-k\frac{x-u}{\alpha})^{1/k} & \text{si } x \geq u \text{ et } \quad 1-k\frac{x-u}{\alpha} > 0 \end{cases} \tag{7.27}$$

où u, v sont fixés arbitrairement à $u = 30, v = 15$, et $F_{v;u}$ est la distribution d'une loi exponentielle de paramètre α' tronquée en u et de seuil v.

La structure de dépendance entre deux variables X_t, X_{t+1} est donnée par la loi bi-variée de (Morgenstern, 1956) :

$$F_M(x,y) = F(x)F(y)(1+\mu\bar{F}(x)\bar{F}(y)), \qquad x > v \tag{7.28}$$

où $F(.)$ est la distribution marginale de X_t et $-1 \leq \mu \leq 1$ le paramètre de la distribution de Morgenstern. Le choix de la loi de Morgenstern est exploratoire. La loi de Morgenstern permet de décrire différents types de corrélation entre les variables modélisées. Le cas $\mu = 0$ correspond à des variables indépendantes, le cas $\mu < 0$ à des variables asymptotiquement indépendantes et négativement associées et le cas $\mu > 0$ correspond à des variables asymptotiquement indépendantes et positivement associées. Le degré de l'association croît avec $|\mu|$.

Remarquons qu'un développement limité de la fonction de survie bi-variée de Morgenstern $\bar{F}_M(x,y)$, avec une marginale $F(x)$ de loi Fréchet standard[1], donne à l'ordre 1 pour x, y tendant vers l'infini :

$$\bar{F}_M(x,y) \approx (1+\mu)x^{-1}y^{-1}. \tag{7.29}$$

Si on modélise des données simulées avec le modèle de Morgenstern, par le modèle de dépendance M_L, on peut donc s'attendre à ce que le paramètre a de l'équation 7.3 soit proche de $1+\mu$, et η proche de $1/2$. Nous retrouvons ainsi que le processus de Morgenstern est asymptotiquement indépendant.

[1] $F(x) = \exp(-1/x)$

Simulation

La simulation d'une telle suite de variables aléatoires est simple : on simule tout d'abord une réalisation de X_1, le premier terme de la suite, selon la loi marginale $F(.)$. Ensuite, connaissant la réalisation de la variable aléatoire X_t, on simule une réalisation de la variable aléatoire X_{t+1} selon la structure de dépendance de Morgenstern. Pour cela, on doit connaître

$$P(X_{t+1} \leq y | X_t = x) = \int_{-\infty}^{y} \frac{f(x,z)}{f(x)} dz. \tag{7.30}$$

Dans le cas de la structure de dépendance de Morgenstern, on a

$$f(x,y) = \frac{\partial^2 F_M}{\partial x \partial y}(x,y) = f(x)f(y)\{1 + \mu(2F(x)-1)(2F(y)-1)\}, \tag{7.31}$$

d'où

$$P(X_{t+1} \leq y | X_t = x) = F(y)^2 \mu(2F(x)-1) + F(y)\{1 - \mu(2F(x)-1)\}. \tag{7.32}$$

Une réalisation y de X_{t+1} connaissant $X_t = x$ est donc tirée en résolvant

$$F(y)^2 \mu(2F(x)-1) + F(y)\{1 - \mu(2F(x)-1)\} = alea \tag{7.33}$$

avec *alea* un réel aléatoire entre 0 et 1. L'équation 7.33 de second degré en $F(y)$ possède deux solutions

$$r_{1,2} = \frac{1 - \mu(2F(x)-1) \pm \sqrt{\Delta}}{-2\mu(2F(x)-1)}, \tag{7.34}$$

où

$$\Delta = \{1 - \mu(2F(x)-1)\}^2 + 4\mu(2F(x)-1)alea \tag{7.35}$$

est le discriminant. Or une seule de ces deux solutions appartient à [0,1] :

$$r_2 = \frac{1 - \mu(2F(x)-1) - \sqrt{\Delta}}{-2\mu(2F(x)-1)}, \text{si } F(y) \neq 1, \tag{7.36}$$

si $2F(x) - 1 = 0$, alors la solution est $F(y) = alea$. Une fois $F(y)$ calculé, on obtient y en inversant $F(y) = r_2$.

Évaluation de la dépendance des extrêmes et de la persistance des valeurs fortes du modèle de Morgenstern

L'objectif est ici de décrire les types dépendance des extrêmes et de persistance présents dans un processus de Morgenstern. Nous voulons savoir si la persistance des valeurs fortes est fortement ou faiblement marquée. Pour cela, on simule un échantillon de taille 100 000, suffisamment grand pour que les estimations de la dépendance des extrêmes et de la persistance du processus soient précises.

Nous analysons tout d'abord la dépendance des extrêmes, via les estimations empiriques de $P\{F(X_{t+1}) > w | F(X_t) > w\}, \chi, \bar{\chi}$ et η.

Comme le montre la figure 7.1, l'indépendance asymptotique du processus semble vérifiée, car $\bar{\chi} \neq 1$, $\chi(w)$ semble s'approcher de 0 lorsque $w \to 1$ et $\eta \neq 1$. A des niveaux finis ($w < 1$), les extrêmes semblent positivement associés, puisque $\bar{\chi}(w)$ est positif, $\eta(w)$ est supérieur à 0.5, et $P\{F(X_{t+1}) > w | F(X_t) > w\} > P\{F(X_t) > w\} = 1 - w$.

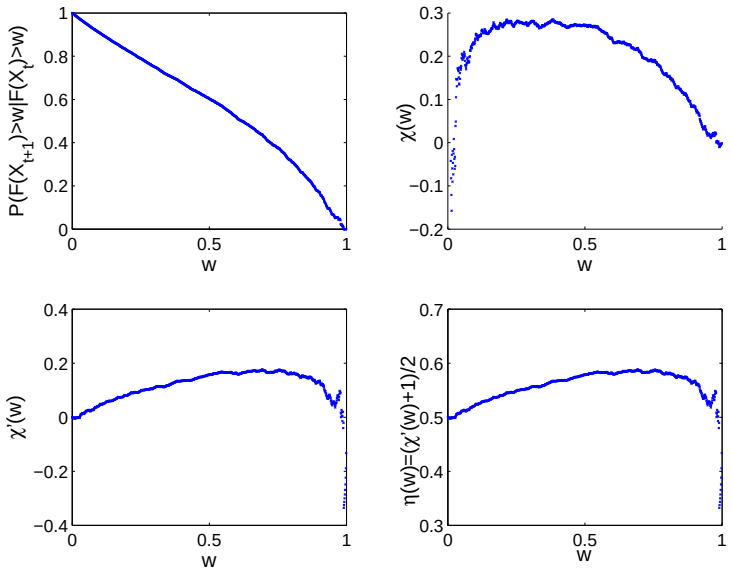


Fig. 7.1: Estimation empirique des fonctions $P\{F(X_{t+1}) > w | F(X_t) > w\}, \chi(w), \bar{\chi}(w)$ et $\eta(w)$ sur un échantillon de taille 100 000.

Ensuite, nous examinons la persistance des valeurs fortes. Une manière d'évaluer cette persistance est d'examiner la distribution de la somme de plusieurs variables consécutives. La figure 7.2 présente les distributions de la somme de deux variables consécutives $X_t + X_{t+1}$ dans le cas du processus de Morgenstern d'équations 7.27 et 7.28[2], et pour différents paramètres μ de dépendance : $\mu = 1$ (le cas de la dépendance avec corrélation positive maximale) ; $\mu = 0.8$ (que l'on choisira par la suite) ; $\mu = 0$ (le cas de l'indépendance) ; $\mu = -1$ (le cas de la dépendance avec corrélation négative maximale[3]). La structure de dépendance de Morgenstern ne crée pas de dépendance asymptotique, donc il n'est pas étonnant que les deux distributions se ressemblent en queue de distribution. Les figures montrent que pour une forte fréquence d'apparition donnée, les quantiles sont croissants avec la valeur de μ, tandis que pour une faible fréquence on observe le contraire. Ceci est un résultat attendu puisque avec μ=1 par exemple, la corrélation entre variables X_t et X_{t+1} est positive et maximale (si X_t prend une valeur forte, X_{t+1} aura tendance à prendre une valeur forte, et si X_t prend une valeur faible, X_{t+1} aura tendance à prendre une valeur faible), tandis que si $\mu = -1$ la corrélation est négative et maximale. Nous pouvons remarquer graphiquement que l'inversion de l'ordre des distributions a lieu autour de la fréquence 0.83 (dans notre cas d'étude, avec F donnée par l'équation 7.27).

Par rapport au cas de l'indépendance ($\mu = 0$), la structure de dépendance de Morgenstern avec $\mu = 1$ ou 0.8 se traduit donc par une hausse des quantiles de $X_t + X_{t+1}$, à partir de la

[2]Pour que la représentation des distributions soit lisible et ait l'allure lisse théorique, et pour éviter les problèmes d'échantillonnage, la figure a été tronquée à la fréquence F=0.993, correspondant à -log(− log F) = −5.

[3]maximale pour la valeur absolue

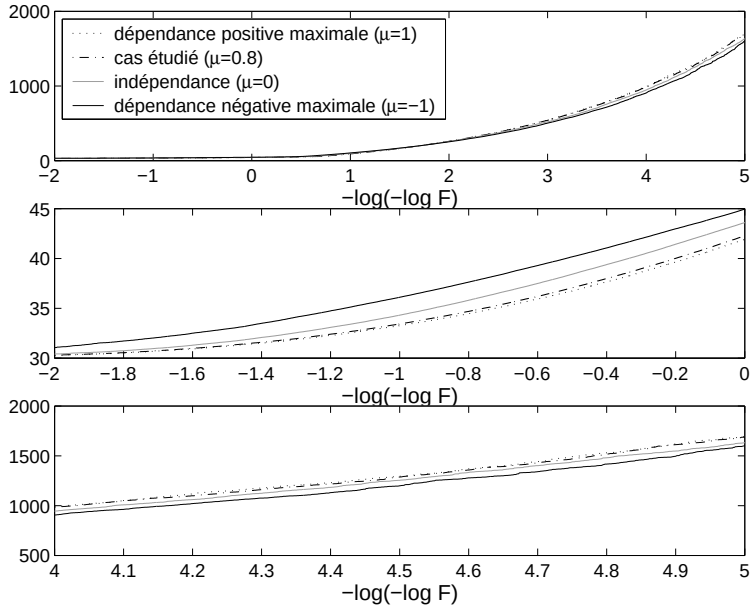


Fig. 7.2: Distribution de la somme de deux variables consécutives du processus de Morgenstern d'équations 7.27 et 7.28. Comparaison de différents cas avec $\mu = 1, \mu = 0.8, \mu = 0, \mu = -1$. Les distributions ont été évaluées à partir d'une simulation d'un échantillon de 100 000 valeurs. Au dessus : tracé des distributions entre les fréquences $6.2.10^{-4}$ et 0.993, au milieu : zoom des distributions entre les fréquences $6.2.10^{-4}$ et 0.37, en bas : zoom des distributions entre les fréquences 0.982 et 0.993.

fréquence ≈ 0.83, et une baisse des quantiles de fréquence inférieure à ≈ 0.83. Cependant, la différence entre les quantiles des différentes distributions ($\mu = 1, \mu = 0.8, \mu = 0, \mu = -1$) est peu marquée : la structure de dépendance de Morgenstern ne marque pas de persistance forte.

En fait, Gumbel (1960) et Schucany *et al.* (1978) ont montré que pour une distribution bi-variée de Morgenstern, la corrélation entre les deux variables est toujours inférieure à 1/3 en valeur absolue. De plus, pour que la persistance soit marquée, il faudrait que, conditionnellement à une forte valeur de X_t, X_{t+1} prenne une forte valeur avec une probabilité forte. Or, étant donné $X_t = x$, X_{t+1} est tiré en résolvant 7.33. Par exemple, si $\mu = 1$, la solution r_2 de l'équation 7.33 est croissante avec $F(x)$ et est donnée par :

$$r_2 = \frac{1 - F(x) - \sqrt{(1 - F(x))^2 + alea(2F(x) - 1)}}{1 - 2F(x)}, \tag{7.37}$$

et si $F(x) = 0.5$, $r_2 = alea$. Si $F(x) < 0.5$, alors $r_2 < alea$, et si $F(x) > 0.5$, $r_2 > alea$. Ainsi, quelle que soit la distribution marginale F, la structure de dépendance de Morgenstern ne permet pas de modéliser une persistance fortement marquée.

Enfin, nous avons également analysé la persistance des valeurs fortes du modèle de Morgenstern en examinant la distribution de la variable conditionnelle $X_t + X_{t+1}|X_t \geq u$. Ce point de vue permet d'évaluer l'effet d'une grande valeur de X_t sur la somme $X_t + X_{t+1}$, et donc de donner une meilleure idée de la persistance. La figure 7.3 représente les distributions de la variable conditionnelle $X_t + X_{t+1}|X_t \geq u$. Cette fois, les différences de persistance sont plus marquées entre les modèles de Morgenstern de différents paramètres μ. La figure 7.3 permet, sinon de justifier, au moins de ne pas rejeter le modèle M_L de persistance que nous avons proposé. En effet, le modèle M_L définit la persistance de manière particulière pour les cas où $X_t \geq u$ (à la fois dans la loi marginale, qui est alors une loi GPD, et dans la loi jointe de $(X_t, X_{t+1})|X_t \geq u$).

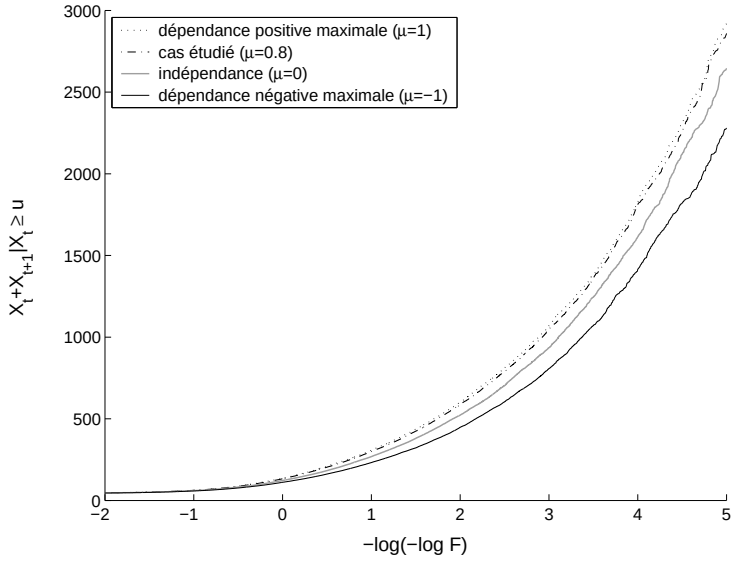


Fig. 7.3: Distribution de la somme de deux variables consécutives $X_t + X_{t+1}$, conditionnellement à $X_t \geq u$ du processus de Morgenstern d'équations 7.27 et 7.28. Comparaison de différents cas avec $\mu = 1, \mu = 0.8, \mu = 0, \mu = -1$. Les distributions ont été évaluées à partir d'une simulation d'un échantillon de 100 000 valeurs.

Nous avons donc mis en évidence l'existence de la notion de persistance dans un processus temporel dont la dépendance entre X_t, X_{t+1} est modélisée par la loi bi-variée de Morgenstern : l'effet de la persistance est visible sur l'estimation des quantiles. Cependant, la structure de Morgenstern ne discrimine pas fortement les différents types de persistance. Certes, il pourrait sembler plus intéressant, du point de vue pratique pour un travail ensuite sur des averses, de travailler avec une loi bi-variée créant une forte persistance entre valeurs consécutives. Mais nous verrons dans le prochain chapitre que les données de pluie de Marseille ne présentent pas une persistance fortement marquée. Les résultats apportés par le modèle de Morgenstern sont une première ébauche de modélisation de la dépendance dans des cas de persistance faiblement marquée.

7.2.2 Estimations des paramètres du modèle M_L

Dorénavant, nous travaillons avec un échantillon $X_1, \ldots, X_{100}$ de 100 valeurs, simulé selon le modèle de Morgenstern avec $\mu = 0.8, \lambda = 0.25, \alpha = 100, k = -0.5, \alpha' = 0.1$. Dans

cette étude, nous avons fixé $\mu = 0.8$, afin d'avoir une dépendance avec association positive relativement forte (μ n'a pas été choisi égal à 1 pour éviter les problèmes de bord lors des estimations). Le choix du seuil u au dessus duquel la loi marginale de l'échantillon peut être modélisée par une loi GPD est reporté dans la figure 7.4. u est choisi égal à 30. Nous ajustons ensuite le modèle M_L à ce processus simulé. Nous présentons deux méthodes pour estimer les paramètres $\lambda, \alpha, k, a, \eta, \alpha'$ du modèle M_L : par maximum de vraisemblance et dans un cadre bayésien.

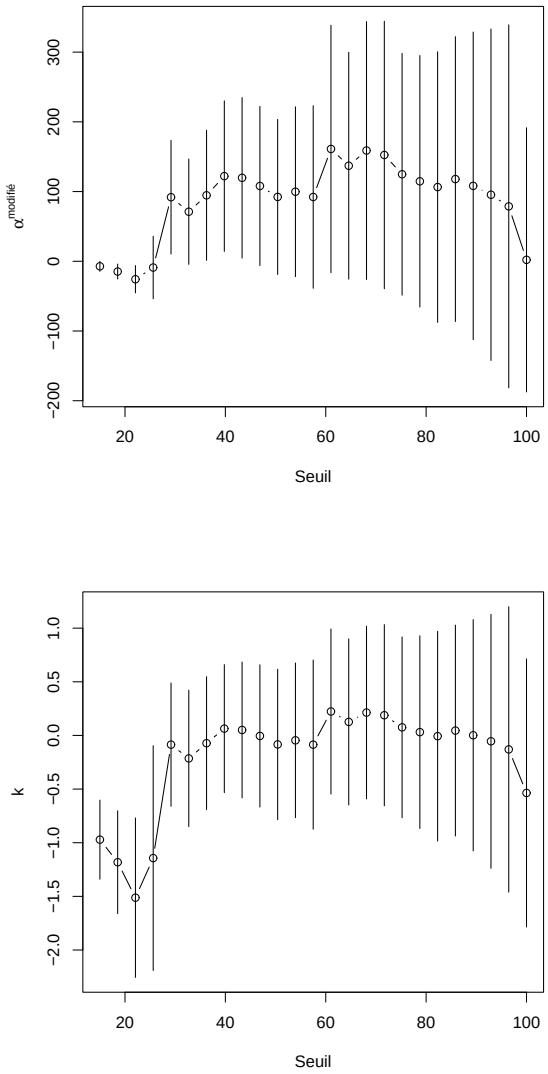


Fig. 7.4: Estimation des paramètres de la GPD en fonction du seuil pour la loi marginale de X_t, avec intervalle de confiance à 95%. Les graphes ont été tracés avec le package POT proposé par Ribatet (2006) sous le logiciel libre R. $\alpha^{modifié}(w) = \alpha(w) + k(w)w$ avec $\alpha(w), k(w)$ les paramètres de la GPD pour des données supérieures à w.

Estimation par maximum de vraisemblance

La vraisemblance du modèle M_L est donnée par :

$$f(x_1, \ldots, x_n) = f(x_1) \prod_{i=1}^{n-1} f(x_i, x_{i+1})/f(x_i), \tag{7.38}$$

où $f(.), f(.,.)$ sont les densités marginale et jointe du processus $\{X_t\}$.

L'estimation par maximum de vraisemblance sur l'échantillon $X_1, \ldots, X_{100}$ fournit $\hat{\lambda} = 0.23, \hat{\alpha} = 89.8, \hat{k} = -0.12, \hat{a} = 2.11, \hat{\eta} = 0.45, \hat{\alpha'} = 0.11$. La loi marginale est bien reproduite, comme le montre la figure 7.5. En ce qui concerne la loi jointe, nous vérifions que le modèle

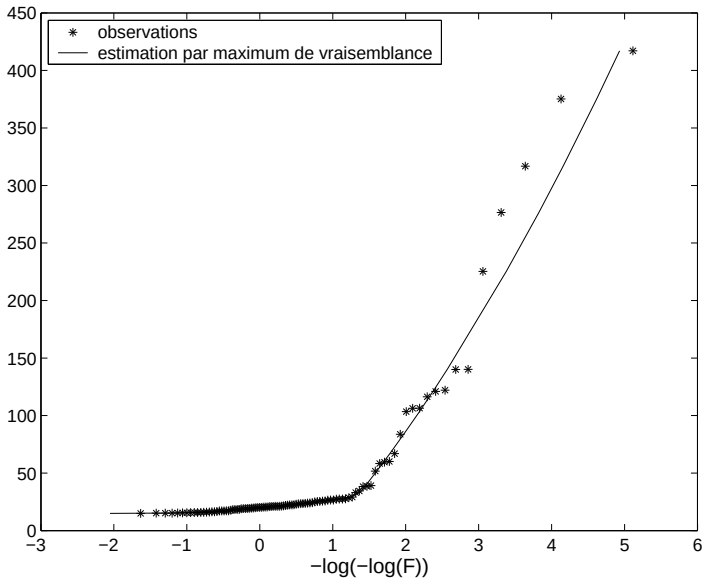


Fig. 7.5: Estimation par maximum de vraisemblance de la loi marginale de X_t.

estime correctement $P(X_{t+1} > u | X_t > u)$. Empiriquement la valeur obtenue est 0.5. Avec le modèle ajusté on a $\hat{a}(-\log(1-\hat{\lambda}))^{1/\hat{\eta}}/\hat{\lambda} = 0.48$, ce qui est proche de la valeur empirique. De plus, on a théoriquement,

$$P(X_{t+1} > u | X_t > u) = \bar{F}(u,u)/\bar{F}(u) = \frac{2\lambda - 1 + (1-\lambda)^2(1+\mu\lambda^2)}{\lambda} = 0.36,$$

et si λ est remplacé par son estimateur du maximum de vraisemblance, on obtient 0.34.

Estimation dans un cadre bayésien

À présent, un algorithme MCMC permet d'estimer la loi a posteriori des paramètres, en considérant la vraisemblance donnée par l'équation 7.38, et une loi a priori large sur les paramètres (des lois uniformes sur [0,1] pour λ, η, sur [-1.5,1.5] pour k, sur [0,10 000] pour α, a, α'). La borne 10 000 de la loi a priori des paramètres α, a, α' est bien supérieure à la valeur maximale simulée dans la loi a posteriori par l'algorithme MCMC. L'intérêt d'une telle loi non informative est seulement de donner une loi a priori non impropre (sinon la loi a priori ne serait pas une loi de probabilité).

La convergence de l'algorithme a été vérifiée avec la statistique R de Gelman *et al.* (1997), calculée sur les paramètres et pour sept chaînes simulées en parallèle avec des points de départs tirés dans les lois uniformes a priori. Après 150 000 itérations, la valeur de la statistique calculée sur les six paramètres est proche de 1 (entre 1.001 et 1.08). Nous travaillons donc avec les 50 000 dernières simulations d'un algorithme de 200 000 itérations.

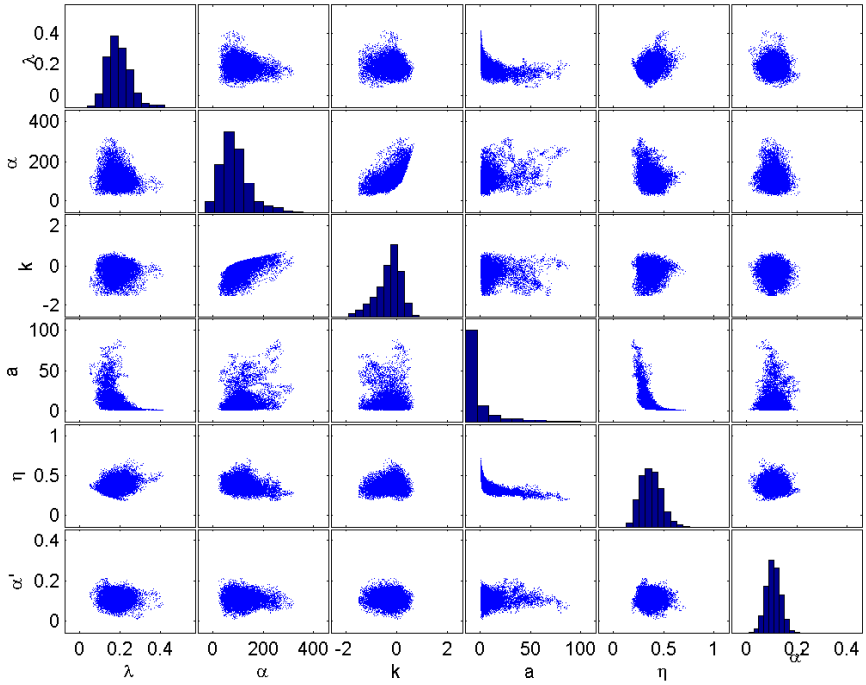


Fig. 7.6: Histogrammes des paramètres simulés par l'algorithme MCMC par les 50 000 dernières itérations d'algorithme MCMC. Représentation de la corrélation entre paramètres par le tracé des paramètres deux à deux.

Les distributions a posteriori des paramètres semblent loin d'être des lois gaussiennes, comme le montre la figure 7.6. Ceci est dû à la complexité de la fonction de vraisemblance et au lien existant entre les paramètres (voir les conditions 7.6, 7.9, 7.16).

Malgré la bonne valeur de la statistique de Gelman, proche de 1, nous observons que les algorithmes MCMC convergent difficilement sur le paramètre a. Comme le montre la figure 7.6, ce paramètre a en effet une queue lourde, et il est fortement dépendant des paramètres η et λ. Nous remarquons que différentes chaînes de Markov simulées en parallèle ont des résultats différents en terme d'intervalles de crédibilité et moyenne sur le paramètre a, tandis que la médiane reste stable (voir la figure 7.7).

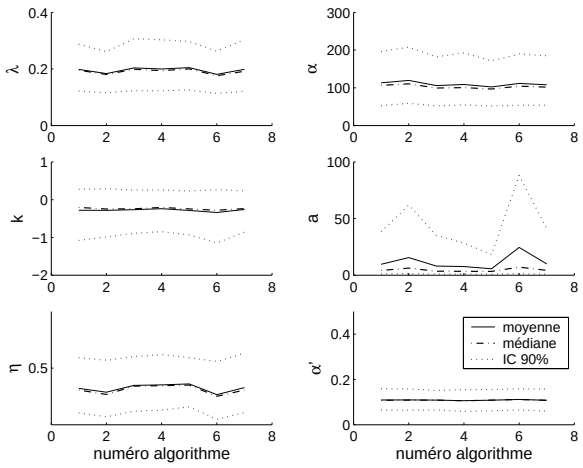


Fig. 7.7: Évolution des estimations des paramètres, sur sept chaînes de Markov simulées en parallèle (points de départs tirés dans la loi a priori). Les longueurs des chaînes de Markov considérées sont 200 000, et les estimations sont réalisées avec les 50 000 dernières simulations de chaque chaîne.

Les résultats sur les autres paramètres sont stables entre les différentes chaînes de Markov simulées en parallèle. Néanmoins, le manque de précision sur l'estimation de la loi a posteriori de a n'a pas de conséquence sur l'estimation de la loi marginale puisque celle-ci ne dépend pas de a. En revanche, l'effet de cette imprécision pourrait être important sur les estimations de la persistance : celle-ci dépend de la loi bi-variée, qui dépend elle-même de a. Or, comme le montre la figure 7.8, l'imprécision sur a n'a presque pas d'effet sur l'estimation de $P(X_{t+1} > u|X_t > u) = a(-\log(1-\lambda))^{1/\eta}/\lambda$. D'autre part, les bons résultats présentés dans la section 7.3 sur la distribution de la somme de plusieurs variables consécutives montrent que l'effet de l'imprécision de a n'est pas important.

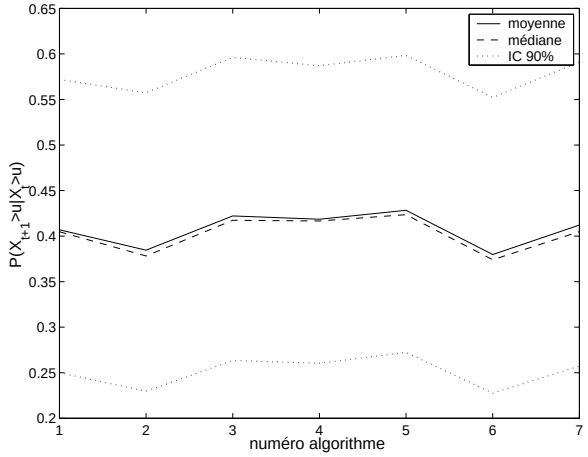


Fig. 7.8: Évolution des estimations de $P(X_{t+1} > u|X_t > u) = a(-\log(1-\lambda))^{1/\eta}/\lambda$, sur les sept chaînes de Markov simulées en parallèle (points de départs tirés dans la loi a priori). Les estimations sont réalisées avec les 50 000 dernières simulations de chaque chaîne, de longueur 200 000.

Les estimations moyennes, médianes et les intervalles de crédibilité à 90% des paramètres sont présentés dans le tableau 7.1. Les paramètres simulés (les 50 000 des sept chaînes simulées

Paramètre	Moyenne	Médiane	Intervalle de crédibilité
λ	0.20	0.19	(0.12,0.29)
α	112.71	105.55	(54.8,193.9)
k	-0.276	-0.220	(-1.023,0.269)
a	13.3	4.1	(1.1,58.7)
η	0.395	0.391	(0.249,0.559)
α'	0.12	0.12	(0.06,0.15)

Tab. 7.1: Estimation des moyennes, médianes et des intervalles de crédibilité des paramètres, réalisés avec l'ensemble des 50 000 itérations de sept chaînes de Markov simulées en parallèle.

en parallèle) sont utilisés ensuite pour estimer certaines distributions prédictives, comme la distribution a posteriori des quantiles (cf. figure 7.9) ou la distribution a posteriori de la quantité $P(X_{t+1} > u | X_t > u)$: on obtient une moyenne et une médiane de 0.4, et un intervalle de crédibilité à 90% de (0.24,0.58), ce qui contient la valeur 0.36 théorique.

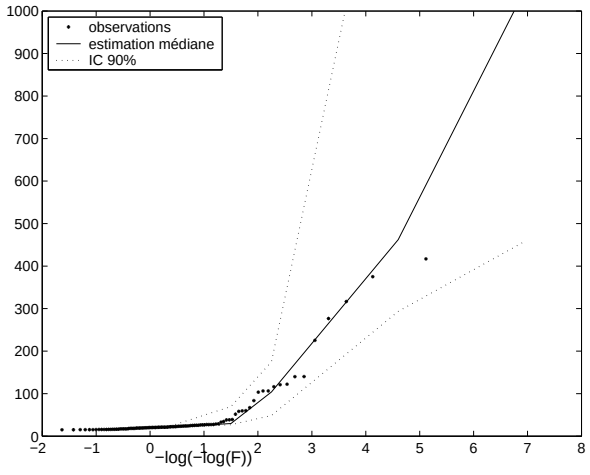


Fig. 7.9: Distribution a posteriori de la loi marginale du processus X_t étudié.

7.3 Représentation de la persistance par le modèle M_L

Pour évaluer la capacité du modèle M_L à représenter la persistance des données simulées avec le modèle théorique de Morgenstern, nous comparons les distributions des sommes de plusieurs variables consécutives ($X_t + X_{t+1}$, $X_t + X_{t+1} + X_{t+2}$ et $X_t + X_{t+1} + X_{t+2} + X_{t+3}$) dans quatre cas différents.

Les deux premiers cas sont théoriques :

– tout d'abord, pour fixer les idées, nous représentons les distributions théoriques. Le calcul des distributions de $X_t + X_{t+1}$, $X_t + X_{t+1} + X_{t+2}$ et $X_t + X_{t+1} + X_{t+2} + X_{t+3}$ n'étant pas trivial, nous estimons ces distributions par la simulation d'un échantillon de taille 100 000, pour avoir une estimation assez précise. De même que précédemment,

pour éviter le problème d'échantillonnage et pour une meilleure lisibilité des figures, nous avons tronqué les figures à la fréquence F=0.993, correspondant à -log($-\log F$) = 5. Afin de mettre en parallèle les résultats sous hypothèse de dépendance et les résultats sous hypothèse d'indépendance ($\mu = 0$), nous représentons les distributions théoriques dans le cas $\mu = 0.8$ et dans le cas $\mu = 0$.

Les deux cas suivants sont liés à l'échantillon de taille 100 $X_1, \ldots, X_{100}$, simulé selon le modèle de Morgenstern, de paramètres $\alpha = 0.8, \lambda = 0.25, \alpha = 100, k = -0.5, \alpha' = 0.1$. Le modèle M_L est ajusté à cet échantillon simulé.

- Nous considérons le modèle de Morgenstern de paramètre $\mu = 0.8$. Les autres paramètres du modèle $(\lambda, \alpha, k, \alpha')$ sont donnés par les simulations MCMC de la section 7.2.2. Pour chaque jeu de paramètres, simulé via l'algorithme MCMC, on simule une série $X_1, \ldots, X_{100}$ selon le modèle de Morgenstern et on en déduit les distributions empiriques des sommes de plusieurs variables consécutives. On en déduit finalement la médiane et les intervalles de crédibilité à 90% des quantiles de la distribution de la somme de plusieurs variables consécutives.
- Nous considérons enfin le modèle M_L avec les paramètres $\lambda, \alpha, k, a, \eta, \alpha'$ simulés par l'algorithme MCMC. Pour chaque jeu de paramètres donné par l'algorithme MCMC, on simule une série $X_1, \ldots, X_{100}$ selon le modèle M_L, avec la méthode de simulation de la section 7.1.3, et on en déduit les distributions empiriques, les médianes et les intervalles de crédibilité à 90% des quantiles de la somme de plusieurs variables consécutives.

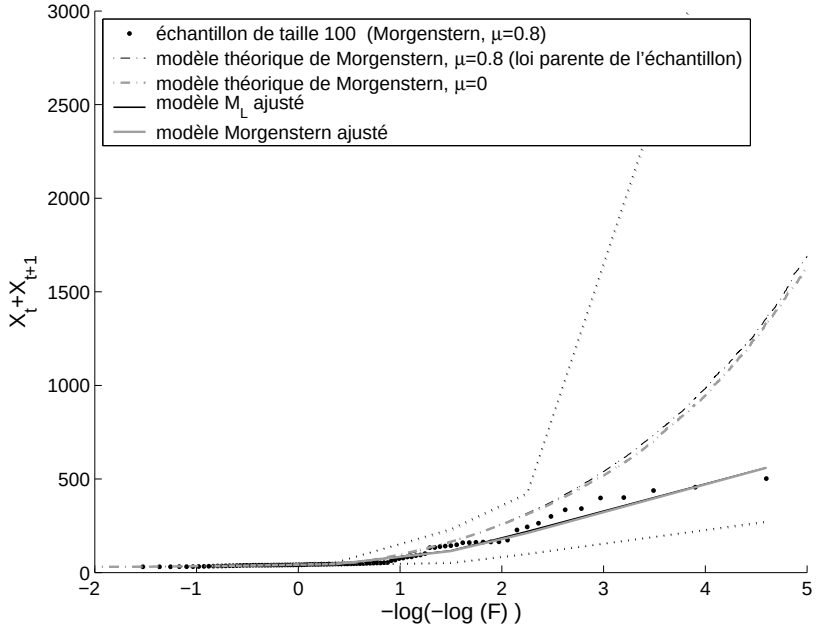


Fig. 7.10: Distribution de la somme de deux variables consécutives : estimation sur l'échantillon de taille 100, dans les cas théoriques du modèle de Morgenstern (avec $\mu = 0.8$) et i.i.d. (c'est-à-dire $\mu = 0$), et dans les cas des modèles de Morgenstern et M_L (de paramètres suivant la loi a posteriori définie à la section 7.2.2, avec les données de l'échantillon de taille 100).

Les résultats sont illustrés dans les figures 7.10, 7.11 et 7.12. Nous constatons que :

- les distributions correspondant à l'échantillon de taille 100 donnent des quantiles assez différents de ceux de la distribution théorique de Morgenstern de paramètre $\mu = 0.8$ et

de laquelle est issu l'échantillon (la distribution théorique est en fait estimée avec un échantillon de taille 100 000). Ceci est un effet d'échantillonnage et de taille d'échantillon (100 reste une taille assez petite). Nous pouvons remarquer visuellement que l'écart entre les distributions théoriques et de l'échantillon de taille 100 devient important pour des fréquences supérieures à environ 0.92.

– le modèle M_L respecte les distributions des sommes de plusieurs variables consécutives. En effet, dans chacune des figures, la distribution 'théorique' de la loi parente de l'échantillon est comprise dans les intervalles de crédibilité des quantiles obtenus avec le modèle M_L.
– De plus, le modèle M_L, dont la structure de dépendance n'est pas explicitement celle de Morgenstern, donne des résultats similaires aux résultats obtenus par le modèle de Morgenstern. Ceci montre que le modèle M_L est capable de modéliser correctement une structure de dépendance particulière.
– Enfin, on constate encore que le modèle d'indépendance n'est pas significativement inadéquat avec les observations, puisque la distribution théorique, sous hypothèse d'indépendance, est non seulement incluse dans l'intervalle de crédibilité du modèle M_L, mais elle est également proche de la distribution théorique avec hypothèse de dépendance.

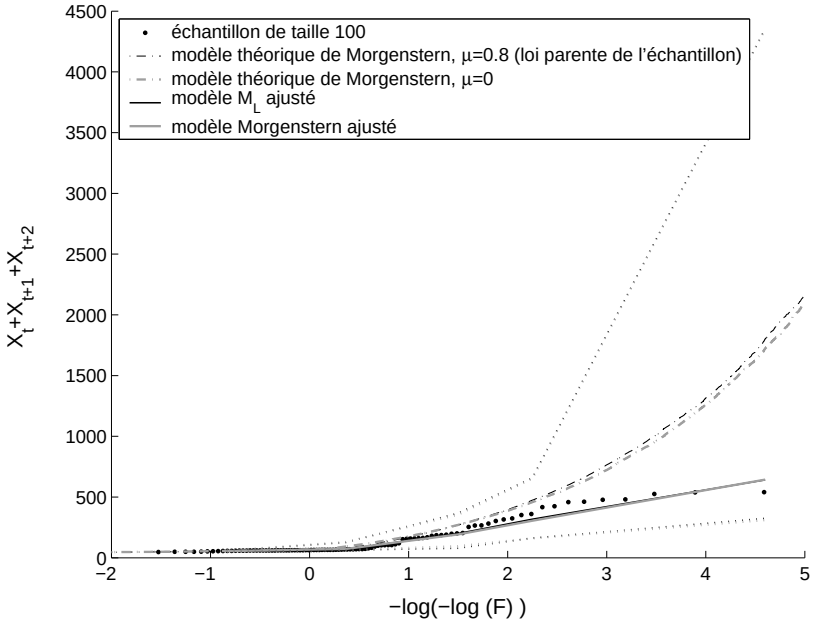


Fig. 7.11: Idem figure 7.10 mais avec la somme de trois variables consécutives.

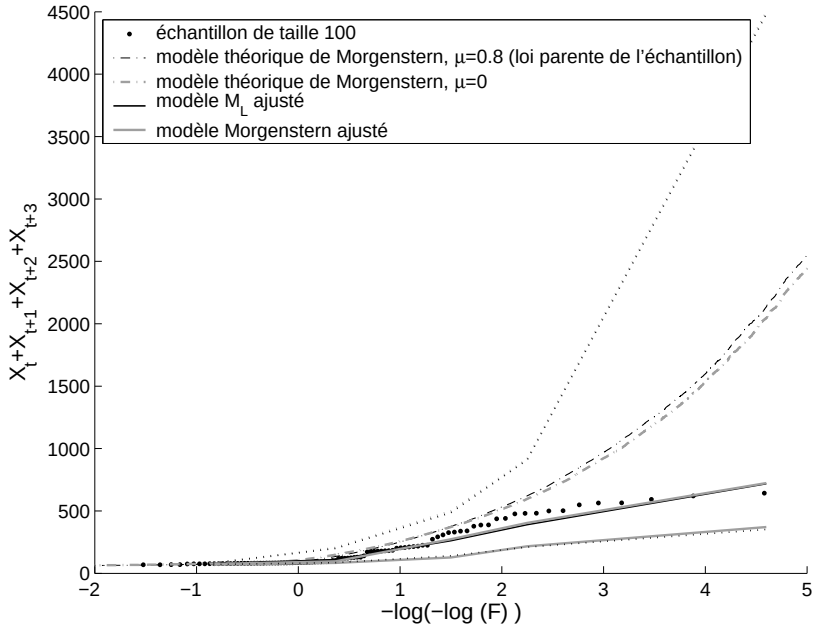


Fig. 7.12: Idem figure 7.10 mais avec la somme de quatre variables consécutives.

7.4 Conclusions

Nous avons proposé un modèle M_L de dépendance pour une série stationnaire. La loi marginale est modélisée par une loi GPD au delà d'un certain seuil élevé u, et exponentielle en deçà du seuil (le choix de la loi exponentielle est repris d'une modélisation de volumes d'averses (Arnaud *et al.*, 1998)). La loi jointe de deux variables consécutives est modélisée au delà de u par $\bar{F}(x,y) = a(-\log F(x))^{1/2\eta}(-\log F(y))^{1/2\eta}$ avec $a > 0, 0 < \eta \leq 1$ et $F(x), F(y)$ les probabilités marginales au non dépassement de X, Y. Ce modèle bi-varié est emprunté à Ledford et Tawn (1996, 1997). Il permet de modéliser des variables aléatoires asymptotiquement indépendantes (pour $\eta < 1$) et dépendantes (pour $\eta = 1$). Une méthode de simulation du modèle M_L a été proposée. Les paramètres du modèle M_L sont estimés par maximum de vraisemblance et dans un cadre bayésien, avec un algorithme MCMC. En pratique, l'algorithme MCMC semble avoir quelques difficultés à estimer le paramètre a, dont la distribution a posteriori est à queue lourde. Cependant, l'effet du paramètre difficilement estimé semble être négligeable sur les résultats du modèle M_L.

Dans ce chapitre, nous avons analysé la pertinence du modèle M_L en l'appliquant à un processus stationnaire dont la loi est connue : la loi jointe de deux variables consécutives du processus étudié ici est décrite par la structure de dépendance de Morgenstern. Le choix de la loi bi-variée de Morgenstern est expérimental.

En pratique, la structure de dépendance des valeurs extrêmes d'une série est en particulier visible dans la persistance de ses valeurs fortes. Cette persistance, définie par la propagation dans le temps de fortes valeurs de la variable de la série, s'exprime par exemple via la distribution de la somme de plusieurs variables consécutives du processus. Pour examiner la capacité du modèle M_L à reproduire la persistance, nous avons comparé les distributions de la somme de plusieurs variables consécutives sous hypothèse du modèle théorique de Morgenstern, et sous hypothèse du modèle M_L ajusté à un échantillon simulé avec le modèle

théorique de Morgenstern. Il s'est avéré que la structure de dépendance de Morgenstern n'entraîne qu'une légère dépendance entre les deux variables modélisées, car les distributions de la somme de plusieurs variables consécutives dans le cas stationnaire sont proches des mêmes distributions dans le cas d'un processus indépendant et identiquement distribué ($\mu = 0$). Cependant, malgré cette faible persistance de Morgenstern, les résultats montrent que le modèle M_L réussit à représenter la structure de dépendance de Morgenstern.

Il serait intéressant, dans des recherches ultérieures, d'appliquer le modèle M_L à d'autres processus où la persistance est plus fortement marquée. De tels processus peuvent être encore donnés par des processus markoviens, avec une loi de saut donnée par exemple par une loi bi-variée extrême, ou une loi multi-variée gaussienne (de dimension supérieure ou égale à 2), ou d'autres lois bi-variées (des exemples sont donnés dans la thèse de Mousavi Nadoshani (1997)).

Chapitre 8

Application : modélisation de la persistance des averses

Dans ce chapitre, on se place dans le contexte des événements pluvieux, tels qu'ils sont définis à la section 5.3.1.

8.1 Présentation du problème

Arnaud (2004) a remarqué que :

- les averses les plus fortes d'une série observée sont parfois regroupées dans un même événement,
- si l'événement contient une averse forte, la fréquence du volume d'une autre forte averse est supérieure à celle obtenue avec une hypothèse d'indépendance entre les fortes averses d'un événement,
- la fréquence moyenne des plus fortes averses d'un événement augmente avec le nombre d'averses de l'événement, ce qui montre encore que l'hypothèse d'indépendance des fortes averses d'un même événement ne peut pas être acceptée.

On parle donc de persistance des averses : si un événement contient une forte averse, les volumes des autres averses de l'événement ont tendance à être également forts.

On cherche à modéliser cette persistance des averses par une dépendance chronologique des averses. Pour cela, on propose d'utiliser les chaînes de Markov : le volume d'une averse dépend des volumes des averses précédentes. Les modèles de chaînes de Markov ont déjà été utilisés par Smith *et al.* (1997) et Bortot et Tawn (1998) sur des températures, par Ledford et Tawn (2003) et Sisson et Coles (2003) sur des extrêmes de pluies ou de taux de change, ou par Ledford et Tawn (1996, 1997) pour des modèles couplant deux variables (pluie-vent ou vagues-déferlement). Les modèles de chaînes de Markov, ainsi que les paramètres de dépendance des extrêmes sont présentés dans le premier chapitre de la thèse. Pour simplifier, nous utilisons dans ce chapitre des chaînes de Markov à l'ordre 1 : connaissant les averses numéro $1,\ldots,i-1$ d'un événement, le volume de l'averse numéro i (notée X_i) dépend seulement de l'averse numéro $i-1$ (notée X_{i-1}). Pour autant, les averses numéro $i-2$ (notée X_{i-2}) et i ne sont

pas indépendantes, puisque

$$\begin{aligned} P(X_{i-2}=x, X_i=y) &= \int P(X_{i-2}=x, X_{i-1}=z, X_i=y)dz \\ &= \int P(X_i=y|X_{i-1}=z)P(X_{i-1}=z|X_{i-2}=x)P(X_{i-2}=x)dz. \end{aligned} \tag{8.1}$$

Le modèle de chaîne de Markov peut être utilisé d'au moins deux manières différentes. En effet, on peut modéliser $P(X_{i+1}=y|X_i=x)$, c'est-à-dire la dépendance entre deux averses successives, ou modéliser $P(X_{i+\tau}=y|X_i=x)$, c'est-à-dire la dépendance entre deux averses séparées d'un nombre $\tau-1$ fixé d'averses. Cette étude permet en outre d'estimer le nombre d'averses nécessaires entre deux averses pour les considérer indépendantes.

8.2 Présentation du modèle

On s'intéresse au comportement bi-varié du couple $(X_t, X_{t+\tau})$ où $X_t, X_{t+\tau}$ sont les volumes des averses t et $t+\tau$. Marginalement, la loi de X_t, conditionnellement à $X_t > u$, avec u un seuil élevé, est une GPD, de paramètres d'échelle et de forme $\alpha > 0$ et k :

$$\begin{aligned} F_u(x) &= P(X_t \leq x | X_t > u) \\ &= \begin{cases} 1-(1-k(x-u)/\alpha)^{1/k}, & \text{pour } x>u \text{ et } 1-k(x-u)/\alpha>0, \text{ si } k\neq 0\,; \\ 1-\exp(-(x-u)/\alpha), & \text{pour } x>u, \text{ si } k=0. \end{cases} \end{aligned} \tag{8.2}$$

La loi non conditionnelle est donc, pour $x \geq u$ et $\lambda = P(X_t > u)$:

$$F(x) = \begin{cases} 1-\lambda(1-k(x-u)/\alpha)^{1/k}, & \text{si } k\neq 0\,; \\ 1-\lambda\exp(-(x-u)/\alpha), & \text{si } k=0. \end{cases} \tag{8.3}$$

D'autre part, on modélise la dépendance temporelle de la série par une chaîne de Markov d'ordre 1. La loi jointe $\bar{F}_\tau(x,y) = P(X_t > x, X_{t+\tau} > y)$ de deux variables successives est donnée, pour $x, y > u$, par la formalisation proposée par Ledford et Tawn (1996, 1997). Cette formalisation est capable de modéliser des variables asymptotiquement dépendantes et asymptotiquement indépendantes.

$$\bar{F}_\tau(x,y) = \mathcal{L}_\tau(-1/\log F(x), -1/\log F(y))\{-\log F(x)\}^{c_{1,\tau}}\{-\log F(y)\}^{c_{2,\tau}} \tag{8.4}$$

avec $c_{1,\tau} > 0, c_{2,\tau} > 0, c_{1,\tau} + c_{2,\tau} \geq 1$, et $\mathcal{L}_\tau$ une fonction bi-variée à variation lente.

Le modèle de variation lente utilisé ici est une constante :

$$\mathcal{L}_\tau(x,y) = a_\tau > 0 \tag{8.5}$$

On pose $\eta_\tau = 1/(c_{1,\tau}+c_{2,\tau})$, et pour simplifier, on utilise $c_{1,\tau} = c_{2,\tau} = c_\tau$.

8.3 Cas d'étude

On travaille sur une série événementielle de volumes d'averses de Marseille, pour la saison été (juin-novembre). La définition des événements est donnée dans la section 5.3. La série est

une suite d'événements, décomposés en averses. Les volumes des averses sont exprimés en dixièmes de mm. Deux événements différents sont séparés d'au moins un jour avec moins de 4 mm de pluie, et sont supposés indépendants. La série de mesures de la saison été contient 231 événements, pour un total de 1105 averses.

En supposant les événements indépendants et en notant $na(E)$ le nombre d'averses de l'événement E, on peut exprimer la densité de probabilité d'un échantillon d'averses contenues dans ne événements

$(x_{1,1}, \ldots, x_{1,na(1)}, x_{2,1}, \ldots, x_{2,na(2)}, x_{ne,1}, \ldots, x_{ne,na(ne)})$ par :

$$\prod_{E=1}^{ne} f(x_{E,1}, \ldots, x_{E,na(E)}) = \prod_{E=1}^{ne} f(x_{E,1}) \prod_{i=1}^{na(E)-1} f(x_{E,i}, x_{E,i+1})/f(x_{E,i}), \tag{8.6}$$

d'après les propriétés de la densité d'une chaîne de Markov.

8.3.1 Loi marginale

On recherche tout d'abord le seuil u à partir duquel il est justifié de modéliser les dépassements de seuil par la loi GPD. Pour cela, on considère la variable Y égale au volume maximal d'une averse pour un événement donné. L'intérêt de cette variable est que ses réalisations sont indépendantes, puisque les événements sont supposés indépendants. On ajuste alors une GPD de paramètres α_Y, k_Y à cette variable. Le seuil u est choisi de manière à ce que les paramètres estimés $k_Y(v)$ et $\alpha_Y^{modifié}(v) = \alpha_Y(v) + k_Y(v)v$ soient stables pour $v \geq u$, d'après l'approche de Coles (2001). Le seuil choisi est $u = 25$ mm (cf fig. 8.1).

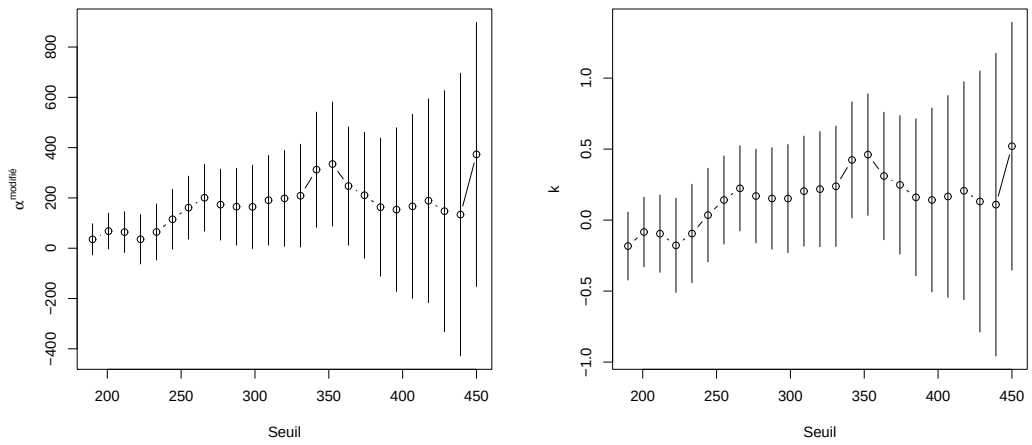


Fig. 8.1: Evolution de $\alpha_Y^{modifié}(u) = \alpha_Y(u) + k_Y(u)u$ et $k_Y(u)$ en fonction du seuil u, en dixièmes de mm, avec intervalles de confiance à 95%. Les graphes ont été tracés avec le package POT de Ribatet (2006) sous R. On choisit u=250 dixièmes de mm.

Alors en notant $\lambda_Y = P(Y > u)$, on a $\widehat{\lambda}_Y \approx 0.24$, calculé empiriquement sur les maxima des 231 événements. λ_Y représente la probabilité que l'averse maximale d'un événement

dépasse 25 mm. Les paramètres α_Y, k_Y de la GPD sont estimés par maximum de vraisemblance, sur les maxima de chaque événement : $\widehat{\alpha}_Y = 106$ et $\widehat{k}_Y = 0.035$. L'ajustement des maxima des averses par événement est présenté dans la figure 8.2.

Il est important de remarquer que la distribution de Y n'est pas la même que la distribution F de X_t. Y n'a servi ici qu'à donner un cadre théorique pour estimer un seuil u au dessus duquel il est justifié de modéliser Y par une GPD. On supposera que ce seuil est encore valable pour la distribution de X_t. On a $P(X_t \geq u) \leq P(Y \geq u) = \lambda_Y$. Même si les valeurs de X_t peuvent être modélisées par une GPD avec un seuil inférieur à 25 mm, la théorie des valeurs extrêmes montre que la valeur du paramètre de forme de la GPD n'est pas changée.

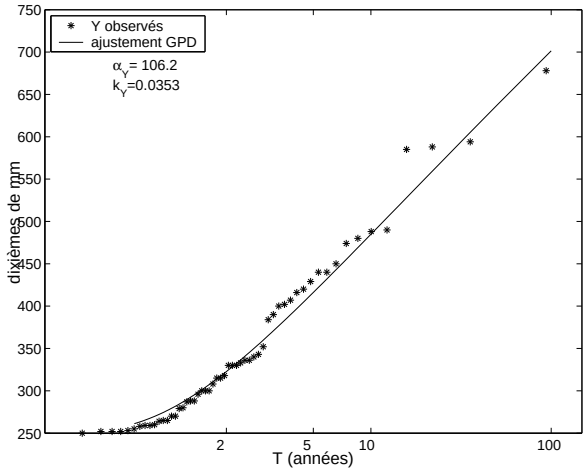


Fig. 8.2: Ajustement de la loi GPD aux averses maximales des événements, supérieures à 25 mm, par maximum de vraisemblance.

8.3.2 Étude de la dépendance entre averses

Cette section présente une étude de la dépendance entre deux averses successives ou séparées d'une ou plusieurs averses. L'intérêt est d'analyser l'effet d'une averse sur une averse suivante (immédiate ou non). Nous allons montrer que deux averses fortes peuvent être séparées de quelques petites averses. Dans ce cas, la chaîne de Markov à l'ordre 1 ne modélise pas bien la dépendance des extrêmes. Or, c'est cette dépendance ou persistance des valeurs fortes qui nous intéresse. Dans la section suivante, nous remédierons à ce problème en considérant seulement les fortes averses (les petites averses seront négligées et les fortes averses agglomérées).

La dépendance des averses séparées par un nombre τ d'averses est donc analysée par l'intermédiaire du modèle de chaîne de Markov. Plus particulièrement, à l'image de Ledford et Tawn (2003), on étudie la série temporelle définie par

$$T_t^\tau = \min\{X_t, X_{t+\tau}\} \tag{8.7}$$

pour $t = 1, \ldots, n - \tau$, où n est la longueur de la série $\{X_t\}$.

Sous les hypothèses du modèle M_L par l'équation 8.4, la loi marginale de T^τ vérifie

$$P(T_t^\tau > x) = P(X_t > x, X_{t+\tau} > x) = a_\tau(-\log F(x))^{1/\eta_\tau}, \tag{8.8}$$

pour $x > u_\tau$ où u_τ est un seuil élevé. La difficulté, soulignée par Ledford et Tawn (2003), est que la série temporelle $\{T_t^\tau\}$ n'est pas indépendante, puisqu'elle est construite à partir de données dépendantes. Néanmoins, pour avoir une idée de la valeur des paramètres a_τ, η_τ et des paramètres de $F(.)$, on propose une vraisemblance fondée sur l'hypothèse (fausse) d'indépendance. La vraisemblance ainsi définie est une vraisemblance censurée, parfois appelée pseudo-vraisemblance (Ledford et Tawn, 2003). Elle est donnée par :

$$\prod_{t:T_t^\tau<u_\tau} G(u_\tau) \prod_{t:T_t^\tau>u_\tau} g(T_t^\tau), \tag{8.9}$$

où $G(.), g(.)$ désigne les fonctions de répartition et de densité de T^τ, et u_τ est le seuil au dessus duquel la loi jointe de $X_t, X_{t+\tau}$ est modélisée par le modèle 8.4.

$$\begin{aligned} G(u) &= 1 - P(T^t > u) = 1 - a(-\log(1-\lambda))^{1/\eta}, \\ g(x) &= \frac{a}{\eta}(-\log F(x))^{1/\eta-1}\frac{f(x)}{F(x)}, \text{ pour} x \geq u \end{aligned} \tag{8.10}$$

avec $F(.)$, et $f(.)$ les fonctions de répartition et de densité marginales du processus $\{X_t\}$. Outre la méthode du maximum de vraisemblance, les paramètres η_τ peuvent également être estimés par l'estimateur non paramétrique de Hill (Ledford et Tawn, 2003), (Embrechts *et al.*, 1997).

Dans le cas des averses de Marseille, nous devons prendre en compte la partition des données en événements. Dans la définition de T_t^τ, nous rajoutons la contrainte du même événement :

$$T_t^\tau = \min\{X_t, X_{t+\tau} \text{ avec } X_t \text{ et } X_{t+\tau} \text{ dans le même événement}\}. \tag{8.11}$$

On considère que tous les seuils u_τ sont égaux à $u = 25$ mm.

Les algorithmes d'estimation du maximum de vraisemblance présentent des difficultés à capturer les paramètres maximisant la vraisemblance. Après examen des vraisemblances, il semble que la surface de vraisemblance a une forme fortement aplatie sur une large gamme de paramètres. On estime alors l'ensemble des cinq paramètres $\lambda, \alpha, k, a, \eta$ dans un cadre bayésien. La loi a priori choisie est uniforme sur [0,1] pour λ, η, sur [-1.5,1.5] pour k, sur [0,10 000] pour α et a. Le peu d'information de cette loi a priori en fait une loi a priori non informative, mais non impropre. Dans la loi a priori, on rajoute les conditions

$$\max\{2\lambda_\tau - 1, 0\} < a_\tau(-\log(1-\lambda_\tau))^{1/\eta_\tau} < \lambda_\tau \tag{8.12}$$

pour respecter le fait que :

$$\begin{aligned} &P(X_1 < u_\tau, X_{1+\tau} < u_\tau), \\ &P(X_1 > u_\tau, X_{1+\tau} > u_\tau), \\ &P(X_1 < u_\tau, X_{1+\tau} > u_\tau), \\ &P(X_1 > u_\tau, X_{1+\tau} < u_\tau) \end{aligned} \tag{8.13}$$

sont entre 0 et 1.

La distribution a posteriori définie par la vraisemblance censurée 8.9 et la loi a priori ne reflètent pas la variance réelle du modèle, puisqu'elle est fondée sur l'hypothèse fausse de

l'indépendance des données. Néanmoins, l'hypothèse d'indépendance entre deux T_t^τ consécutifs est vraie s'ils correspondent à des événements distincts (deux événements distincts sont indépendants).

Les distributions a posteriori des paramètres du modèle sont simulées avec un algorithme MCMC, de 200 000 itérations.

Laissant de côté une période de chauffe de 150 000 itérations, on calcule $P(X_{t+\tau} > u | X_t > u)$ de manière théorique, pour chacun des 50 000 derniers paramètres simulés par l'algorithme MCMC :

$$P(X_{t+\tau} > u | X_t > u) = a_\tau \{-\log(1-\lambda_\tau)\}^{1/\eta_\tau} / \lambda_\tau. \tag{8.14}$$

Ensuite, on détermine les moyennes et les intervalles de crédibilité de $P(X_{t+\tau} > u | X_t > u)$.

Les résultats obtenus avec le modèle estimé dans le cadre bayésien sont comparés avec la probabilité $P(X_{t+\tau} > u | X_t > u)$ estimée sur les observations. De manière plus générale, on note, pour $x > v$:

$$p_{\tau,x,v} = P(X_{t+\tau} > x | X_t > x, X_t \text{ et } X_{t+\tau} \text{ dans le même événement, } X_t > v, X_{t+\tau} > v) \tag{8.15}$$

on s'intéresse ici à $p_{\tau,u,0}$.

On choisit une méthode empirique pour estimer $p_{\tau,x,v}$, avec $x > v$, par l'expression :

$$\sum_{E=1}^{ne} \sum_{i=1}^{na(E)-\tau} \mathbb{1}_{X_i^E > x, X_{i+\tau}^E > x} \Big/ \sum_{E=1}^{ne} \sum_{i=1}^{na(E)-\tau} \mathbb{1}_{X_i^E > x, X_{i+\tau}^E > v}. \tag{8.16}$$

Les résultats des estimations bayésiennes et des estimations sur les observations sont présentés dans la figure 8.3.

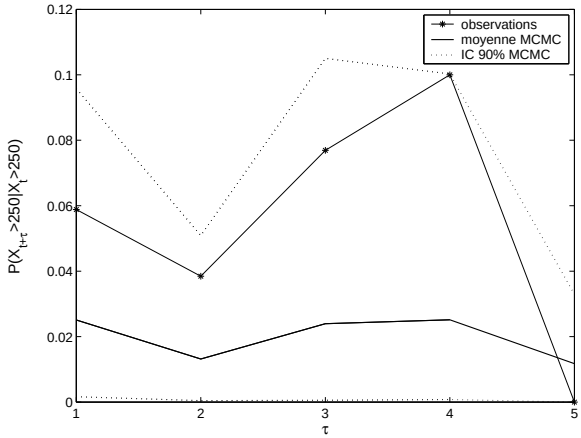


Fig. 8.3: Estimation de $P(X_{t+\tau} > u | X_t > u)$, pour différents τ entre 1 et 5. Les 'observations' correspondent aux estimations sur la série des averses d'été de Marseille. La courbe nommée 'moyenne' correspond à la moyenne de la distribution a posteriori de $P(X_{t+\tau} > u | X_t > u)$, simulée avec un algorithme MCMC. 'IC 90%' correspond à l'intervalle de crédibilité de $P(X_{t+\tau} > u | X_t > u)$ estimé par l'algorithme MCMC.

La figure 8.3 nous informe sur plusieurs points :

– sur les averses observées, la dépendance entre deux averses consécutives est moins forte qu'entre deux averses séparées de deux ou trois averses.
– Même si l'estimation de $P(X_{t+\tau} > x | X_t > x)$ par le modèle est biaisée, le modèle reproduit bien le fait que les extrêmes des averses séparées par deux ou trois averses sont plus dépendants que les averses consécutives.

Afin de représenter correctement les dépendances entre averses, les résultats précédents indiquent qu'il est préférable d'utiliser une chaîne de Markov d'ordre 4. Mais étant donné la complexité d'une chaîne de Markov d'ordre 4, nous préférons une méthode plus simple, et nous laissons la modélisation par chaîne de Markov d'ordre supérieur en perspectives. Nous proposons de ne pas tenir compte des petites averses intermédiaires entre les fortes averses. Pour cela, un seuil doit être défini, et seules les averses de volume supérieur au seuil sont conservées.

8.3.3 Agglomération des fortes averses

La chaîne de Markov à l'ordre 1 ne modélise pas de façon optimale la dépendance entre fortes averses successives si une ou plusieurs faibles averses s'intercalent entre les deux fortes averses. Pour palier ce problème, on agglomère les fortes averses entre elles, en ignorant les averses de faibles volumes. Pour cela, on doit fixer un seuil $v < u$ avec $u = 25$ mm en deçà duquel les averses sont dites de faible volume. Ce seuil doit satisfaire certains critères. D'une part, v ne doit pas être trop grand, sinon le nombre d'événements sera faible, ainsi que le nombre d'averses par événement, et l'estimation sera très imprécise. D'autre part, pour que la modélisation par chaîne de Markov à l'ordre 1 soit pertinente, la dépendance entre deux averses consécutives doit être plus forte que la dépendance entre deux averses séparées d'une ou plusieurs averses : la fonction $p_{\tau,x,v}$ doit décroître avec τ.

Remarquons que le modèle présenté ci-dessus par les équations 8.3 et 8.4, est encore valable si l'on agglomère les fortes averses : soit $x > u > v$,

$$P(X > x | X > v) = \lambda(1 - k(x-u)/\alpha)^{1/k} / P(X > v) \tag{8.17}$$
$$= \lambda'(1 - k(x-u)/\alpha)^{1/k}, \tag{8.18}$$

et

$$P(X_{t+\tau} > x, X_t > x | X_{t+\tau} > v, X_t > v) = P(X_{t+\tau} > x, X_t > x) / P(X_{t+\tau} > v, X_t > v), \tag{8.19}$$

et $P(X_{t+\tau} > v, X_t > v)$ est une constante.

Nous donnons ci-dessous les lois marginale et jointe des averses, lorsque l'on ne considère plus que les averses de volume supérieure à un volume v donné. Ces lois nous seront utiles dans la suite pour simuler des averses. Pour ne pas surcharger les notations, les variables X_t des volumes d'averses sont implicitement conditionnées par $X_t \geq v$.

Loi marginale des averses agglomérées

Soit $F_{v;u}$ la fonction de distribution des averses X_t, conditionnellement à $v < X_t < u$, avec $u = 25$ mm, alors

$$F(x) = \begin{cases} (1-\lambda) F_{v;u}(x) & \text{si } v < x < u; \\ 1 - \lambda(1 - k\frac{x-u}{\alpha})^{1/k} & \text{si } x \geq u. \end{cases} \tag{8.20}$$

Dans la suite, nous considérons que $F_{v;u}$ est une loi exponentielle, la loi déjà utilisée par Arnaud (2004) pour simuler les volumes des averses.

Loi jointe des averses agglomérées

On considère la loi donnée par l'équation 8.4 et on note : $F(x,y) = P(X_t \leq x, X_{t+1} \leq y)$, et $\bar{F}(x,y) = P(X_t > x, X_{t+1} > y)$ où X_t, X_{t+1} sont deux averses consécutives dans un même événement.

Choix du seuil v

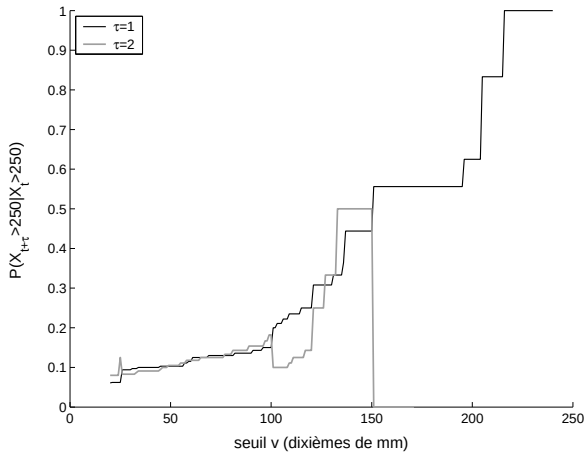


Fig. 8.4: Recherche du seuil en deçà duquel les averses sont considérées de faible volume.

La croissance de $p_{\tau,x,v}$ n'est vérifiée que pour $\tau = 1$ et 2, en raison de la forte incertitude d'estimation pour $\tau \geq 3$. D'après les résultats de la figure 8.4, le seuil $v = 10$ mm paraît satisfaire la condition $p_{1,u,v} \geq p_{2,u,v}$. Cependant, l'analyse des dépendances simulées avec les averses supérieures à 10 mm n'a pas donné de bons résultats. Finalement, le seuil v est choisi égal à 13 mm. On remarque cependant que

$$\widehat{p}_{2,x,v} > \widehat{p}_{1,x,v}$$

pour des valeurs de v supérieures à 13 mm, mais ces valeurs sont calculées avec un faible nombre de données lorsque v augmente : par exemple, pour $v = 13$ mm, il y a 13 couples (X_t, X_{t+1}) vérifiant $X_t > u, X_{t+1}$ appartenant au même événement que X_t, dont 4 tels que $X_t > u, X_{t+1} > u$; et il y a 3 couples (X_t, X_{t+2}) vérifiant $X_t > u, X_{t+2}$ appartenant au même événement que X_t, dont un tel que $X_t > u, X_{t+2} > u$. Cela illustre le peu de données extrêmes disponibles, et nous avertit d'éventuels problèmes dans les estimations par des approches empiriques.

La série ainsi définie contient 263 averses, réparties dans 190 événements :
- 134 événements d'une seule averse,
- 43 événements de deux averses,
- 10 événements de trois averses,
- 2 événements de quatre averses,

– 1 événement de cinq averses.

Le fait de négliger les petites averses est susceptible d'introduire des dépendances artificielles entre les averses. Néanmoins ces dépendances artificielles ne devraient pas être trop importantes puisque les dépendances étudiées sont limitées aux averses d'un même événement. Dorénavant, la dépendance entre valeurs successives (supérieures à v) est modélisée par une chaîne de Markov d'ordre 1, et $\tau = 1$. Pour simplifier les notations, on note η et a pour η_1 et a_1.

8.3.4 Analyse des dépendances, Estimations

L'étude des dépendances des extrêmes de la série des averses est réalisée à partir de paramètres et de mesures de dépendance $\bar{\chi}(u)$ (défini ci-dessous), $\eta, P(X_t > x, X_{t+1} > x)$. **Lorsque l'on écrit la loi jointe de X_t, X_{t+1}, on considère X_t, X_{t+1} dans le même événement. On rappelle que les événements sont indépendants entre eux. Cette indépendance a été prise en compte dans toutes les estimations. De plus, toutes les estimations sont réalisées sur les averses supérieures à 13 mm, agglomérées entre elles au sein de chaque événement.**

– L'estimation empirique de $P(X_t > x, X_{t+1} > x)$ est :

$$\sum_{E=1}^{ne} \sum_{i=1}^{na(E)} \mathbb{1}_{X_i^E > x, X_{i+1}^E > x} / \sum_{E=1}^{ne} (na(E) - 1) \tag{8.21}$$

et l'estimation empirique de $P(X_{t+1} > x | X_t > x)$ est :

$$\sum_{E=1}^{ne} \sum_{i=1}^{na(E)} \mathbb{1}_{X_i^E > x, X_{i+1}^E > x} / \sum_{E=1}^{ne} \sum_{i=1}^{na(E)} \mathbb{1}_{X_i^E > x, X_{i+1}^E \neq 0} \tag{8.22}$$

– $\bar{\chi}$ est défini ici par :

$$\bar{\chi} = \lim_{w \to 1} \bar{\chi}(w), \tag{8.23}$$

avec $\bar{\chi}(w)$ lui-même défini par :

$$\bar{\chi}(w) = \frac{2 \log P(U_t > w)}{\log P(U_t > w, U_{t+1} > w)} - 1, \tag{8.24}$$

où $U_t = F(X_t)$ suit une loi uniforme. X_t, X_{t+1} appartiennent au même événement.

– Le paramètre η mesure, s'il est différent de 1, le degré d'indépendance asymptotique entre les valeurs extrêmes consécutives de la série. On a $\eta = \frac{\bar{\chi}+1}{2}$ (voir le chapitre 1). On définit alors $\eta(w) = \frac{\bar{\chi}(w)+1}{2}$.

La figure 8.5 illustre une estimation de η via $\bar{\chi}$.

On peut également estimer η par l'estimateur de Hill (Embrechts *et al.*, 1997). Pour cela, on considère $\tilde{T}_t = \min(Z_t, Z_{t+1})$, avec $Z_t = -1/\log F(X_t)$, $Z_{t+1} = -1/\log F(X_{t+1})$, et X_t, X_{t+1} dans le même événement, et F la loi marginale des volumes d'averses :

$$\begin{aligned} P(\tilde{T}_t > x) &= P(Z_t > x, Z_{t+1} > x) \\ &= P(X_t > F^{-1}(\exp(-1/x)), X_{t+1} > F^{-1}(\exp(-1/x))) = a x^{-1/\eta}. \end{aligned} \tag{8.25}$$

Parmi les 263 averses, il reste seulement 73 couples d'averses dont les deux averses sont dans le même événement. La figure 8.6 montre l'estimateur de Hill $\hat{\eta} = (j-1)^{-1} \sum_{i=1}^{j-1} \{\log \tilde{T}_{(i)} - \log \tilde{T}_{(j)}\}$ avec $\tilde{T}_{(j)} \leq \ldots \leq \tilde{T}_{(2)} \leq \tilde{T}_{(1)}$ les j plus grandes statistiques d'ordre de $\tilde{T}$. Cet estimateur ne semble pas converger lorsque j approche 73 : on constate deux paliers (autour de 0.62 et 0.8). Le manque de stabilité peut s'expliquer par la petite taille de l'échantillon (73 couples), et par le fait que les statistiques d'ordres les plus élevés ont une forte variance. Néanmoins, il apparaît que $\eta > 0.5$.

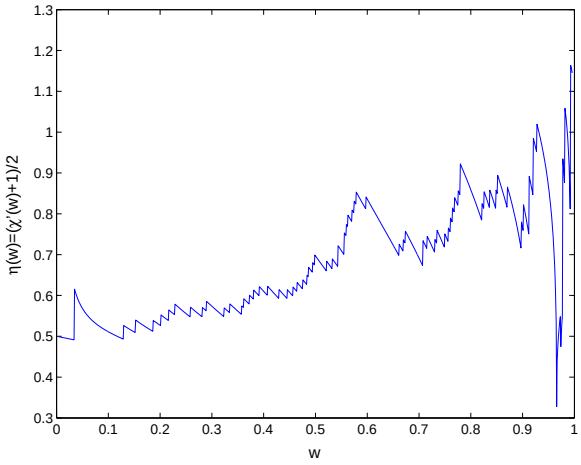


Fig. 8.5: Estimation de η, à partir de l'estimation de $\bar{\chi}_1(w)$: $\eta(w) = \frac{\bar{\chi}_1(w)+1}{2}$, sur les averses de Marseille supérieures à 13 mm.

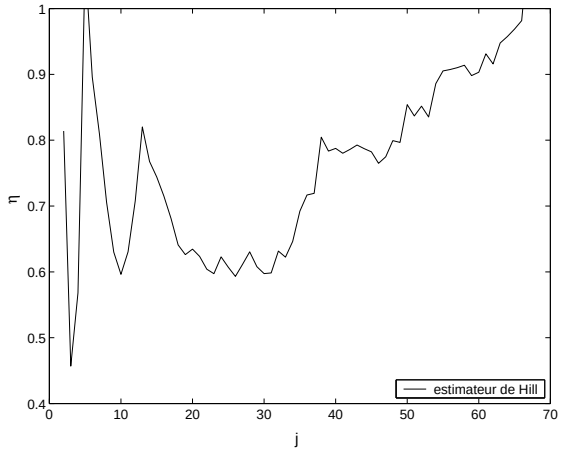


Fig. 8.6: Estimateur de Hill de η : j est le nombre des plus grandes valeurs de l'échantillon utilisées pour la construction de l'estimateur de Hill.

D'autre part, η et les autres paramètres du modèle peuvent être estimés par maximum de vraisemblance. La vraisemblance est donnée par l'équation 8.6, où la densité marginale est donnée par la dérivée de la fonction de répartition marginale 8.20, et les densités jointes sont données aux équations 7.10, 7.12, 7.14 et 7.17. Les paramètres ainsi estimés sont : $\widehat{\lambda} = 0.236, \widehat{\alpha} = 93, \widehat{k} = -0.048, \widehat{a} = 1.05, \widehat{\eta} = 0.536$ et $\widehat{\alpha'} = 0.008$. Le calcul théorique de $p_{1,u} = P(X_{t+1} > u | X_t > u) = \frac{a(-\log(1-\lambda))^{1/\eta}}{\lambda}$ donne $\widehat{p}_{1,u} = 0.386 > \widehat{\lambda} = P(X_t > u)$.

Globalement, η semble supérieur à 0.5 : les fortes averses sont donc positivement associées.

Enfin, dans le cadre bayésien, les distributions a priori choisies sont encore des lois uniformes sur [0,1] pour λ, η, sur [-1.5,1.5] pour k et sur [0,10 000] pour α, a, α'. On impose de plus les conditions 7.6, 7.9, 7.16 et 8.12 sur les paramètres. On simule sept chaînes de Markov en parallèle, de 200 000 itérations. La statistique R de Gelman est calculée sur les sept chaînes, pour les six paramètres. La valeur de la statistique est inférieure à 1.04 à partir de 150 000 itérations et pour les six paramètres. Les 50 000 dernières itérations sont considérées comme des réalisations de la loi a posteriori. De même que dans le cas théorique, on rencontre encore le problème du manque de stabilité de l'estimation du paramètre a dans les sept chaînes de Markov. Cependant, même avec cette instabilité, l'estimation de $P(X_{t+1} > u | X_t > u) = \frac{a(-\log(1-\lambda))^{1/\eta}}{\lambda}$ est stable sur les sept chaînes de Markov, comme dans l'étude théorique du chapitre précédent. On regroupe alors les 50 000 dernières itérations des sept chaînes pour donner les estimations du tableau 8.1.

Paramètre	Moyenne	Médiane	Intervalle de crédibilité à 90%
λ	0.23	0.23	(0.19,0.28)
α	95.0	93.3	(62.3,133.8)
k	-0.128	-0.103	(-0.516,0.174)
a	2.86	1.31	(0.41,8.91)
η	0.503	0.484	(0.278,0.814)
α'	0.008	0.008	(0.004,0.011)

Tab. 8.1: Moyennes, médianes et intervalles de crédibilité à 90% des paramètres du modèle de persistance des averses.

8.3.5 Validation du modèle par calculs théoriques

Capacité du modèle à reproduire les valeurs extrêmes

À présent, nous utilisons les simulations des paramètres selon leur loi a posteriori, pour en déduire les distributions prédictives a posteriori de la probabilité conditionnelle $P(X_{t+1} > x | X_t > x) = \frac{a[-\log(1-\lambda(1-k(x-u)/\alpha)^{1/k})]^{1/\eta}}{\lambda(1-k(x-u)/\alpha)^{1/k}}$. Cette étude est réalisée avec les paramètres simulés par les sept algorithmes MCMC en parallèle. On compare les estimations de $p_{1,x,v}$ pour $x \geq u$ (c'est-à-dire la probabilité conditionnelle de dépassement d'une forte valeur) calculées à partir de la série observée des averses supérieures à 13 mm, et à partir des paramètres simulés par MCMC selon leur loi a posteriori. Les résultats sont présentés sur la figure 8.7. On constate des résultats certes différents pour les sept chaînes de Markov, mais relativement proches les uns des autres (les différences relatives entre les valeurs obtenues pour une même variable sont inférieures à 20%). Les instabilités observées sont dues à l'instabilité du paramètre a. Les probabilités conditionnelles de dépassement d'une forte valeur estimées à partir des observations sont cohérentes avec la distribution a posteriori théorique des probabilités conditionnelles. Cependant, la courbe des probabilités conditionnelles présente une allure monotone dans le cas du modèle, qui n'est pas reproduite dans le cas des observations. Cette allure pourrait remettre en cause l'hypothèse d'indépendance asymptotique des averses. Néanmoins, l'estimation sur les observations est très imprécise : par exemple, pour $x = 27.5$ mm, on observe seulement neuf couples d'averses (X_t, X_{t+1}) avec $X_t > x$ (sur les

190 événements de la série), dont un seul couple tel que $X_t > x, X_{t+1} > x$[1]. Pour des x supérieurs, le nombre d'observations est encore plus réduit.

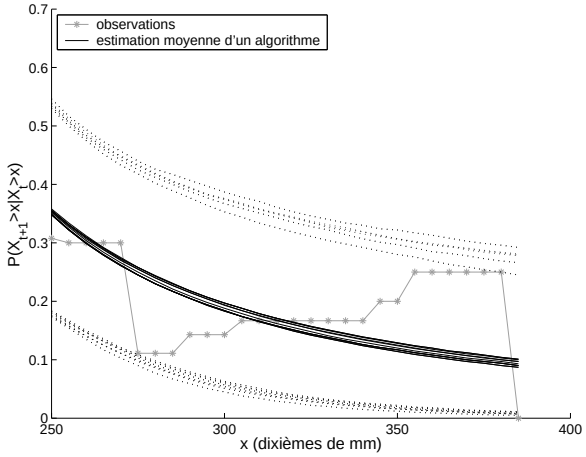


Fig. 8.7: Comparaison de $P(X_{t+1} > x|X_t > x)$ estimé sur les observations et à partir du modèle dont les paramètres sont simulés par sept algorithmes MCMC simulés en parallèle. Pour chaque algorithme MCMC, les intervalles en pointillés sont les intervalles de crédibilité à 90% de $P(X_{t+1} > x|X_t > x)$. Les observations utilisées sont les averses de plus de 13 mm agglomérées au sein de chaque événement.

Dans la suite, les résultats sont calculés avec les 50 000 dernières itérations de l'un des sept algorithmes MCMC.

Capacité du modèle à reproduire le nombre de dépassements de seuil élevé

Il est possible de calculer théoriquement la probabilité d'observer j dépassements, $0 \leq j \leq na$ dans un événement de taille na (noté $P(j|na)$). On note $p_1 = P(X_{t+1} < u|X_t > u), p_2 = P(X_{t+1} < u|X_t < u)$ les probabilités données par les équations 7.18 et 7.19.

Alors, on a pour $na = 1$:

$$P(0|1) = 1 - \lambda, \qquad P(1|1) = \lambda. \tag{8.26}$$

Pour $na = 2$:

$$\begin{aligned} P(0|2) &= P(X_1 < u)P(X_2 < u|X_1 < u) = (1-\lambda)p_2 \\ P(1|2) &= P(X_1 < u, X_2 > u) + P(X_1 > u, X_2 < u) = \lambda p_1 + (1-\lambda)(1-p_2). \end{aligned} \tag{8.27}$$

De manière générale, on a :

$$P(0|na) = (1-\lambda)p_2^{na-1}, \qquad P(na|na) = \lambda(1-p_1)^{na-1} \tag{8.28}$$

[1]Ce couple correspond aux valeurs (38.4 mm, 38.4 mm). Le fait que les deux valeurs du couple soient égales est dû à un problème de dépouillement des données. En effet, si une averse dure plus de 20 h, elle est décomposée en deux averses de même volume. On voit ici l'incertitude due aux données.

et $P(j|na)$ se calcule facilement, mais a une forme compliquée, à na et $0 < j < na$ fixés.

Nous avons donc reconsidéré les cinq paramètres $\lambda, \alpha, k, a, \eta$ du modèle, simulés selon leur loi a posteriori par une des chaînes de Markov. Pour chaque jeu de paramètres, nous avons calculé $P(j|na)$ de manière théorique. Puis nous en avons déduit des estimations moyennes et des intervalles de crédibilité à 90% de $P(j|na)$ et les fonctions de répartition $F(j|na) = \sum_{i=0}^{j} P(i|na)$. Les résultats théoriques sont confrontés aux résultats observés. La probabilité $P(j|na)$ est estimée empiriquement par :

$$\sum_{E=1}^{ne} \mathbb{1}_{na(E)=na, nd(E)=j} / \sum_{E=1}^{ne} \mathbb{1}_{na(E)=na}, \tag{8.29}$$

en notant $nd(E)$ le nombre de dépassements du seuil u dans l'événement E. Les résultats sont présentés dans la figure 8.8.

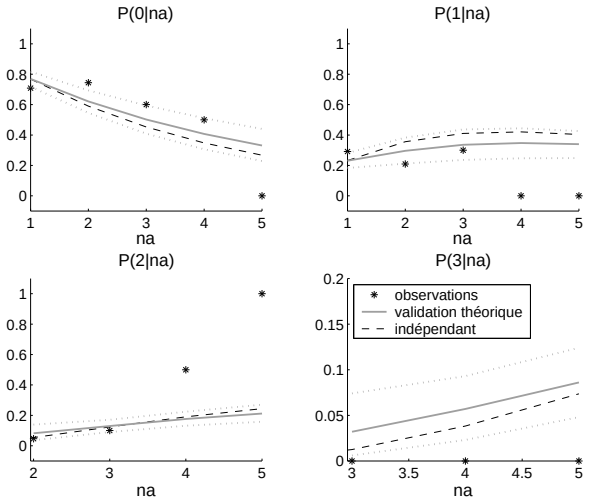


Fig. 8.8: Comparaison de $P(j|na)$ estimé sur les observations (points) et à partir du modèle (trait plein) dont les paramètres sont simulés par un algorithme MCMC. Les intervalles de crédibilités des probabilités $P(j|na)$ sont tracés en courbes pointillées. Les observations utilisées sont les averses de plus de 13 mm agglomérées au sein de chaque événement. Remarque : la valeur observée $P(2|5) = 1$ est due au fait que la série observée ne possède qu'un événement de cinq averses, et cet événement contient deux dépassements. Dans le cas de $na \geq 3$, il n'y a aucun événement observé avec 3 averses dépassant le seuil 25 mm, c'est pourquoi l'estimation de $P(j|na)$ sur les observations est 0.

Puisque les nombres d'événements de trois averses ou plus sont très faibles (inférieurs à 10), les estimations empiriques des observations sont très incertaines. Le fait que ces observations soient en dehors des intervalles peut s'expliquer ainsi. D'autre part, il est intéressant de comparer les résultats observés et simulés avec les résultats obtenus sous condition d'indépendance des averses d'un même événement. Sous hypothèse d'indépendance des averses d'un même événement, le nombre de dépassement du seuil u d'un événement contenant na averses, suit théoriquement une loi binomiale de paramètres na et $\lambda = P(X_t > u)$. Pour na=1 (l'événement ne contient qu'une averse), les modèles d'indépendance et de dépendance sont évidemment les mêmes.

Le modèle de persistance simule plus souvent les grands nombres de dépassements que sous hypothèse d'indépendance des averses. Par exemple, dans le cas d'un événement de taille $na = 5$, la probabilité d'observer 5 dépassements de 25 mm est égale à $\lambda^5 \approx 0.73\ 10^{-4}$, tandis que dans le cas du modèle de persistance proposé ici, on observe en moyenne prédictive (c'est-à-dire, d'après l'analyse bayésienne) une probabilité de 0.006 : ce qui est 12 fois plus grand que dans le cas indépendant. De même, la probabilité d'observer 3 dépassements dans un événement de taille $na = 3, 4$ ou 5 est supérieure dans le cas du modèle de persistance que dans le cas du modèle d'indépendance, comme le montre la figure 8.8.

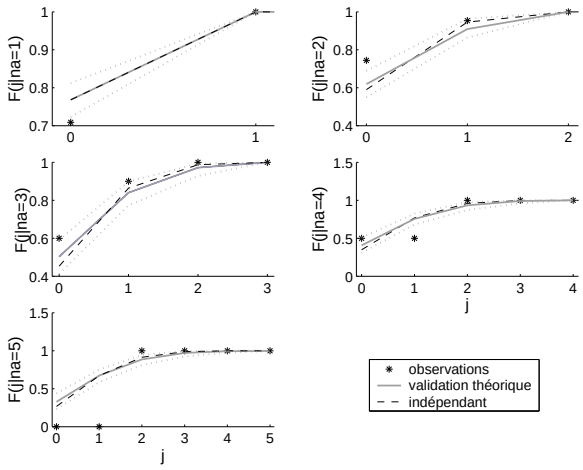


Fig. 8.9: Comparaison de $F(j|na)$ estimé sur les observations et à partir du modèle dont les paramètres sont simulés par un algorithme MCMC. Les intervalles de crédibilités des probabilités $F(j|na)$ sont tracés en courbes pointillées. Les observations utilisées sont les averses de plus de 13 mm agglomérées au sein de chaque événement.

L'intérêt de la figure 8.9 est d'illustrer autrement la persistance. Notons pour simplifier $\bar{F}(j|na), \bar{F}_{indep}(j|na)$ les fonctions de survie respectivement estimée de façon moyenne par le modèle, et calculée de manière théorique sous hypothèse d'indépendance. Pour tout $1 \leq na \leq 5$, les probabilités au dépassement vérifient l'inégalité $\bar{F}(j|na) > \bar{F}_{indep}(j|na)$ pour $j \geq 1$, et l'inégalité contraire pour $j = 0$. Cela signifie que :

- le modèle de persistance simule plus fréquemment des non-dépassements que le modèle d'indépendance,
- mais lorsqu'il y a dépassement, le modèle simule plus fréquemment un grand nombre de dépassements que le modèle d'indépendance.

D'autre part, les valeurs des probabilités $P(j|na)$ sous hypothèse d'indépendance sont incluses dans les intervalles de crédibilité à 90% des valeurs de $P(j|na)$ prédites par le modèle. Ceci montre, comme dans la figure 8.8, que les valeurs estimées sous hypothèse d'indépendance ne sont pas significativement différentes des valeurs estimées par le modèle. Ceci peut s'expliquer par une faible persistance des valeurs fortes de la série de Marseille.

8.3.6 Validation du modèle par simulation

Technique de simulation

Une chronique d'événements constitués d'averses est simulée avec la méthode de simulation donnée à la section 7.1.3. Nous rappelons que les événements sont indépendants. A l'intérieur de chaque événement, la première averse est tirée selon la loi marginale du modèle M_L, les autres averses sont simulées à partir du schéma de simulation donné à la section 7.1.3.

Pour travailler sur des variables comparables, et puisque la série des averses de plus de 13 mm compte 190 événements, on simule des séries d'événements $E_1, \ldots, E_{190}$ possédant les mêmes nombres d'averses de plus de 13 mm que dans la série observée :

$$na(E_1), \ldots, na(E_{190}). \tag{8.30}$$

Capacité du modèle à reproduire les valeurs extrêmes

Nous reconsidérons l'analyse des dépendances réalisée de manière théorique dans la section précédente, et illustrée dans la figure 8.7. Dans la section précédente, nous avons calculé $P(X_{t+1} > x | X_t > x)$ de manière théorique, selon le modèle : $P(X_{t+1} > x | X_t > x) = \frac{a(-\log(1-\lambda(1-k(x-u))^{1/k}))^{1/\eta}}{\lambda(1-k(x-u))^{1/k}}$. Ici, on recalcule $P(X_{t+1} > x | X_t > x)$ de manière empirique (voir l'estimateur 8.22) à partir de simulations du modèle. L'intérêt de cette approche est d'une part, de vérifier que l'estimateur empirique de $P(X_{t+1} > x | X_t > x)$ est proche de la valeur théorique de $P(X_{t+1} > x | X_t > x)$. D'autre part, cette phase de la validation permet de comparer les probabilités $P(X_{t+1} > x | X_t > x)$ des données observées et du modèle, avec une méthode d'estimation identique dans les deux cas.

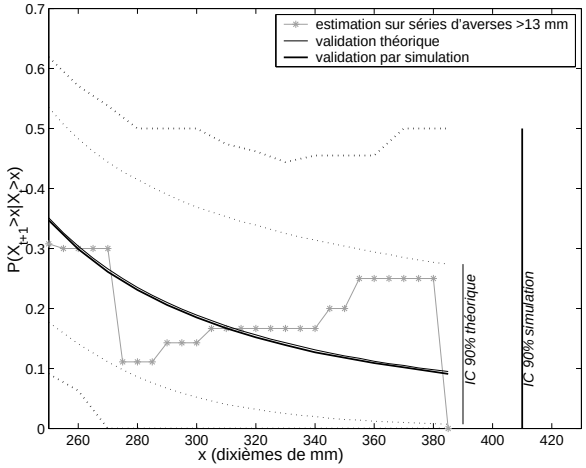


Fig. 8.10: Estimation de $P(X_{t+1} > x | X_t > x)$: comparaison des résultats observés et des résultats du modèle (moyenne en traits continus, et intervalles de crédibilité à 90% en traits pointillés). La validation théorique est reprise de la figure 8.7.

On reprend les paramètres simulés selon la loi a posteriori par l'algorithme MCMC. Pour chaque jeu de paramètres, on simule une série d'événements $E_1, \ldots, E_{190}$. On estime alors $P(X_{t+1} > x | X_t > x)$ de manière empirique. On reproduit cette procédure sur les jeux de paramètres simulés par l'algorithme MCMC. On peut alors en déduire une estimation moyenne

et des intervalles de crédibilité pour $P(X_{t+1} > x | X_t > x)$ (voir le résultat sur la figure 8.10). La figure montre la bonne cohérence des résultats du modèle avec les observations.

Le fait de simuler des séries avec le même nombre d'événements et le même nombre d'averses que la série observée permet de restituer les incertitudes dues aux données. De plus, pour un jeu de paramètres donné, la série des événements simulée est aléatoire : deux séries d'événements simulées avec le même jeu de paramètres sont différentes, et les estimations empiriques de $P(X_{t+1} > x | X_t > x)$ sont donc différentes. Les intervalles de crédibilité reflètent donc également l'incertitude de simulation, c'est pourquoi les intervalles de crédibilité sont plus larges que lors de la phase de validation théorique.

La méthode d'estimation empirique semble non biaisée : les moyennes des estimations de $P(X_{t+1} > x | X_t > x)$ sur les paramètres simulés par MCMC, dans le cas théorique et dans le cas de validation par simulation sont semblables.

Capacité du modèle à reproduire le nombre de dépassements de seuil élevé

On étudie ici le nombre de dépassements du seuil 25 mm par événement, conditionnellement au nombre d'averses dans l'événement.

De même que précédemment, on simule des séries d'événements $E_1, \ldots, E_{190}$ possédant les mêmes nombres d'averses que dans la série observée

$$na(E_1), \ldots, na(E_{190}). \tag{8.31}$$

Ainsi, les nombres d'événements de na averses seront les mêmes dans la série observée et dans la série simulée. Par exemple, la série observée ne contient qu'un événement de cinq averses. La distribution du nombre de dépassements est donc estimée de manière plus qu'incertaine dans la série observée. Cette incertitude due à l'échantillonnage est conservée dans la série simulée, et est reflétée par les intervalles de crédibilités calculés en analyse bayésienne.

On reprend les paramètres simulés par l'algorithme MCMC. Pour chaque paramètre, on simule une série de 190 événements, possédant le même nombre d'averses que dans la série observée des averses de plus de 13 mm. On calcule alors les probabilités $P(j|na)$ pour chaque jeu de paramètre (via l'estimateur 8.29). On peut également calculer la fonction de répartition des dépassements :

$$F(j|na) = \sum_{i=0}^{j} P(i|na). \tag{8.32}$$

De manière classique, on estime alors la moyenne et les intervalles de crédibilité à 90% de ces probabilités ponctuelles $P(j|na)$ et des probabilités au non dépassement $F(j|na)$.

Les figures 8.11 et 8.12 représentent respectivement les probabilités conditionnelles $P(j|na)$ et les probabilités au non dépassement conditionnelles $F(j|na)$, ainsi que les probabilités sous hypothèse d'indépendance.

On compare les résultats obtenus à partir de la série observée des averses supérieures à 13 mm, à partir des simulations du modèle, et sous hypothèse d'indépendance des averses d'un même événement. Les résultats du modèle sont cohérents avec les observations. Les résultats présentés ici incluent les incertitudes dues à l'échantillonnage : les intervalles de crédibilités sont donc plus larges que dans le cas des figures 8.8 et 8.9.

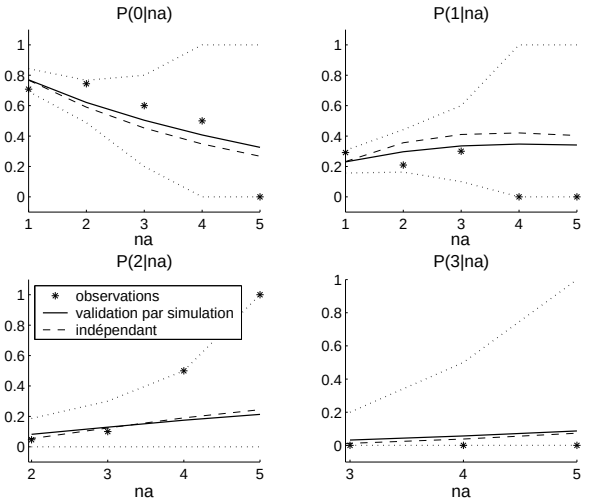


Fig. 8.11: Estimation des probabilités conditionnelles de dépassement du seuil u=25 mm $p(j|na)$. Les courbes en pointillés représentent les intervalles de crédibilités à 90% des probabilités $p(j|na)$ sous hypothèse du modèle de persistance.

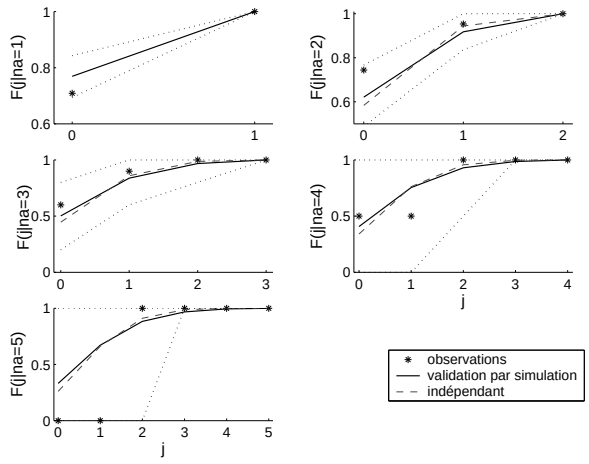


Fig. 8.12: Estimation des fonctions de répartition conditionnelles de dépassement du seuil u=25 mm $F(j|na)$. Les courbes en pointillés représentent les intervalles de crédibilité à 90% des probabilités $F(j|na)$ sous hypothèse du modèle de persistance.

Enfin, nous examinons l'effet de la modélisation de la persistance sur le nombre moyen d'averses de volume supérieur à 25 mm, dans un événement de taille na donnée. Pour cela, on compare les résultats sur la série observée (cf. tableau 8.2), avec les résultats obtenus avec le modèle (cf. tableau 8.3), et les résultats obtenus si l'on suppose l'indépendance des averses (cf. tableau 8.4). En ce qui concerne les observations, ces moyennes sont calculées empiriquement. De même, les moyennes sont calculées empiriquement pour chaque chronique simulée (de même caractéristiques que la série observée) par le modèle avec chacun des paramètres échantillonnés par l'algorithme MCMC. On donne la moyenne et l'intervalle de crédibilité à 90% du nombre moyen de dépassements dans le tableau 8.3. Dans le cas où les averses sont supposées indépendantes entre elles, le nombre moyen de dépassements du seuil 25 mm par événement de taille na se calcule théoriquement par la moyenne d'une loi binomiale de paramètres (na,λ), c'est-à-dire $na \cdot \lambda$. On suppose que $\lambda \approx 0.23$.

Lorsque l'événement ne contient qu'une averse, la modélisation de la persistance n'entre pas en jeu, et on peut s'attendre à observer en moyenne λ dépassements de 25 mm. Or, on observe dans le tableau 8.2 qu'environ 29% des averses des événements mono-averse sont supérieures à 25 mm. Cela peut s'expliquer par un nombre important d'événements forts isolés, de type orage par exemple.

na	Nombre moyen de dépassements de u par événement de taille na	Nombre d'événements de taille na
1	0.29	134
2	0.3	43
3	0.5	10
4	1	2
5	2	1

Tab. 8.2: Nombre moyen d'averses de volume supérieur à u=25 mm, dans un événement de taille na donnée. Estimations sur la série observée.

na	Nombre moyen de dépassements de u par événement de taille na
1	0.23 (0.157,0.306)
2	0.31 (0.162,0.511)
3	0.39 (0.1,0.8)
4	0.48 (0,1.5)
5	0.57 (0,2)

Tab. 8.3: Nombre moyen d'averses de volume supérieur à 25 mm, dans un événement de taille na donnée. Estimations moyennes et intervalles de crédibilité à 90%, obtenus sur les séries simulées par le modèle, dans le cadre bayésien.

na	Nombre moyen de dépassements de u par événement de taille na
1	0.23
2	0.46
3	0.69
4	0.92
5	1.15

Tab. 8.4: Nombre moyen d'averses de volume supérieur à 25 mm, dans un événement de taille na donnée. Estimations dans le cas où les averses sont supposées indépendantes.

Là encore, on constate que l'indépendance ne donne pas de résultats significativement différents de l'hypothèse de persistance du modèle. Ceci peut être dû à la faible persistance des averses de Marseille.

Capacité du modèle à reproduire la distribution des maxima des volumes d'averses par événement

Nous présentons ici une analyse des maxima des volumes des averses par événement. Plus particulièrement, le volume maximal d'une averse, le volume maximal du cumul de deux, trois ou quatre averses successives sont analysés.

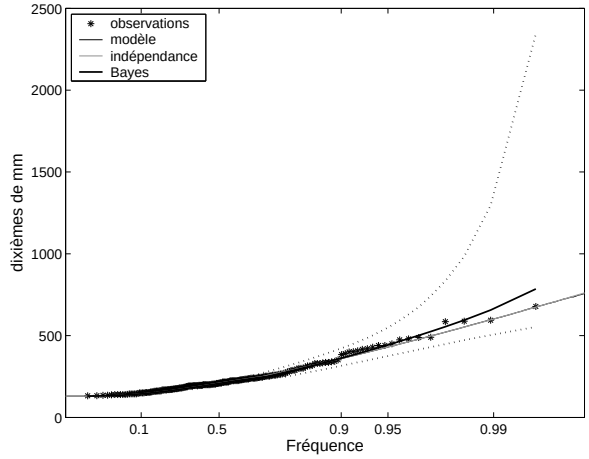


Fig. 8.13: Distribution du volume maximum d'une averse par événement : observation, modèle avec et sans persistance. Les averses utilisées sont uniquement celles de plus de 13 mm.

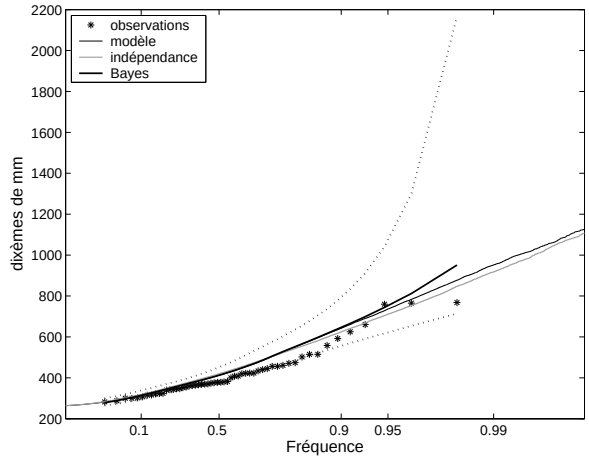


Fig. 8.14: Distribution du cumul maximum de deux averses consécutives par événement : observation, modèle avec et sans persistance. Les averses utilisées sont uniquement celles de plus de 13 mm.

La figure 8.13 montre la distribution du maximum de volumes d'averses par événement. Les figures 8.14, 8.15, 8.16 représentent les distributions du maximum de la somme de 2, 3, 4 averses successives dans un événement.

– Les points sont les distributions empiriques observées,

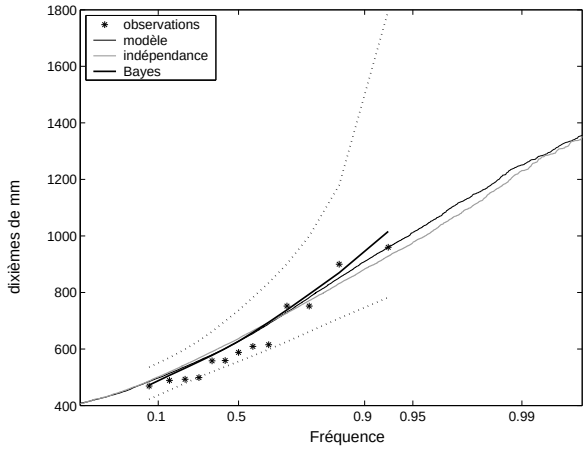


Fig. 8.15: Distribution du cumul maximum de trois averses consécutives par événement : observation, modèle avec et sans persistance. Les averses utilisées sont uniquement celles de plus de 13 mm.

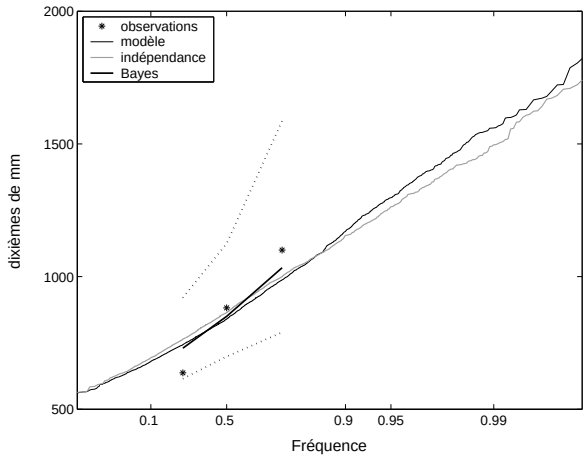


Fig. 8.16: Distribution du cumul maximum de quatre averses consécutives par événement : observation, modèle avec et sans persistance. Les averses utilisées sont uniquement celles de plus de 13 mm.

- les courbes noires en trait fin sont les distributions obtenues par 1000 simulations de chroniques des 190 événements de Marseille, avec le modèle dont les paramètres sont estimés par maximum de vraisemblance,
- les courbes grises en trait fin sont les distributions obtenues par 1000 simulations de chroniques des 190 événements de Marseille, en utilisant seulement la loi marginale de paramètres estimés par maximum de vraisemblance (c'est-à-dire sous hypothèse d'indépendance des averses entre elles),
- les courbes en trait plein épais et en pointillés sont obtenues par analyse bayésienne : pour chaque réalisation de la loi a posteriori des paramètres, on simule une chronique des 190 événements de Marseille. Ensuite les maxima des événements et de la somme de deux, trois ou quatre averses consécutives sont triés dans l'ordre croissant. Les médianes et intervalles de crédibilité à 90% de ces statistiques d'ordre sont ensuite calculées.

La figure 8.13 représente la distribution du maximum du volume d'une averse dans un événement. Il n'y a pas de différence entre le modèle avec persistance et le modèle sans persistance, ce qui n'est pas surprenant puisque si on ne fait pas la somme d'averses, la modélisation de la persistance n'est pas visible. Cette figure montre que le modèle reproduit bien la loi marginale des volumes des averses.

Encore une fois, le modèle d'indépendance ne semble pas significativement différent du modèle de persistance, puisque les distributions correspondant au modèle d'indépendance sont incluses dans les intervalles de crédibilité des distributions des maxima, estimées dans le cadre bayésien (cf. figures 8.14, 8.15 et 8.16). De même que dans le cas théorique étudié au chapitre précédent, la modélisation de la persistance induit des quantiles d'averses légèrement plus élevés que la modélisation sans persistance, pour des fréquences élevées (correspondant à la fréquence au delà de laquelle $X_t > u$). Sur des fréquences basses, le contraire est observé car la dépendance est essentiellement gouvernée par le paramètre $p_2 = P(X_{t+1} < u | X_t < u)$ (estimé à environ 0.8, dans le cas de l'estimation par maximum de vraisemblance).

La comparaison des résultats des modèles avec ou sans persistance et des observations montre des résultats corrects, sauf pour les maxima de la somme de deux averses consécutives. En effet, les modèles avec ou sans persistance semblent légèrement surestimer la distribution observée.

Marginalement, on a bien $P(X_t > u) \approx 0.23$ dans les deux modèles avec ou sans persistance. Mais on a $P(X_{t+1} > u | X_t > u) \approx 0.35$ en moyenne (d'après les estimations bayésiennes, et avec un intervalle de crédibilité à 90% de [0.18,0.54]) dans le cas de la modélisation avec persistance, et $P(X_{t+1} > u | X_t > u) \approx 0.23$ dans le cas de la modélisation sans persistance. On obtient 0.31 dans le cas des observations.

Au vu des figures, et de ces résultats, il semble que la modélisation de la persistance peut légèrement sur-estimer les quantiles d'averses. Le regroupement des averses fortes introduit peut-être des dépendances artificielles.

8.4 Conclusions

L'analyse des averses a montré que la dépendance entre averses séparées de trois ou quatre averses est plus forte qu'entre deux averses consécutives. Cela signifie que des petites averses s'intercalent entre les fortes averses. Or, c'est la dépendance des fortes averses qui est responsable des plus fortes pluies observées, et c'est la dépendance des fortes averses que nous

cherchons à modéliser. Nous avons d'abord éliminé les petites averses, inférieures à un seuil fixé à 13 mm, puis avons uniquement considéré les averses supérieures à ce seuil. L'analyse de la persistance des fortes averses a été réalisée via le modèle M_L, présenté au chapitre précédent. Le modèle M_L définit le comportement marginal des averses, ainsi que le comportement bi-varié de deux averses consécutives via une modélisation markovienne au premier ordre de la dépendance temporelle. La modélisation proposée traite explitement les valeurs extrêmes en séparant les averses en deux populations par l'intermédiaire d'un seuil, fixé à 25 mm. En particulier, la loi marginale des averses supérieures à 25 mm est une loi GPD, tandis que la loi marginale des averses inférieures à 25 mm est une loi exponentielle. La loi bi-variée des averses consécutives supérieures à 25 mm est modélisée via le modèle de Ledford et Tawn (1996, 1997). Les paramètres du modèle M_L sont estimés par maximum de vraisemblance et dans un cadre bayésien.

Pour un jeu de paramètres donné, il est possible de simuler une chronique d'événements constitués d'averses. La simulation de chroniques peut permettre ensuite d'estimer les distributions de certaines variables des averses extrêmes (nombre de dépassements de seuil donné par événement, valeurs des dépassements, valeurs des cumuls des dépassements, etc.)

Par ailleurs, la modélisation proposée considère deux populations d'averses selon que ces averses sont supérieures ou inférieures à un seuil assez élevé (dans le cas de Marseille, ce seuil a été fixé à 25 mm). Cette distinction est une alternative à celle des 'averses principales' et 'averses ordinaires' proposée par Arnaud (1997). La définition de deux populations d'averses, par une loi exponentielle et une loi GPD semble donner de bons résultats en terme de valeurs extrêmes simulées.

Nous avons constaté, comme dans le cas théorique du chapitre précédent, que la modélisation de la persistance n'a pas beaucoup d'effet sur la distribution de la somme de plusieurs averses successives. Ceci peut être dû au fait que les données issues de simulation du modèle de Morgenstern et les données de Marseille présentent une faible persistance et sont asymptotiquement indépendantes (car 1 n'appartient pas à l'intervalle de crédibilité à 90% de η). De plus, sur les averses de Marseille, 134 événements sur 190 possèdent une seule averse, et la dépendance ne peut s'étudier que sur des événements de taille supérieure à deux averses. Nous avons également vu que le nombre moyen de dépassements du seuil u est plus grand dans le cas des événements mono-averses que dans l'ensemble des événements de toutes tailles confondues. Cela suggère peut-être la nécéssité d'introduire une modélisation spécifique au nombre d'averses de l'événement.

Pour permettre d'analyser le modèle de façon plus probante, il serait intéressant de l'appliquer à des données présentant une plus forte persistance (par exemple, données de poste des Cévennes, ou en milieu tropical).

Enfin, le modèle proposé ici utilise des chaînes de Markov d'ordre 1 avec une modélisation bi-variée empruntée à un modèle de Ledford et Tawn (1996, 1997). Nous laissons en perspectives l'utilisation de modèles d'ordres supérieurs, ou de lois multi-variées avec dépendance asymptotique si les données présentent une dépendance asymptotique. Citons également Mousavi Nadoshani (1997) qui présente plusieurs lois bi-variées appliquées à l'étude des régimes de crue, en précisant les avantages et inconvénients de chacune.

Conclusion générale

Nous nous sommes intéressés dans cette thèse au comportement de la distribution locale des pluies extrêmes en France. Les valeurs extrêmes ont été considérées au travers de diverses variables : les maxima annuels ou saisonniers de pluie journalière ou horaire, les valeurs de pluie journalière dépassant un seuil élevé, et enfin le processus temporel de la succession des événements pluvieux. Les lois de ces variables et des processus ont essentiellement été étudiées avec des modèles issus de la théorie des valeurs extrêmes, ou avec des modèles de génération de pluie.

Avant de conclure sur les résultats des différents modèles, précisons que la validation et les applications des modèles à des données réelles sont soumises aux erreurs et incertitudes associées aux données et au choix des modèles eux-mêmes. Nous avons donc inclus l'analyse des incertitudes dans les analyses des modèles, via des méthodes fréquentistes ou bayésiennes. Nous avons souvent préféré les méthodes bayésiennes, qui présentent l'avantage de ne pas reposer sur des hypothèses asymptotiques et permettent de prendre en compte une information exogène aux données via la loi a priori (par exemple : information régionale, avis d'expertise). Il faut cependant être conscient de l'effet important de la loi a priori sur la quantification de l'incertitude. De même que l'apport d'informations a priori sur le comportement d'un modèle modifie le point de vue de l'expert, un changement de loi a priori n'est pas sans conséquences sur les bornes des intervalles de crédibilité de quantiles.

Nous avons tout d'abord analysé la distribution des maxima annuels de pluie journalières au travers du modèle général des valeurs extrêmes : la loi GEV. Une analyse préliminaire, avec des données simulées, a montré l'incertitude considérable de l'estimateur du paramètre de forme, incertitude qui se propage dans l'estimation des quantiles de valeurs extrêmes. Afin de réduire l'incertitude des estimations des quantiles extrêmes, de nombreuses publications ont modélisé les maxima annuels de pluie journalière par une loi Gumbel, fixant ainsi le paramètre de forme à 0. Récemment, certains auteurs ont montré l'inadéquation de la loi Gumbel aux données de maxima annuels de pluie journalière. En outre, les publications récentes ont en général supprimé l'hypothèse simplificatrice de loi Gumbel, et utilisé une loi GEV sans imposer la nullité du paramètre de forme. L'enjeu d'un tel débat entre la loi Gumbel et la loi GEV est considérable puisqu'il est directement lié à la sécurité des structures hydrauliques, à l'établissement des zones d'inondation et aux estimations des événements extrêmes. Cependant, il est difficile de donner une réponse tranchée à la question 'Que choisir entre une loi Gumbel et une loi GEV ?'. En effet, la variabilité de l'estimateur du paramètre de forme de la loi GEV est telle que l'hypothèse d'une loi Gumbel contre une loi GEV non bornée supérieurement et non Gumbel est souvent acceptée. Ce résultat, donné par quelques publications récentes sur différents postes à l'échelle mondiale, a été vérifié dans cette thèse par l'analyse des maxima annuels journaliers de 22 postes français de plus de 100

ans de mesures, fournis par Météo-France. Pour ces 22 postes, l'hypothèse de la loi Gumbel est rejetée sur environ la moitié des postes, et acceptée sur l'autre moitié. Ces résultats soulignent la difficulté de rejeter la loi Gumbel au profit de la loi GEV. Enfin, signalons une proposition formulée dans la littérature : afin de réduire l'incertitude de l'estimateur du paramètre de forme de la loi GEV, tout en conservant l'hypothèse d'une loi GEV non bornée supérieurement et non Gumbel, Koutsoyiannis (2004b) suggère de considérer une loi GEV de paramètre de forme constant sur une région donnée (en l'occurrence, l'Hémisphère Nord, et avec un paramètre de forme égal à -0.15).

Nous avons ensuite proposé une analyse plus détaillée des maxima annuels ou saisonniers de pluie, via une analyse multi-variée des pluies mesurées à différents pas de temps mobiles entre 1 heure et 72 heures, et des pluies journalières mesurées à un pas de temps fixe de 24 heures. Les lois marginales des distributions des pluies sont supposées être des lois GEV. En ce qui concerne la modélisation multi-variée des variables, trois modèles de dépendance entre les pluies de différentes durées ont été proposés et comparés. Deux analyses bibliographiques nous ont permis de justifier théoriquement une relation d'égalité entre les paramètres de forme des lois GEV des maxima annuels de différentes durées (Nadarajah *et al.*, 1998) et une relation entre les paramètres de la loi GEV des pluies maximales en 24 heures (Robinson et Tawn, 2000), mesurées sur des pas de temps mobiles de 24 heures, et des pluies maximales journalières, mesurées sur des pas de temps fixes de 24 heures. Une référence bibliographique d'hydrologie (Koutsoyiannis *et al.*, 1998) a permis de définir d'autres relations entre les paramètres des lois GEV des pluies de différentes durées. Nous avons en particulier modélisé la relation de dépendance entre les maxima de pluie en 24 heures et les maxima de pluie en 72 heures, via une loi extrême bi-variée (la loi logistique). Enfin, pour des raisons physiques, nous avons supposé l'indépendance des pluies de courtes durées et des pluies de longues durées. Les trois modèles proposés ont huit ou neuf paramètres, dont l'estimation nécessite des données de pluies horaires et journalières. Ces modèles présentent un intérêt en ingénierie hydrologique puisque les données journalières sont souvent mieux renseignées (en couverture spatiale et en taille des séries) que les données horaires. Dans notre cas d'étude situé à Marseille, les pluies journalières sont en outre de meilleure qualité en particulier au niveau des valeurs extrêmes, puisque la série horaire de Marseille présente des lacunes importantes lors des événements extrêmes, tandis que la série journalière, longue de 122 années, contient les événements extrêmes en lacune dans la série horaire. L'intérêt de l'analyse multi-durée de la pluie est réel puisque les trois modèles multi-variés ont permis d'estimer des quantiles plus forts qu'avec des estimations marginales. De plus et excepté pour l'un des trois modèles, les résultats montrent que la loi Gumbel n'est pas adaptée pour modéliser les valeurs extrêmes, tandis que la loi GEV non bornée supérieurement l'est (à un niveau de risque de 10%). Dans le cas du modèle pour lequel la loi Gumbel est acceptée au niveau 10%, il est à noter que l'intervalle de crédibilité à 90% du paramètre de forme ([-0.18, 0.01]) contient 0 de manière presque frontalière et donne une préférence à une loi GEV non bornée supérieurement. Enfin, différents tests ont montré que la dépendance entre les pluies maximales en 24 heures et 72 heures est significativement mieux représentée par la loi extrême bi-variée logistique que par une hypothèse d'indépendance.

Le comportement des pluies extrêmes peut également être étudié avec un générateur de pluie. Nous avons présenté une analyse bibliographique des générateurs de pluie existants. Plus particulièrement, nous avons étudié les valeurs extrêmes de pluie des hyétogrammes simulés par le modèle Shypre. Nous avons également analysé les incertitudes associées aux quantiles estimés par le modèle. Cette analyse a été fondée sur un cas d'étude, avec la

longue série de données journalières de Marseille, et comparée avec une modélisation probabiliste issue de la théorie des valeurs extrêmes : la loi GPD ajustée aux dépassements d'un seuil élevé. Les incertitudes d'échantillonnage et la sensibilité des modèles Shypre et GPD à l'échantillonnage ont été analysés avec une méthode fréquentielle. Les résultats ont montré que le modèle Shypre est peu sensible à l'échantillonnage et respecte les valeurs extrêmes observées. Au niveau des quantiles de période de retour moyenne (10 ans par exemple), les estimations issues de Shypre et de la GPD sont assez similaires. En revanche, sur les grandes périodes de retour (1000 ans), l'analyse des incertitudes de la loi GPD a montré des résultats contraires à ceux de Shypre : la loi GPD est plus sensible à l'échantillonnage, et les quantiles estimés sont plus incertains. Enfin, nous avons noté encore une fois la difficulté de donner un avis tranché entre une loi GPD et une loi exponentielle. L'analyse fréquentielle sur la série entière de Marseille (122 années) a montré que la loi GPD modélisant les dépassements de seuil élevé a un paramètre de forme significativement négatif : la loi est donc non bornée supérieurement et non exponentielle. En revanche, la même analyse sur des sous séries de 20 ans est teintée d'une plus grande incertitude. Les intervalles de confiance du paramètre de forme sont alors plus larges, ils contiennent la valeur 0, et l'on ne peut plus conclure au rejet de la loi exponentielle. Nous avons ensuite analysé les incertitudes de paramétrisation des deux modèles, avec une méthode bayésienne. Les résultats, à prendre avec précaution car la loi a priori des paramètres de Shypre est informative, tandis qu'elle l'est très peu dans le cas des paramètres de la GPD, montrent que les incertitudes de paramétrisation de Shypre sont faibles, tandis qu'elles sont importantes dans le cas de la loi GPD.

Enfin, la dernière partie de la thèse aborde une description événementielle du processus pluvieux, via une succession temporelle supposée stationnaire d'une série d'averses. La particularité d'une telle série est la dépendance temporelle entre ses valeurs consécutives, et plus encore, la dépendance des valeurs extrêmes : lorsqu'une averse forte compose un événement composé de plusieurs averses, d'autres fortes averses peuvent apparaître dans l'événement avec une plus forte probabilité. Nous appelons ce phénomène la persistance des averses fortes. Nous avons proposé de modéliser cette persistance via un processus markovien au premier ordre. Plus précisément, la dépendance temporelle entre valeurs fortes est modélisée avec une loi bi-variée paramétrée avec un coefficient de dépendance de queue, directement lié au type de dépendance des averses fortes. De plus, au delà d'un certain seuil, la loi marginale du modèle est supposée être une loi GPD. Le modèle ainsi proposé (noté M_L) a été appliqué à deux cas d'étude : un cas d'étude théorique, de loi parfaitement connue, avec une structure de dépendance de Morgenstern et une loi marginale GPD au delà d'un seuil élevé, et un cas d'étude réel avec des averses de Marseille. Dans les deux cas étudiés (cas théorique de Morgenstern et réel de Marseille), les variables consécutives des processus sont asymptotiquement indépendantes, et la persistance est faiblement marquée, mais bien représentée par le modèle M_L. Cependant, la persistance étant faible, le modèle d'indépendance n'est pas significativement différent du modèle M_L. L'intérêt du modèle M_L par rapport à un modèle d'indépendance n'a donc pas pu être complètement mis en valeur, par manque de temps. En perspectives, une application du modèle M_L à des données présentant une forte persistance (données de postes des Cévennes, ou de climat tropical) permettrait de conclure de façon plus précise quant à l'intérêt et la pertinence du modèle.

Perspectives

Nous avons utilisé les outils de la modélisation multi-variée pour représenter la distribution des extrêmes de pluies mesurées sur différents pas de temps, au niveau local. Ces mêmes méthodes d'analyse multi-variée des valeurs extrêmes, associées à des techniques de géostatistique peuvent être appliqués en analyse régionale pour étudier les données de différents postes. Ce point de vue régional permettrait d'enrichir les résultats par une analyse plus globale, comprenant un nombre plus important de données. Le risque régional de dépassement de valeurs de référence pourrait être évalué et comparé avec le risque local, estimé avec la distribution locale des pluies maximales.

Les modèles multi-variés proposés pourraient également être appliqués à l'analyse spatiale des pluies. Cette analyse concerne l'abattement des intensités de pluies en fonction de la superficie touchée par l'événement. L'étude de la distribution des valeurs maximales de la lame d'eau spatiale de bassins peut permettre d'évaluer le risque pluviométrique à différentes échelles spatiales.

Le choix d'une loi multi-variée adéquate reste cependant un point délicat : il existe un nombre infini de modèles multi-variés des valeurs extrêmes (il n'existe pas de paramétrisation finie de ces modèles), et ces modèles impliquent la dépendance asymptotique des co-variables. En dimension supérieure à trois, ces modèles peuvent devenir rapidement compliqués. Il existe également un nombre infini de modèles n'appartenant pas aux modèles multi-variés des valeurs extrêmes. Par exemple, les lois multi-variées les plus simples à manipuler sont les lois gaussiennes, mais elles reposent sur une hypothèse de normalité des vecteurs modélisés. Des estimations non-paramétriques peuvent être une alternative au problème du choix d'un modèle paramétrique particulier.

Le phénomène de persistance des averses fortes, observé par Arnaud (1997, 2004), a été modélisé dans la thèse via un modèle markovien au premier ordre (noté M_L). Le modèle M_L a été appliqué à des données théoriques ou réelles présentant une faible persistance des valeurs fortes. Bien que le modèle donne des résultats corrects, la faible persistance des données ne met pas en valeur l'intérêt d'utiliser un tel modèle au lieu d'un modèle, plus simple, d'indépendance. Des résultats plus probants pourraient être obtenus en appliquant le modèle proposé à des données présentant une forte persistance (par exemple des données théoriques issues d'un modèle impliquant une dépendance importante entre les fortes valeurs, ou des données réelles de postes des Cévennes ou de milieu tropical). La modélisation proposée, si elle apporte des résultats satisfaisants, peut être insérée dans le modèle Shypre. Le cadre bayésien proposé au chapitre 6, pourra être repris pour comparer les incertitudes issues de différentes formulations de la persistance des averses (voire de différentes définitions des averses).

Enfin, la loi de saut du processus markovien inclus dans le modèle M_L a été définie avec

une loi bi-variée proposée par Ledford et Tawn (1996, 1997). Ce choix peut être modifié, par exemple en utilisant une loi bi-variée des valeurs extrêmes. Une telle loi implique la dépendance asymptotique entre averses successives. Cette hypothèse serait justifiée pour des types particuliers de pluies (par exemple sur des postes soumis à des pluies intenses de longues durées). Par ailleurs, l'application du modèle sur les averses des données de Marseille a nécessité le choix d'un seuil au delà duquel les averses peuvent être simplement modélisées par un processus markovien d'ordre 1. Une modélisation markovienne d'ordre supérieur pourrait peut-être apporter une meilleure modélisation de la persistance et éviter la définition du seuil nécessaire dans le cas de la modélisation markovienne au premier ordre.

Bibliographie

Acreman, M. (1990). A simple stochastic model of hourly rainfall of Farnborough, Engeland. *Hydrological Sciences Journal-Journal Des Sciences Hydrologiques*, 35:119–148.

Akaike, H. (1974). A new look at statistical model identification. *IEEE Trans. Auto. Control*, 19:716–722.

Ancona-Navarrete, M. et Tawn, J. A. (2000). A comparison of methods for estimating the extremal index. *Extremes*, 3(1):5–38.

Angus, J. (1993). Asymptotic theory for bootstrapping the extremes. *Comm. Statist. Theory Methods*, 22(1):19–30.

Arnaud, P. (1997). *Modèle de prédetermination de crues basé sur la simulation stochastique des pluies horaires*. Thèse de doctorat, Montpellier II.

Arnaud, P. (2004). Extension en métropole de la méthode Shypre, adaptation du modèle de pluie. Rapport technique, Cemagref.

Arnaud, P., Finé, J. et Lavabre, J. (2003). Cartographie des pluies à différentes échelles de temps et de fréquence d'apparition. Rapport technique, Engees, Météo-France, Cemagref.

Arnaud, P., Lang, M. et Lavabre, J. (1998). Comparaison des méthodes Shypre, Agregee et Qdf. *Rapport Cemagref , Programme Fédérateur Risque*.

Ashkar, F. et Bobee, B. (1988). Confidence intervals for flood events under a Pearson 3 or log Pearson 3 distribution. *American Water Resources association*, 24(3):639–650.

Athreya, K. B. et Fukuchi, J. (1997). Confidence intervals for endpoints of a cdf via bootstrap. *Journal of Statistical Planning and Inference*, 58(2):299–320.

Bâ, K., Diaz-Delgado, C. et Cârsteanu, A. (2001). Confidence intervals of quantiles in hydrology computed by an analytical method. *Natural Hazards*, 24:1–12.

Bacro, J. (2005). Dépendance des extrêmes. Rapport technique.

Bacro, J. et Brito, M. (1998). A tail bootstrap procedure for estimating the tail Pareto-index. *Journal of Statistical Planning and Inference*, 71:245–260.

Bacro, J. et Chaouche, A. (2006). Incertitude d'estimation des pluies extrêmes du pourtour méditerranéen : illustration par les données de Marseille. *Hydrological Sciences Journal-Journal Des Sciences Hydrologiques*, 51(3):389–405.

Barakat, H., Nigm, E. et Askar, M. (2004). On the trivariate extreme value distributions. *Journal of Statistical Planning and Inference*, 118(1-2):19–35.

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Phil. Trans. Roy. Soc.*, 53:370–418.

Beirlant, J., Goegebeur, Y., Segers, J. et Teugels, J. (2004). *Statistics of Extremes, Theory and Applications*. Wiley series in probability and statistics.

Bernard, M. (1932). Formulas for rainfall intensities of long durations. *Trans. ASCE*, 96:592–624.

Bernardara, P. (2004). *Scale invariance in raifall fields modeling*. Thèse de doctorat, Milan.

Beven, K. (1987). Towards the use of a catchement geomorphology in flood frequency prediction. *Earth Surface Processes and Landforms*, 12:69–82.

Beven, K. et Binley, A. (1992). The future of distributed models : model calibration and uncertainty prediction. *Hydrological Processes*, 6:279–298.

Beven, K. et Freer, J. (2001). Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the glue methodology. *Journal of Hydrology*, 249:11–29.

Blazkova, S. et Beven, K. (1997). Flood frequency prediction for data limited catchments in the Czech republic using a stochastic rainfall model and TOPMODEL. *Journal of Hydrology*, 195(1-4):256–278.

Bo, Z., Islam, S. et Eltahir, E. (1994). Aggregation-disaggregation properties of a stochastic rainfall model. *Water Resour. Res.*, 30(12):3423–3435.

Borga, M., Vezzani, C. et Dalla Fontana, G. (2005). Regional rainfall depth-duration-frequency equations for an alpine region. *Natural Hazards*, 36(1-2):221–235.

Bortot, P. et Tawn, J. A. (1998). Models for the extremes of Markov chains. *Biometrika*, 85(4):851–867.

Brandsma, T. et Buishand, T. (1998). Simulation of extreme precipitation in the Rhine basin by nearest-neighbour resampling. *Hydrol. Earth Syst. Sci.*, 1(2-3):195–209.

Buishand, T. (1977). *Stochastic modelling of daily rainfall sequences*. Mededelingen Landbouwhogeschool. Wageningen, Nederland.

Buishand, T. (1978). Some remarks on the use of daily rainfall models. *Journal of Hydrology*, 36:295–308.

Buishand, T. et Brandsma, T. (2001). Multi-site simulation of daily precipitation and temperature in the Rhine basin by nearest-neighbour resampling. *Water Resour. Res.*, 37:2761–2776.

Burlando, P. et Rosso, R. (1996). Scaling and multiscaling models of depth-duration-frequency curves for storm precipitation. *Journal of Hydrology*, 187(1-2):45–64.

Burn, D. H. (2003). The use of resampling for estimating confidence intervals for single site and pooled frequency analysis. *Hydrological Sciences Journal-Journal Des Sciences Hydrologiques*, 48(1):25–38.

Burnham, K. et Anderson, D. (2004). Multimodel inference : understanding AIC and BIC in model selection. *In Workshop on Model Selection*, Universiteit van Amsterdam.

Cahill, A. T. (2003). Significance of AIC differences for precipitation intensity distributions. *Advances in Water Resources*, 26(4):457–464.

Cameron, D., Beven, K. et Tawn, J. (2000a). An evaluation of three stochastic rainfall models. *Journal of Hydrology*, 228(1-2):130–149.

Cameron, D., Beven, K. et Tawn, J. (2000b). Modelling extreme rainfalls using a modified random pulse Bartlett-Lewis stochastic rainfall model (with uncertainty). *Advances in Water Resources*, 24(2):203–211.

Cameron, D., Beven, K., Tawn, J., Blazkova, S. et Naden, P. (1999). Flood frequency estimation by continuous simulation for a gauged upland catchement (with uncertainty). *Journal of Hydrology*, 219:169–187.

Cameron, D., Beven, K., Tawn, J. et Naden, P. (2000c). Flood frequency estimation by continuous simulation (with likelihood based uncertainty estimation). *Hydrology and Earth System Sciences*, 4(1):23–34.

Capéraà, P., Fougères, A. et Genest, C. (1997). A non parameteric estimation procedure for bivariate extreme value copulas. *Biometrika*, 84:567–577.

Cernesson, F. (1993). *Modèle simple de prédétermination des crues de fréquences courantes à rares sur petits bassins versants méditerranéens.* Thèse de doctorat, Université Montpellier 2.

CFGB (1994). Les crues de projet des barrages : méthode du Gradex. Design flood determination by the Gradex method. *In* Barrages, B. d. C. F. d. G., éditeur : *18éme congrès CIGB-ICOLD*, volume 2.

Chaouche, A. et Parent, E. (1999). Inférence et validation bayesiennes d'un modèle de pluie journalière en régime de mousson. *Hydrological Sciences Journal-Journal Des Sciences Hydrologiques*, 44(2):199–220.

Chaouche, K., Hubert, P. et Lang, G. (2002). Graphical characterisation of probability distribution tails. *Stochastic Environmental Research and Risk Assessment*, 16(5):342–357.

Chapman, T. (1998). Stochastic modelling of daily rainfall : the impact of adjoining wet days on the distribution of rainfall amounts. *Environmental Modelling and Software*, 13(3-4):317–324.

Chapman, T. G. (1997). Stochastic models for daily rainfall in the western Pacific. *Mathematics and Computers in Simulation*, 43(3-6):351–358.

Chib, S. et Greenberg, E. (1995). Understanding the Metropolis-Hastings Algorithm. *The American Statistician*, 49(4):327–335.

Chin, E. et Miller, J. (1980). On the conditionnal distribution of daily precipitation amounts. *Monthly Weather Review*, 108:1462–1464.

Chow, V., Maidment, D. et Mays, L. (1988). *Applied hydrology*. McGraw-Hill.

Christopeit, N. (1994). Estimating parameters of an extreme value distribution by the method of moments. *Journal of Statistical Planning and Inference*, 41:173–186.

Coles, S. (1994). A temporal study of extreme rainfall. *In Vic Bartnett, K. Feridun Turkman*, Statistics for the Environment 2 : Water related Issues.

Coles, S. (2001). *An introduction to statistical modeling of extreme values*. Springer-Verlag. London.

Coles, S., Heffernan, J. et Tawn, J. A. (2002). Dependence measures for extreme value analyses. *Extremes*, 2(4):339–365.

Coles, S. et Pauli, F. (2002). Models and inference for uncertainty in extremal dependence. *Biometrika*, 89(1):183–196.

Coles, S. et Pericchi, L. (2003). Anticipating catastrophes through extreme value modelling. *Journal of the Royal Statistical Society. Series C : Applied Statistics*, 52(4):405–416.

Coles, S., Pericchi, L. et Sisson, S. (2003). A fully probabilistic approach to extreme rainfall modelling. *J. Hydrol.*, 273(1-4):35–50.

Coles, S. et Tawn, J. (1994). Statistical methods for multivariate extremes : an application to structural design (with discussion). *Applied Statistics.*, 43:1–48.

Coles, S. et Tawn, J. (1996). A bayesian analysis of extreme rainfall data. *Applied Statistics.*, 45(4):463–478.

Coles, S. et Tawn, J. A. (1991). Modelling extreme multivariate events. *J. R. Statist. Soc. B.*, 53:377–392.

Colin, E. (1977). Quelques applications sur la distribution d'échantillonnage. Rapport technique Etude num 21, Cemagref Antony, nov, 21p.

Cowpertwait, P. S. P. (1998). A poisson-cluster model of rainfall : high-order moments and extreme values. *Proc. R. Soc. Lond. Series A : Mathematical and Physical Sciences*, 454:885–898.

Cowpertwait, P. S. P., Kilsby, C. G. et O'Connell, P. E. (2002). A space-time Neyman-Scott model of rainfall : Empirical analysis of extremes. *Water Resources Research*, 38(8).

Cowpertwait, P. S. P. et O'Connell, P. E. (1997). A regionalised Neyman-Scott model of rainfall with convective and stratiform cells. *Hydrol. Earth Syst. Sci.*, 1:71–80.

Cowpertwait, P. S. P., O'Connell, P. E., Metcalfe, A. et Mawdsley, J. (1996). Stochastic point process modelling of rainfall. 2. Regionalisation and disaggregation. *Journal of Hydrology*, 175:47–65.

Cunnane, C. (1978). Unbiased plotting position -a review. *Journal of Hydrology*, 37:205–222.

Davison, A., Hinkley, D. et Schechtman, E. (1986). Efficient bootstrap simulation. *Biometrika*, 73:555–566.

de Haan, L. (1970). *On regular variation and its applications to weak convergence of sample extremes*, volume 32 de *CWI Tract*. Amsterdam.

de Lima, M. I. P. et Grasman, J. (1999). Multifractal analysis of 15-min and daily rainfall from a semi-arid region in Portugal. *Journal of Hydrology*, 220(1-2):1–11.

Deheuvels, P., Masson, D. et Shorack, G. (1993). Some results on the influence of the extreme on bootstrap. *Ann. Inst. H. Poincaré*, 29:83–103.

Dekkers, A., Einmahl, J. et de Haan, L. (1989). A moment estimator for the index of an extreme-value distribution. *Ann. Statist.*, 17:1795–1832.

Diaz-Granados, M., Valdes, J. et Bras, R. (1984). A physically based flood frequency distribution. *Water Resour. Res.*, 20:995–1002.

Dubrovsky, M., Buchtele, J. et Zalud, Z. (2004). High-frequency and low-frequency variability in stochastic daily weather generator and its effect on agricultural and hydrologic modelling. *Climatic Change*, 63:145–179.

Eagleson, P. (1972). Dynamics of flood frequency. *Water Resour. Res.*, 8:878–898.

Efron, B. (1979). Bootstrap methods : Another look at the Jacknife. *Annals of Statistics*, 7:1–26.

Efron, B. (1981). Nonparametric standard errors and confidence intervals (with discussion). *Canad. J. Statist.*, 9:139–172.

Embrechts, P., Klüppelberg, C. et Mikosch, T. (1997). *Modelling Extremal Events*. Springer-Verlag.

Engeland, K., Xu, C. Y. et Gottschalk, L. (2005). Assessing uncertainties in a conceptual water balance model using bayesian methodology. *Hydrological Sciences Journal-Journal Des Sciences Hydrologiques*, 50(1):45–63.

Favre, A. C., Musy, A. et Morgenthaler, S. (2004). Unbiased parameter estimation of the Neyman-Scott model for rainfall simulation with related confidence interval. *Journal of Hydrology*, 286(1-4):168–178.

Ferro, C. et Segers, J. (2003). Inference for clusters of extreme values. *Journal of the Royal Statistical Society Series B*, 65:545–556.

Fisher, R. et Tippett, L. (1928). On the estimation of the frequency distributions of the largest or smallest member of a sample. *Proceedings of the Cambridge Philosophical Society*, 24:180–190.

Foufoula-Georgiou, E. et Lettenmeier, D. (1987). Markov renewal model for rainfall occurrence. *Water Resources Research*, 23:875–884.

Freer, J., Beven, K. et Ambroise, B. (1996). Bayesian estimation of uncertainty in runoff prediction and the value of data : An application of the GLUE approach. *Water Resources Research*, 32(7):2161–2173.

Frost, A., Srikanthan, R. et Cowpertwait, P. S. P. (2005). Stochastic generation of point rainfall data at sub-daily timescales : a comparison of DRIP and NSRP. *In International congress on modelling and simulation society of Australia and New Zealand*, Melbourne.

Galambos, J. (1975). Order statistics of samples from multivariate distributions. *J. Amer. Statist. Assoc.*, 70:674–680.

Garcia-Bartual, R. et Schneider, M. (2001). Estimating maximum expected short-duration rainfall intensities from extreme convective storms. *Physics and Chemistry of the Earth, Part B : Hydrology, Oceans and Atmosphere*, 26(9):675–681.

Gaume, E., Villeneuve, J. et Desbordes, M. (1998). Uncertainty assessment and analysis of the calibrated parameter values of an urban storm water quality model. *Journal of Hydrology*, 210:39–50.

Gelman, A., Carlin, J., Stren, H. et Rubin, D. (1997). *Bayesian Data Analysis*. Chapman and Hall. London.

Geyer, C. (1992). Practical Markov Chain Monte Carlo. *Statistical Science*, 7(4).

Gnedenko, B. (1943). Sur la distribution limite du terme maximum d'une série aléatoire. *Annals of Mathematics*, 44:423–453.

Greenwood, J., Landwehr, J., Matalas, N. et Wallis, J. (1979). Probability weighted moments : definition and relation to parameters of several distributions expressable in inverse form. *Water Resources Research*, 15:1049–1054.

Gregory, J., Wigley, T. et Jones, P. (1993). Application of Markov models to area-average daily precipitation series and interannual variability in seasonal totals. *Climate Dynamics*, 8:299–310.

Guillot, P. et Duband, D. (1967). La méthode du Gradex pour le calcul de la probabilité des crues à partir des pluies. *In* Colins, F., éditeur : *Colloque International sur les crues et leur évaluation*, volume publication n 84, 560-56, Leningrad. IASH.

Gumbel, E. (1960). Bivariate exponential distributions. *J. Amer. Statist. Assoc.*, 55:698–707.

Guntner, A., Olsson, J., Calver, A. et Gannon, B. (2001). Cascade-based disaggregation of continuous rainfall time series : the influence of climate. *Hydrology and Earth System Sciences*, 5(2):145–164.

Guo, S. (1990). A discussion on unbiased plotting positions for the extreme value distribution. *Journal of Hydrology*, 121(1-4):33–44.

Gupta, V. et Waymire, E. (1993). A statistical analysis of mesoscale rainfall as a random cascade. *Journal of Applied Meteorology*, 32:251–267.

Guttorp, P. (1995). *Stochastic modeling of scientific data*. Chapman and Hall. London.

Gyasi-Agyei, Y. (2005). Stochastic disaggregation of daily rainfall into one-hour time scale. *Journal of Hydrology*, 309(1-4):178–190.

Haan, C., Allen, D. et Street, J. (1976). A Markov chain model of daily rainfall. *Water Resour. Res.*, 12(3):443–449.

Harrold, T., Sharma, A. et Sheather, S. (2003a). A nonparametric model for stochastic generation of daily rainfall occurrence. *Water Resources Research*, 39(10):SWC101–SWC1011.

Harrold, T. I., Sharma, A. et Sheather, S. J. (2003b). A nonparametric model for stochastic generation of daily rainfall amounts. *Water Resources Research*, 39(12).

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

Hebson, C. et Wood, E. (1982). A derived flood frequency distribution. *Water Resour. Res.*, 18:1509–1518.

Heffernan, J. et Resnick, S. (2005). Hidden regular variation and the rank transform. *Advances in Applied Probability*, 37(2):393–414.

Heneker, T. M., Lambert, M. F. et Kuczera, G. (2001). A point rainfall model for risk-based design. *Journal of Hydrology*, 247(1-2):54–71.

Hershfield, D. (1961). Rainfall frequency atlas of the United States for durations from 30 minutes to 24 hours and return periods from 1 to 100 years. Rapport technique, Weather Bureau Technical Paper 40, U.S. Department of Commerce. Washington D.C.

Hill, B. (1975). A simple general approach to inference about the tail of a distribution. *Ann. Statist.*, 3:1163–1174.

Hornberger, G. et Spear, R. (1981). An approach to the preliminary analysis of environmental systems. *Journal of Environmental Management*, 12:7–18.

Hosking, J. et Wallis, J. (1987). Parameter and quantile estimation for the Generalized Pareto Distribution. *Technometrics*, 29(3):339–349.

Hosking, J., Wallis, J. et Wood, E. (1985). Estimation of the Generalized Extreme Value Distribution by the method of probability - weighted moments. *Technometrics*, 27(3):251–261.

Hossain, F. et Anagnostou, E. (2005). Assessment of a stochastic interpolation based parameter sampling scheme for efficient uncertainty analyses of hydrologic models. *Computers and Geosciences*, 31(4):497–512.

Hsing, T., Hüsler, J. et Leadbetter, M. (1988). On the exceedance point process for a stationary sequence. *Probab. Theory Reltd Flds*, 78:97–112.

Hubert, P., Tessier, Y., Lovejoy, S., Shertzer, D., Schmitt, F., Ladoy, P., Carbonell, J., Violette, S. et Desurosne, I. (1993). Multifractals and extreme rainfall events. *Geophys. Res. lett.*, 20(10):931–934.

Husler, J. et Reiss, R. (1989). Maxima of normal random vectors : between independence and complete dependence. *Statist. Probab. Letters*, 7:283–286.

Joe, H. (1990). Families of min-stable multivariate exponential and multivariate extreme value distributions. *Statist. Probab. Letters*, 9:75–81.

Joe, H. (1993). Parametric families of multivariate distributions with given margins. *Journal of Multivariate Analysis*, 46(2):262–282.

Joe, H., Smith, R. et Weissman, I. (1992). Bivariate threshold methods for extremes. *J. R. Statist. Soc. B.*, 54:171–183.

Katz, R. (1977). Precipitation as a chain-dependent process. *Journal of Applied Meteorology*, 16:671–676.

Katz, R. et Parlange, M. (1995). Generalizations of chain-dependent processes : application to hourly precipitation. *Water Resour. Res.*, 31:1331–1341.

Katz, R. et Parlange, M. (1998). Overdispersion phenomenon in stochastic modeling of precipitation. *Journal of Climate*, 11:591–601.

Keesman, K. et Van Straten, G. (1990). Set-membership approach to identification and prediction of lake eutrophication. *Water Resources Research*, 262643-2652.

Kendall, M. (1975). *Rank correlation methods*. London.

Kendall, M. et Stuart, A. (1987). *The advanced theory of statistics*, volume 1 de *Charles Griffin*. London.

Kendall, M., Stuart, A. et Ord, J. (1983). *The advanced theory of statistics*, volume 3 de *Charles Griffin*. London.

Khaliq, M. N. et Cunnane, C. (1996). Modelling point rainfall occurrences with the modified Bartlett-Lewis rectangular pulses model. *Journal of Hydrology*, 180(1-4):109–138.

Kieffer Weisse, A. (1998). *Etude des précipitations exceptionnelles de pas de temps court en relief accidenté (Alpes françaises). Méthode de cartographie des précipitations extrêmes.* Thèse de doctorat, Institut National Polytechnique de Grenoble.

Kite, G. (1975). Confidence limits for design events. *Water Resour. Res.*, 11(1):48–53.

Klepper, O. et Hendrix, E. (1994). A method for robust calibration of ecological models under different types of uncertainty. *Ecological Modelling*, 74:161–182.

Koutsoyiannis, D. (2004a). Statistics of extremes and estimation of extreme rainfall : 1. Theoretical investigation. *Hydrological Sciences Journal-Journal Des Sciences Hydrologiques*, 49(4):575–590.

Koutsoyiannis, D. (2004b). Statistics of extremes and estimation of extreme rainfall : 2. Empirical investigation of long rainfall records. *Hydrological Sciences Journal-Journal Des Sciences Hydrologiques*, 49(4):591–610.

Koutsoyiannis, D. (2006). An entropic-stochastic representation of rainfall intermittency : the origin of clustering and persistence. *Water Resour. Res.*, 42.

Koutsoyiannis, D. et Baloutsos, G. (2000). Analysis of a long record of annual maximum rainfall in Athens, Greece, and design rainfall inferences. *Natural Hazards*, 22(1):31–51.

Koutsoyiannis, D., Kozonis, D. et Manetas, A. (1998). A comprehensive study of rainfall intensity-duration-frequency relationships. *Journal of Hydrology*, 206(1-2):118–135.

Koutsoyiannis, D. et Onof, C. (2001). Rainfall disaggregation using adjusting procedures on a Poisson cluster model. *Journal of Hydrology*, 246(1-4):109–122.

Krzysztofowicz, R. (1999). Bayesian theory of probabilistic forecasting via deterministic hydrologic model. *Water Resources Research*, 35(9):2739–2750.

Kuczera, G. et Parent, E. (1998). Monte Carlo assessment of parameter uncertainty in conceptual catchment models : the Metropolis algorithm. *Journal of Hydrology*, 211:69–85.

Lall, U., Rajagopalan, B. et Tarboton, D. G. (1996). A nonparametric wet/dry spell model for resampling daily precipitation. *Water Resources Research*, 32(9):2803–2823.

Lall, U. et Sharma, A. (1996). A nearest-neighbor bootstrap for resampling hydrologic time series. *Water Resour. Res.*, 32:679–693.

Lang, M. (1995). *Les chroniques en hydrologie : modélisation comparée par un système de gestion de bases de données relationnem et orienté-objet. Traitements de base et intervalles de confiance des quantiles de crue. Techniques d'échantillonnage par la méthode du renouvellement.* Thèse de doctorat, Joseph Fourier.

Lang, M. (1997). New developments with Agregee, a statistical model using hydrometeorological information. *In G. Oberlin, E. Desbos.* FRIEND projects H-5-5 et H-1-1, third report 1994-1997.

Lavallee, D., Schertzer, D. et Lovejoy, S. (1991). On the determination of the codimension function. *In Scaling, fractals and nonlinear variability in geophysics*, pages 99–109. eds. D. Schertzer and S. Lovejoy, Kluwer press.

Le Cam, L. (1953). On some asymptotic properties of maximum likelihood estimates and related bayes estimates. *Univerity of California, Publications in Statistics*, 1(11):277–330.

Leadbetter, M. (1974). On extreme values in stationnary sequences. *Zeitschrift für Wahrscheinlichkeitstheorie and Verwandte Gebiete*, 28:289–303.

Leadbetter, M. (1983). Extremes and local dependence in stationary sequences. *Zeit. Wahrscheinl. -theorie*, 65:291.

Leadbetter, M. (1991). On a basis for 'peak over threshold' modeling. *Statist. Probab. Letters*, 12:357–362.

Leadbetter, M., Lindgren, G. et Rootzen, H. (1983). *Extremes and related properties of random sequences and series.* New York.

Lebarbier, E. et Mary-Huard, T. (2004). Le critère BIC : fondements théoriques et interprétation. Rapport technique, INRIA.

Lebel, T. (1984). *Moyenne spatiale de la pluie sur un bassin versant : estimation optimale, génération stochastique et Gradex des valeurs extrêmes.* Thèse de doctorat, Institut National Polytechnique de Grenoble.

Ledford, A. W. et Tawn, J. A. (1996). Statistics for near independence in multivariate extreme values. *Biometrika*, 83(1):169–187.

Ledford, A. W. et Tawn, J. A. (1997). Modelling dependence within joint tail regions. *Journal of the Royal Statistical Society Series B-Methodological*, 59(2):475–499.

Ledford, A. W. et Tawn, J. A. (2003). Diagnostics for dependence within time series extremes. *Journal of the Royal Statistical Society Series B-Statistical Methodology*, 65:521–543. Part 2.

Lee, S. M. S. (2000). Nonparametric confidence intervals based on extreme bootstrap percentiles. *Statistica Sinica*, 10(2):475–496.

Lu, L. et Stedinger, J. R. (1992). Variance of two- and three-parameter GEV/PWM quantile estimators : formulae, confidence intervals, and a comparison. *Journal of Hydrology*, 138:247–267.

Lye, L., Hapuarachchi, K. et Ryan, S. (1993). Bayes estimation of the extrema-value reliability function. *IEEE Transactions on Reliability*, 42:641–644.

Madden, R., Shea, D., Katz, R. et Kidson, J. (1999). The potential long-range predictability of precipitation over New Zealand. *International Journal of Climatology*, 19(4):405–421.

Mann, H. (1945). Nonparametric tests against trend. *Econometrica*, 13:245–259.

Margoum, M., Oberlin, G., Lang, M. et Weingartner, R. (1994). Estimation des crues rares et extrêmes : principes du modèle Agregee. *Hydrol. continent.*, 9(1):85–100.

Marshall, L., Nott, D. et Sharma, A. (2004). A comparative study of Markov chain Monte Carlo methods for conceptual rainfall-runoff modeling. *Water Resources Research*, 40(2).

Menabde, M. et Sivapalan, M. (2000). Modeling of rainfall time series and extremes using bounded random cascades and Levy-stable distributions. *Water Resources Research*, 36(11):3293–3300.

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

Metropolis, N. et Ulam, S. (1949). The Monte Carlo method. *Journal of the American Statistical Association*, 44:335–341.

Mo, X. G., Pappenberger, F., Beven, K., Liu, S. X., De Roo, A. et Lin, Z. H. (2006). Parameter conditioning and prediction uncertainties of the LISFLOOD-WB distributed hydrological model. *Hydrological Sciences Journal-Journal Des Sciences Hydrologiques*, 51(1):45–65.

Molnar, P. et Burlando, P. (2005). Preservation of rainfall properties in stochastic disaggregation by a simple random cascade model. *Atmospheric Research*, 77:137–151.

Montanari, A. et Brath, A. (2004). A stochastic approach for assessing the uncertainty of rainfall-runoff simulations. *Water Resources Research*, 40(1):W011061–W0110611.

Moran, P. (1957). The statistical treatment of floods. *Trans. of the American Geophysical Union*, 38(4):519–523.

Morgenstern, D. (1956). Einfache Beispiele zweidimensionaler Verteilungen. *Mitt. Math. Statist.*, 8:234–235.

Moughamian, M., MCLaughlin, D. et Bras, R. (1987). Estimation of flood frequency : an evaluation of two derived distribution procedures. *Water Resour. Res.*, 23:1309–1319.

Mouhous, N. (2003). *Intérêt des modèles de cascades multiplicatives pour la simulation de séries chronologiques de pluies ponctuelles adaptées à l'hydrologie urbaine.* Phd, Ecole Nationale des Ponts et Chaussées.

Mousavi Nadoshani, S. (1997). *Composition des lois élémentaires en hydrologie régionale : application à l'étude des régimes de crues.* Thèse de doctorat, Université Joseph Fourier, Grenoble 1.

Muller, A., Bacro, J. et Lang, M. (2006). Bayesian comparison of different rainfall depth-duration-frequency relationships. *Stochastic Environmental Research and Risk Assessment, Accepted.*

Nadarajah, S., Anderson, C. et Tawn, J. A. (1998). Ordered multivariate extremes. *J.R. Statistic. Soc. B*, 60(2):473–496.

Naulet, R. (2002). Eléments de statistiques appliquées à l'hydrologie. Rapport technique, Cemagref.

Neyman, J. et Scott, E. (1952). A theory of the spatial distribution of galaxies. *Astrophys. J.*, 116:144–163.

O'Brien, G. (1987). Extreme values for stationary and Markov chains sequences. *Ann. Probab.*, 15:281–291.

Olsson, J. (1995). Limits and characteristics of the multifractal behavior of a high-resolution rainfall time series. *Nonlinear Processes in Geophysics*, (2):23–29.

Olsson, J. (1998). Evaluation of a cascade model for a temporal rainfall disaggregation. *Hydrol. Earth Syst. Sci.*, 2:19–30.

Olsson, J. et Burlando, P. (2002). Reproduction of temporal scaling by a rectangular pulses rainfall model. *Hydrological Processes*, 16(3):611–630.

Olsson, J., Niemczynowicz, J. et Berndtsson, R. (1993). Fractal analysis of high-resolution rainfall time series. *J. Geophys. Res.*, 98(D12):23265–23274.

Onof, C. et Wheater, H. S. (1994). Improvements to the modeling of British rainfall using a modified random parameter Bartlett-Lewis Rectangular Pulse Model. *Journal of Hydrology*, 157(1-4):177–195.

Over, T. et Gupta, V. K. (1994). Statistical analysis of mesoscale rainfall : dependence of a random cascade generator on large scale forcing. *Journal of Applied Meteorology*, 33:1526–1542.

Over, T. et Gupta, V. K. (1996). A space-time theory of mesoscale rainfall using random cascades. *Journal of geophysical Research*, 101(D21):26319–26331.

Pandey, G., Lovejoy, S. et Schertzer, D. (1998). Multifractal analysis of daily river flows including extremes for basins of five to two million square kilometres, one day to 75 years. *Journal of Hydrology*, 208(1-2):62–81.

Pavlopoulos, H. et Gupta, V. (2003). Scale invariance of regional wet and dry duration of rain fields : a diagnostic study. *J. Geophys. Res.*, 108(D8).

Perfekt, R. (1994). Extremal behaviour of stationary Markov chains with applications. *Ann. Appl. Prob.*, 30:197–215.

Pickands, J. (1975). Statistical inference using extreme order statictic. *Ann. Statist.*, 3:119–131.

Pickands, J. (1981). Multivariate extreme value distributions. *In Bulletin of the International Statistical Institute, Proceedings of the 43rd Session*, pages 859–878, Buenos Aires.

Pickands, J. (1989). Multivariate negative exponential and extreme value distributions. *In Extreme Value Theory : Proceedings*, pages 262–274, Oberwolfach.

Press, W., Flannery, B., Teukolsky, S. et Vetterling, W. (1987). *Numerical Recipes : the art of scientific computing*. Cambridge.

Racsko, P., Szeidl, L. et Semenov, M. (1991). A serial approach to local stochastic weather models. *Ecological Modelling*, 57:27–41.

Raftery, A. (1994). Bayesian model selection in social research (with discussion). Rapport technique, University of Washsington Demography Center Working. A revised version appeared in Sociological Methodology 1995.

Rajagopalan, B. et Lall, U. (1995). A nearest-neighbor bootstrap for resampling daily precipitation and other weather variables. Rapport technique, Working Paper WP 95 HWR UL/013, Utah State University.

Rajagopalan, B., Lall, U. et Tarboton, D. G. (1996). A nonhomogeneous Markov model for daily precipitation simulation. *J. Hydrol. Eng. -ASCE*, 1:33–40.

Rajagopalan, B., Lall, U., Tarboton, D. G. et Bowles, D. (1997). Multivariate nonparametric resampling sheme for generation of daily weather variables. *Stoch. Hydrol. Hydraul.*, 11(65-93).

Renard, B., Garreta, V. et Lang, M. (2006). An empirical comparison of MCMC methods used in bayesian inference. Application for regional trend detection. *Water Resources Research, in press.*

Resnick, S. (1987). *Extremes values, regular variation and point processes*. New-York.

Ribatet, M. (2006). The POT Package. *http ://cran.r-project.org/doc/packages/POT.pdf.*

Robinson, M. E. et Tawn, J. A. (2000). Extremal analysis of processes sampled at different frequencies. *Journal of the Royal Statistical Society Series B-Statistical Methodology*, 62: 117–135. Part 1.

Rodriguez-Iturbe, Cox, I. (1987). Some models for rainfall based on stochastic point process. *Proc. R. Soc. Lond.*, A 410:269–288.

Rodriguez-Iturbe, Cox, I. (1988). A point rainfall process model for rainfall : further developments. *Proc. R. Soc. Lond.*, A 417:283–298.

Roldan, J. et Woolhiser, D. (1982). Stochastic daily precipitation models. 1. A comparison of occurrence processes. *Water Resour. Res.*, 18:1451–1549.

Romanowicz, R., Beven, K. et Tawn, J. (1996). Bayesian calibration of flood inundation models. *In* Malcolm G. Anderson, Des E. Walling, P. D. B., éditeur : *Floodpalin Processes*, pages 333–360. John Wiley and Sons Ltd.

Salvadori, G. et De Michele, C. (2006). Statistical characterization of temporal structure of storms. *Advances in Water Resources*, 29:827–842.

Saporta, G. (1990). *Probabilités, analyse des données et statistiques*. Editions Technip.

Schertzer, D. et Lovejoy, S. (1987). Physical modeling and analysis of rain and clouds by anisotropic scaling multiplicative processes. *Journal of geophysical Research*, 92(D8):9693 – 9714.

Schucany, W., Parr, W. et Boyer, J. (1978). Correlation structure in Farlie-Gumbel-Morgenstern distribution. *Biometrika*, 65(3):650–653.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Stat.*, 6:461–464.

Seong, K. et Lee, Y. (2003). Two practical approaches for determining the confidence interval of rainfall depth percentiles. *Journal of the American Water Ressources Association*, 39(2):369–380.

Sharma, A. et Lall, U. (1999). A nonparametric approach for daily rainfall simulation. *Mathematics and Computers in Simulation*, 48(4-6):361–371.

Sharma, A. et O'Neill, R. (2002). A nonparametric approach for representing interannual dependence in monthly streamflow sequences. *Water Resour. Res.*, 38(7).

Sharma, T. (1996). Simulation of the longest dry and wet spells and the largest rain-sums using a Markov model. *Journal of Hydrology*, 178:55–67.

Sisson, S. et Coles, S. (2003). Modelling dependence uncertainty in the extremes of Markov chains. *Extremes*, 6(4):283–300.

Sisson, S. A., Pericchi, L. R. et Coles, S. G. (2006). A case for a reassessment of the risks of extreme hydrological hazards in the Caribbean. *Stochastic Environmental Research and Risk Assessment*, 20(4):296–306.

Small, M. et Morgan, D. (1986). The relationship between a continuous-time renewal model and a discrete Markov chain model of precipitation occurrence. *Water Resour. Res.*, 22:1422–1430.

Smith, R. (1990). Extreme value theory. *In* Ledermann, W., éditeur : *Handbook of applicable mathematics*, volume 7, pages 437–471.

Smith, R. (1992). The extremal index for a Markov chain. *J. Appl. Prob.*, 29:37–45.

Smith, R., Tawn, J. A. et Coles, S. (1997). Markov chain models for threshold exceedances. *Biometrika*, 84(2):249–268.

Smith, R. et Weissman, I. (1994). Estimating the extremal index. *J. R. Statist. Soc. B.*, 56:515–528.

Smithers, J. C., Pegram, G. G. S. et Schulze, R. E. (2002). Design rainfall estimation in South Africa using Bartlett-Lewis rectangular pulse rainfall models. *Journal of Hydrology*, 258(1-4):83–99.

Spear, R. et Hornberger, G. (1980). Eutrophication in peel inlet.ii identification of critical uncertainties via generalised sensitivity analysis. *Water Resources Research*, 14:43–49.

Srikanthan, R. et McMahon, T. (1985). Stochastic generation of rainfall and evaporation data. Rapport technique.

Srikanthan, R. et McMahon, T. A. (2001). Stochastic generation of annual, monthly and daily climate data : A review. *Hydrology and Earth System Sciences*, 5(4):653–670.

Stedinger, J. R. (1983). Confidence intervals for design events. *J. Hydraul. Div.*, 109:13–27.

Stern, R. et Coe, R. (1984). A model fitting analysis of daily rainfall data. *R. Statist. Soc.*, 147:1–34.

Storm, B., Jensen, K. et Refsgaard, J. C. (1988). Estimation of catchment rainfall uncertainty and its influence on runoff predictions. *Nordic Hydrol.*, 19:77–88.

Svensson, C., Olsson, J. et Berndtsson, R. (1996). Multifractal properties of daily rainfall in two different climates. *Water Resources Research*, 32(8):2463–2472.

Swanepoel, J. (1986). A note on proving that the (modified) bootstrap method works. *Comm. Statist. Theory Methods*, 15(11):3193–3203.

Tanner, M. (1992). *Tools for statistical inference : observed data and data augmentation methods.* Lecture Notes in Statistics 67. New-York.

Tarboton, D. G., Sharma, A. et Lall, U. (1998). Disaggregation procedures for stochastic hydrology based on nonparametric density estimation. *Water Resources Research*, 34(1): 107–119.

Tawn, J. (1988). Bivariate extreme value theory : models and estimation. *Biometrika*, 75:397–415.

Tessier, Y., Lovejoy, S., Hubert, P., Schertzer, D. et Pecknold, S. (1996). Multifractal analysis and modeling of rainfall and river flows and scaling, causal transfer functions. *Journal of Geophysical Research-Atmospheres*, 101(D21):26427–26440.

Tessier, Y., Lovejoy, S. et Schertzer, D. (1993). Universal multifractals - theory and observations for rain and clouds. *Journal of Applied Meteorology*, 32(2):223–250.

Thorsen, M., Refsgaard, J. C., Hansen, S., Pebesma, E., Jensen, J. B. et Kleeschulte, S. (2001). Assessment of uncertainty in simulation of nitrate leaching to aquifers at catchment scale. *Journal of Hydrology*, 242(3-4):210–227.

Thyer, M. et Kuczera, G. (2000). Modeling long-term persistence in hydroclimatic time series using a hidden state Markov model. *Water Resources Research*, 36(11):3301–3310.

Thyer, M. et Kuczera, G. (2003). A hidden Markov model for modelling long-term persistence in multi-site rainfall time series. 1. Model calibration using a bayesian approach. *Journal of Hydrology*, 275:12–26.

Thyer, M., Kuczera, G. et Wang, Q. J. (2002). Quantifying parameter uncertainty in stochastic models using the Box-Cox transformation. *Journal of Hydrology*, 265(1-4):246–257.

Tourasse, P. (1981). *Analyses spatiales et temporelles des précipitations et utilisation opérationnelle dans un système de prévision des crues. Applications aux régions cévenoles.* Thèse de doctorat, IMG Univ. Scientifique et Medical, Institut National Polytechnique de Grenoble.

van Montfort, M. A. J. (1997). Concomitants of the Hershfield factor. *Journal of Hydrology*, 194(1-4):357–365.

Velghe, T., Troch, P. A., Detroch, F. P. et Vandevelde, J. (1994). Evaluation of cluster-based rectangular pulses point process models for rainfall. *Water Resources Research*, 30(10):2847–2857.

Veneziano, D., Bras, R. et Niemann, J. (1996). Nonlinearity and self-similarity of rainfall in time and a stochastic model. *J. Geophys. Res.*, 101(D21):26371–26392.

Veneziano, D. et Furcolo, P. (2002). Multifractality of rainfall and scaling of intensity-duration-frequency curves. *Water Resources Research*, 38(12):421–4212.

Venugopal, V. et Foufoula-Geoorgiou, E. (1996). Energy decomposition of rainfall in the time-frequency-scale domain using wavelet packets. *J. Hydrol.*, 187:3–27.

Verhoest, N., Troch, P. A. et De Troch, F. P. (1997). On the applicability of Bartlett-Lewis rectangular pulses models in the modeling of design storms at a point. *Journal of Hydrology*, 202(1-4):108–120.

Vrugt, J. A., Gupta, H. V., Bouten, W. et Sorooshian, S. (2003). A shuffled complex evolution Metropolis algorithm for optimization and uncertainty assessment of hydrologic model parameters. *Water Resources Research*, 39(8).

Wald, A. et Wolfowitz, J. (1943). An exact test for randomness in the non parametric case based on serial correlation. *Ann. Math. Stat.*, 14:378–388.

Wang et Nathan (2002). http ://www.toolkit.net.au/cgi-bin/webobjects/toolkit.woa/wa/downloadpublication ?id=1000014.

Weibull, W. (1939). *A statistical theory of strenght of materials.* Ing. Vet. Ak. Handl. Stockholm.

Weiss, L. (1964). Ratio of true to fixed-interval maximum rainfall. *Journal of the Hydraulic Division of ASCE*, 90:77–82.

Weissman, I. et Novak, S. (1998). On blocks and runs estimators of the extremal index. *Journal of Statistical Planning and Inference*, 66:281–288.

Wilby, R., Wigley, T., Conway, D., Jones, P., Hewitson, B., Main, J. et Wilks, D. (1998). Statistical downscaling of general circulation model output : a comparison of methods. *Water Resour. Res.*, 34:2995–3008.

Wilby, R., Wigley, T., Wilks, D., Hewitson, B., Conway, D. et Jones, P. (1996). Statistical downscaling of general circulation model output. Rapport technique, National Centr of Atmospheric Research.

Wilks, D. (1993). Comparison of three-parameter probability distributions for representing annual extreme and partial duration precipitation series. *Water Resour. Res.*, 29(10):3543–3549.

Wilks, D. (1998). Multisite generalization of a daily stochastic precipitation generation model. *Journal of Hydrology*, 210:178–191.

Wilks, D. (1999). Interannual variability and extreme-value characteristics of several stochastic daily precipitation models. *Agricultural and forest Meteorology*, 93:153–169.

Wojcik, R. et Buishand, T. (2003). Simulation of 6-hourly rainfall and temperature by two resampling schemes. *Journal of Hydrology*, 273:69–80.

Woolhiser, D. et Osbortn, H. (1985). A stochastic model of dimensionless thunderstorm rainfall. *Water Resour. Res.*, 21(4):511–522.

Woolhiser, D. et Roldan, J. (1982). Stochastic daily precipitation models. 2. A comparison of distribution amounts. *Water Resour. Res.*, 18:1461–1498.

Young, K. (1994). A multivariate chain model for simulating climatic parameters from daily data. *Journal of Applied Meteorology*, 33:661–671.

Young, P. (1978). General theory of modelling badly defined systems. *Modelling, Identification and Control in Environmental Systems*, pages 103–106.

Zelterman, D. (1993). A semiparametric bootstrap technique for simulating extreme order-statistics. *Journal of the American Statistical Association*, 88(422):477–485.

Zin, I. (2002). *Incertitudes et ambiguité dans la modélisation hydrologique*. Thèse de doctorat, INPG.

Zorita, E., Hughes, J., Lettenmeier, D. et von Storch, H. (1995). Stochastic characterization of regional circulation patterns for climate model diagnosis and estimation of local precipitation. *Journal of Climate*, 8:1023–1042.

Annexe : Article en cours de révision, pour le journal *Stochastic Environmental Research and Risk Assessment*

Bayesian comparison of different rainfall depth-duration-frequency relationships

Aurélie Muller [(1✉)], Jean-Noël Bacro [(2)], Michel Lang [(1)]

(1) Cemagref Centre de Lyon, U.R. Hydrologie-Hydraulique, 3 bis Quai Chauveau, CP 220, 69336 Lyon cedex 09, France

(2) Université Montpellier II, I3M, UMR CNRS 5149

Tel. : 33 4 72 20 87 72

Fax : 33 4 78 47 78 75

e-mail : muller@lyon.cemagref.fr

Abstract

Depth-Duration-Frequency curves estimate the rainfall intensity patterns for various return periods and rainfall durations. An empirical model based on the Generalized Extreme Value Distribution is presented for hourly maximum rainfall, and improved by the inclusion of daily maximum rainfall, through the extremal indexes of 24 hourly and daily rainfall data. The model is then divided into two sub-models for the short and long rainfall durations. Three likelihood formulations are proposed to model and compare independence or dependence hypotheses between the different durations. Dependence is modelled using the bivariate extreme logistic distribution. The results are calculated in a Bayesian framework with a Markov Chain Monte Carlo algorithm. The application to a data series from Marseille shows an improvement of the hourly estimations thanks to the combination between hourly and daily data in the model. Moreover, results are significantly different with or without dependence hypotheses: the dependence between 24 hours and 72 hours durations is significant, and the quantile estimates are more severe in the dependence case.

Keywords

Depth-Duration-Frequency; Extreme value distributions; Bivariate extreme distributions; Extremal index;Bayesian framework

Abbreviations

DDF, Depth-Duration-Frequency; MCMC, Markov Chain Monte Carlo; GEV, Generalized Extreme Value Distribution; h, hour

1. Introduction

The rainfall intensity patterns for various return periods are required for designing hydraulic structures (dams, levees, drainage systems, bridges, etc.) or for flood mapping and zoning. The objective of the rainfall depth-duration-frequency (DDF) curves is to estimate the maximum amount of rainfall for any duration and return period. This frequency analysis uses annual or seasonal maximum series, or independent values above a high threshold selected for different durations. If each duration is treated separately, contradictions between rainfall estimates can occur. DDF analysis takes into account the different durations in a single study, and prevents curves from intersecting.

The first relationship goes back as early as 1932 (Bernard, 1932). The classical approach for building DDF curves has three steps (Chow et al., 1988). In the first step, a probability distribution function is fitted to each duration sample. In the second step, the quantiles of several return periods T are calculated using the estimated distribution function from step one. Lastly, the DDF curves are determined by fitting a parametric equation for each return period, using regression techniques between the quantile estimates and the duration. The disadvantages of this procedure are the need to have a large number of parameters, and the calculation of a regression based on dependent values (since the estimated quantiles come from the same observed series, but aggregated into different time scales).

Several empirical models have been proposed (see Garcia-Bartual and Schneider, 2001 for a review). More recently, some approaches have been derived from a multifractal process (Burlando and Rosso, 1996; de Lima and Grasman, 1999; Veneziano and Furcolo, 2002; Borga et al., 2005). All these approaches need fewer parameters than the classical one, but the dependence problem remains. In section 2, two models are presented: an empirical classical model and an improved empirical model including a relation between the daily and 24 hourly maximum rainfall distributions. Section 3 presents theoretical and practical methods for estimating model parameters, quantiles and confidence intervals in a Bayesian framework, using a Markov Chain Monte Carlo (MCMC) algorithm.

Section 4 gives an application to a rainfall series for Marseille, in southern of France. Section 5 gives the conclusions of this study.

2. Depth-Duration-Frequency (DDF) relationships

2.1. Distribution of annual maximum rainfall

If $X(t)$ is the rainfall intensity at time t, then $Y_i(\delta)=\int_i^{i+\delta} X(t)dt$ is the aggregated rainfall from time i over δ hours. Then the hourly and daily observations correspond to the time series $\{Y_i(1)\}$ and $\{Y_{24i}(24)\}$ respectively. The studied variables are $H_d=\max\{Y_i(d)\}$, the annual maximum rainfall depth measured in a moving window of d hours width, and $H_D=\max\{Y_{24i}(24)\}$ the daily annual maximum rainfall depth.

A traditional approach for estimating the annual maximum rainfall H in France is based on the Gumbel distribution (Gumbel 1958):

$$G(x) = P(H \le x) = \exp\left(-\exp\left\{-(x-\beta)/\alpha\right\}\right), \qquad (1)$$

which is a particular form ($k = 0$) of the GEV distribution:

$$G(x) = P(H \le x) = \exp\left(-\left\{1-k(x-\beta)/\alpha\right\}^{1/k}\right), \quad with\ k(\beta - x) + \alpha > 0 \qquad (2)$$

This paper will show an example where a GEV distribution with a negative shape parameter k is more suitable than the Gumbel distribution.

2.1.1. The empirical DDF model

The following model attempts to estimate the behavior of the hourly variables H_d. Garcia-Bartual and Schneider (2001) give a review and a comparison of nine empirical models, with two or three parameters. Koutsoyiannis et al. (1998) give the general formula:

$$I_d(T)=a(T)/b(d) \qquad (3)$$

where $I_d(T)$ is the annual maximum rainfall intensity at the return period T for the duration d; $b(d)=(d+\theta)^{\eta}$, with $\theta>0$, $\eta\in(0,1)$, and $a(T)=F_Y^{-1}$ (1-1/T) where F_Y is a distribution function (for example GEV, lognormal, Gamma, log Pearson III, generalized Pareto distribution) of the normalized process of intensity $I_d(.)b(d)$.

In this study F_Y will be the GEV distribution of the annual or seasonal maximum rainfall. Then, H_d has a GEV distribution, with a quantile $H_d(T)$ given by:

$$H_d(T)=dI_d(T)=d\left(\beta+\alpha/k\left\{1-\left\{-\log(1-1/T)\right\}^{k}\right\}\right)/(d+\theta)^{\eta}. \qquad (4)$$

The parameters α_d, β_d, k_d of the distribution of H_d are simply expressed with α, β, k, θ, η:

$$\alpha_d = d\alpha/(d+\theta)^{\eta}, \quad \beta_d = d\beta/(d+\theta)^{\eta}, \quad k_d = k. \qquad (5)$$

Before using these relationships, it needs to be determined whether one DDF model can be applied to the whole range of durations, rather than several DDF sub-models on different sub-ranges of durations.

2.1.2. The extremal index DDF model

This second model improves the first one and attempts to estimate the behavior of the variables H_d and H_D. The extremal index (noted *EI*) is a measure of the extremal dependence in the series. Namely, the dependence of extreme values can be measured through the size of clusters of extreme values. A cluster definition is the following: a cluster of extreme values begins with a value above a high threshold u, and finishes when r consecutive values are under the threshold u (Beirlant et al. 2004). Let n_u denote the number of times an upper threshold u is exceeded, and n_c the number of clusters above u; n_c depends on u and r. Careful choices of u and r are needed, as if r is too small, clusters can be dependent and if r is too large, n_c becomes too small.

Leadbetter (1983) showed that *EI*=(mean cluster size)$^{-1}$. Several methods exist to estimate *EI*, the extremal index of a stationary series (Beirlant et al., 2004).

According to Robinson and Tawn (2000), the following estimator generally produces good estimates:

$$\hat{E}I(u,r) = n_c / n_u . \tag{6}$$

The asymptotic value $EI = \lim_{u\to\infty} \hat{E}I(u,r)$ can be approached using a sequence of thresholds (u_1, …, u_n) that increase with n. The limit is considered to have been reached when estimations of $\hat{E}I(u_n,r)$ are stable for u_n above some threshold u.

When daily data are available, their series are often longer and should be included in the model. An empirical relation can be used (Weiss 1964):

$$H_{24} = 1.14 H_D . \tag{7}$$

Where 1.14 is an estimation of the Hershfield factor (Hershfield, 1961). Van Montfort (1997) proposed a method for estimating this factor. A theoretical relation between distributions of H_{24} and H_D has been proposed by Robinson and Tawn (2000):

$$P(H_{24} \le x) = P\left(H_D \le x\right)^{24 EI_{24} / EI_D} \tag{8}$$

where $0 \le EI_D, EI_{24} \le 1$ are the extremal indexes of the daily and 24 hourly series. Relation (8) is based on hypotheses of stationarity and strong-mixing dependence of the series. Stationarity is taken in the strict sense: a process $X_1, X_2, \ldots$ is said to be stationary if, for any subset of integers $\{i_1,\ldots,i_k\}$, and any integer m, the joint distributions of $(X_{i_1},\ldots,X_{i_k})$ and of $(X_{i_1+m},\ldots,X_{i_k+m})$ are identical. The strong-mixing dependence is defined by: for all $i_1 < \ldots < i_p < j_1 < \ldots < j_q$ with $j_1 - i_p > l_n$

$$\begin{aligned} &\Big| P(X_{i1} \le u_n,\ldots, X_{ip} \le u_n, X_{j1} \le u_n,\ldots, X_{jp} \le u_n) \\ &- P\left(X_{i1} \le u_n,\ldots, X_{ip} \le u_n\right) P\left(X_{j1} \le u_n,\ldots, X_{jq} \le u_n\right) \Big| \le \alpha(n, l_n) \end{aligned} \tag{9}$$

where $\alpha(n,l_n)\to 0$ for a sequence l_n such that $l_n\to 0$ as $n\to\infty$, and a sequence of thresholds u_n that increase with n.

Let $\Theta=24EI_{24}/EI_D$, the equation (8) implies relations between GEV parameters of both distributions (Ancona-Navarrete and Tawn, 2000; Coles, 2001):

$$\begin{aligned} &\text{if } k_D=0: \quad \beta_{24}=\beta_D+\log(\Theta)\alpha_D, \quad \alpha_{24}=\alpha_D, \quad k_{24}=0 \\ &\text{if } k_D\neq 0: \quad \beta_{24}=\beta_D+\alpha_D/k_D(1-\Theta^{-k_d}), \quad \alpha_{24}=\alpha_D\Theta^{-k_d}, \quad k_{24}=k_D \end{aligned} \tag{10}$$

The daily data are included in the model (4). A new model is then defined, whose parameters are α_D, β_D, k_D, Θ, θ and η. All the parameters α_d, β_d and k_d of the GEV distribution of H_d are simple functions of the model parameters. For example, in the case $k_D\neq 0$, model (4) becomes:

$$H_d(T)=(d/24)\left[\beta_D+\alpha_D/k_D\left\{1-\Theta^{-k_D}\left(-\log(1-1/T)\right)^{k_D}\right\}\right](24+\theta)^{\eta}/(d+\theta)^{\eta}. \tag{11}$$

In the model, the shape parameter k_d is constant for the different durations, and equal to the shape parameter k_D. Nadarajah et al. (1998) showed theoretically, with a study of ordered multivariate extremes, that the relationship $H_d \le H_{d'} \le (d'/d)\, H_d$ imposes restrictions on the marginal distributions. In particular,

$$k_d = k_d' \le 0 \quad \text{or} \quad k_d > 0, k_d' > 0 \tag{12}$$

In our case, the rainfall is assumed not to be upwardly bounded, thus $k_d \le 0$, and all the shape parameters are equal. Moreover, the relationship (8) between daily and 24 hours maximum rainfall implies equality between k_{24} and k_D.

2.2. Selection of two duration ranges

Since cumulative rainfalls on short and long durations are derived from different rainfall processes, two duration ranges will be considered. The empirical model from eq. (4) is chosen for the short duration rainfalls. Since long duration rainfalls are assumed to contain daily rainfall, the extremal index model from eq. (11) is

used for the long duration rainfalls. Let d_b be the boundary duration that separates the short and long durations. To ensure consistency between short and long durations, the estimated parameters of both ranges have to satisfy continuity in d_b. The shape parameter is constant in both ranges, according to the theoretical study of Nadarajah et al. (1998).

Let f_d ($x; \alpha_d, \beta_d, k_d$) be the GEV density of the maximum annual or seasonal rainfall in d hours, where α_d, β_d and k_d are the scale, location and shape parameters. Therefore, the relationships between the parameters (α_d, β_d, k_d) and the duration d are asfollows:

- for short durations, $d \leq d_b$, and $\alpha_s, \beta_s, \eta_s, \theta_s$ denote the parameters of eq. (4):

$$\alpha_d = d\alpha_s /(d+\theta_s)^{\eta_s} \quad ; \qquad \beta_d = \; d\beta_s /(d+\theta_s)^{\eta_s} ; \qquad k_d = \; k_D \qquad (13)$$

- for long durations, $d \geq d_b$, and α_D, β_D, k_D, Θ, θ, η denote the parameters of eq. (11), for example if $k_d \neq 0$:

$$\alpha_d = (d/24)\alpha_D \Theta^{-k_D} (24+\theta)^{\eta}/(d+\theta)^{\eta}; \; \beta_d = (d/24)\left\{\beta_D + \alpha_D/k_d(1-\Theta^{-k_D})\right\}(24+\theta)^{\eta}/(d+\theta)^{\eta}; \; k_d = k_D \qquad (14)$$

Continuity hypotheses on the boundary d_b imply that $\alpha_{d_b}, \beta_{d_b}$ have the same values in both equations (13) and (14). This implies:

$$\begin{aligned} \beta_s &= \alpha_s \beta_{24}/\alpha_{24} \\ \eta_s &= \left\{ \log[24\alpha_s (d_b+\theta)^{\eta}] - \log[\alpha_{24}(24+\theta)^{\eta}] \right\} / \log(d_b+\theta_s) \end{aligned} \qquad (15)$$

With two ranges of durations, eight parameters (α_D, β_D, k_D, Θ, θ, η, α_s and θ_s) are sufficient to calculate α_d, β_d, k_d, for all d in the ranges of durations.

3. Bayesian framework

3.1. Presentation of the Bayesian method

In the Bayesian paradigm, a prior density of parameters can express the knowledge or a physical constraint about the parameters, without reference to the data. The Bayes' theorem projects this prior information on parameters onto the additional information provided by the data, to calculate a posterior distribution. Since the posterior distribution is not easily tractable, a Markov Chain Monte Carlo algorithm (MCMC) is required to simulate this distribution. In this paper, the algorithm proposed by Renard et al. (2006) is used, with a combination of the Gibbs-Metropolis and Metropolis algorithms. The interest of this mixed MCMC algorithm is to combine the efficiency of the Gibbs sampler, and the speed of the Metropolis algorithm. The steps of the algorithm are the following:

(i) Begin with a set of parameters ($\alpha_D^{(0)}$, $\beta_D^{(0)}$, $k_D^{(0)}$, $\Theta^{(0)}$, $\theta^{(0)}$, $\eta^{(0)}$). A normal jumping distribution is chosen, with a mean equal to 0, and a variance given by prior knowledge.

(ii) As a first step, 1000 iterations of the Gibbs sampler are run. Each iteration cycles through the six parameters (α_D, β_D, k_D, Θ, θ, η), sampling each parameter conditionally on the value of all the others. Since sampling from the conditional distribution is impossible, 100 iterations of Metropolis algorithm are required to construct an approximation of the conditional distribution (Gelman et al., 1997).

(iii) As a second step, 80 000 iterations of the Metropolis algorithm are run, with a starting point given by the mean of the 500 last parameter sets of the first step, and a normal jumping distribution whose covariance matrix is defined by the covariance of the 500 last parameter sets of the first step. The convergence of the algorithm is monitored by the R statistic calculated on several parallel sequences (Gelman et al., 1997).

3.2. Prior elicitation

A priori distribution of the parametersgives the different prior distributions, which are similar to the choice of Coles and Pericchi (2003) for the GEV parameters.

Parameter	Distribution
α_D	lognormal with mean 0 and variance 100
β_D	normal with mean 0 and variance 100
k_D	uniform on [-1,1]
Θ	uniform on [1,24]
θ	normal with mean 0 and variance 100, truncated at 0
η	lognormal with mean 0 and variance 100
Φ	uniform on [0,1]

Table 1: A priori distribution of the parameters

Since the clusters are larger in the time series of 24 hourly rainfall data than in the time series of daily rainfall, $EI_{24} \leq EI_D$, therefore $\Theta = 24EI_{24}/EI_D \leq 24$. Moreover, $\Theta \geq 1$. Indeed, let $N_{r_n}(u_n)$ denote the number of excess of u_n in r_n consecutive measures of 24 hourly rainfall ($H_{24,1},...,H_{24,r_n}$), for r_n such that $\lim_{n\to\infty} r_n/n = 0$. A cluster of extremes is defined to occur when $N_{r_n}(u_n) > 0$, with the values in the cluster being the excedances of u_n. The cluster size distribution $\pi_{24,n}$ is defined by:

$$\pi_{24,n}(j) = P(N_{r_n}(u_n) = j \,|\, N_{r_n}(u_n) > 0), \ for\ j = 1,...,r_n \qquad (16)$$

The limiting cluster size distribution is:

$$\pi_{24}(j) = \lim_{n\to\infty} \pi_{24,n}(j), \ for\ j = 1,...,\infty \qquad (17)$$

Robinson and Tawn (2000) showed that:

$$EI_D \leq 24EI_{24}\left[1 - \sum_{i=1}^{23}(1 - i/24)\pi_{24}(i)\right] \qquad (18)$$

Therefore, as:

$$1-\sum_{i=1}^{23}(1-i/24)\pi_{24}(i) \geq 1-\sum_{i=1}^{\infty}\pi_{24}(i)+\sum_{i=1}^{23}i/24\pi_{24}(i)=\sum_{i=1}^{23}i/24\pi_{24}(i)>0 \quad (19)$$

this implies that:

$$\Theta = 24EI_{24}/EI_D \geq 1/\left(1-\sum_{i=1}^{23}(1-i/24)\pi_{24}(i)\right) \geq 1. \quad (20)$$

Therefore $\Theta \in [1,24]$.

As θ must be positive, a normal distribution truncated at 0 is chosen. Moreover, parameters of both models have to satisfy some physical constraints:

$$\begin{gathered} d<d' \Rightarrow \quad H_d \leq H_{d'} \leq (d'/d)H_d \\ H_D \leq H_{24} \leq 2H_D \end{gathered} \quad (21)$$

If the quantiles calculated for different return periods T=2, 5, 10, 100 and 1,000 years do not verify these two relations, the parameters are rejected in the MCMC algorithm.

3.3. Likelihood definition

Three different likelihood formulations L_1, L_2 and L_3 will be used.

- Likelihood L_1: independence between seven durations.

Firstly, independence is supposed between durations d = 1 h, 6 h, 12 h, 24 h, 48 h, 72 h and daily (D) observations. The likelihood is expressed as

$$L_1 = \prod_{d=1,6,12,24,48,72,D}\prod_{i=1}^{id} f_d(x_i^{(d)};\alpha_d,\beta_d,k_d) \quad (22)$$

where $x_i^{(d)}$, i=1,…,i_d are the annual or seasonal maximum rainfall measured in d hours and α_d, β_d, k_d are given by equations (13) to (15). The advantage of this likelihood is the use of a large set of available data.

- Likelihood L_2: independence between four durations

Secondly, since independence between all these durations is an unlikely hypothesis, only four durations will be considered. Since the one-hour maximum rainfall generally occurs during a thunderstorm, whereas the 72-hours maximum rainfall occurs generally during a frontal rainfall event, both maxima are assumed to originate from different processes. Both durations may be considered independent, as shown by Kieffer Weisse (1998). Moreover, 24 hourly data, and daily data of years without hourly measurements will be used. The likelihood formula is given by:

$$L_2 = \prod_{d=1,24,72,D} \prod_{i=1}^{i'_d} f_d(x_i^{(d)};\alpha_d,\beta_d,k_d) \qquad (23)$$

where α_d, β_d, k_d are given by equations (13) to (15), $i'_1=i_1$, $i'_{24}=i_{24}$ $i'_{72}=i_{72}$ and i'_D is the number of years of daily measurements without hourly measurements.

- Likelihood L_3: four durations, with dependence between both of them.

Lastly, a generalization of the second likelihood is introduced: the dependence between 24 hours and 72 hours maximum rainfall is considered through a bivariate extreme distribution, from the logistic family (Coles, 2001).

$$G(x,y) = P(u_{24}(H_{24}) \le x,\ u_{72}(H_{72}) \le y) = \exp\left\{-(x^{-1/\Phi}+y^{-1/\Phi})^{\Phi}\right\},\quad x>0,\quad y>0, \qquad (24)$$

for a dependence parameter $\Phi\in(0,1)$. As $\Phi\to 1$, $G(x,y)\to\exp\{-(x^{-1}+y^{-1})\}$, corresponding to independent variables; as $\Phi\to 0$, $G(x,y)\to\exp\{-\max(x^{-1}, y^{-1})\}$, corresponding to perfectly dependent variables. This is the most widely used model in bivariate extreme value analysis. The transformed variable $u_d(H_d)$=-

$1/\log(G_d(H_d))$, with G_d the GEV distribution of rainfall H_d is standard Fréchet distributed.

(proof: $P(-1/\log(G_d(H_d)) \le x) = P(G_d(H_d) \le \exp(-1/x)) = \exp(-1/x)$ since $G_d(H_d)$ is uniformly distributed between 0 and 1).

The dependence structure of any bivariate extreme value distribution function G can be described in several ways. A quite popular way is the Pickands dependence function A, satisfying some properties (Beirlant et al., 2004). The Pickands dependence function $A(t)$, is defined for $t \in [0,1]$ by:

$$A(t) = -\log\left(G\left[G_1^{-1}\{\exp(-1+t)\}, G_2^{-1}\{\exp(-t)\}\right]\right) \tag{25}$$

where G_1, G_2 are the two marginal distributions. G is completely determined by its margins G_1, G_2 and its Pickands dependence function $A(t)$ through equation (25). A can be estimated by non-parametric (Pickands 1981, 1989; Capéraà et al. 1997), or by parametric methods. The comparison between non-parametric and parametric estimators of A is a way to validate the parametric model.

The likelihood is given by:

$$L_3 = \prod_{i=1}^{i_{24}} g\left(u_{24}(x_i^{(24)}), u_{72}(x_i^{(72)}); \alpha_{24}, \beta_{24}, k_{24}, \alpha_{72}, \beta_{72}, k_{72}, \Phi\right) u'_{24}(x_i^{(24)}) u'_{72}(x_i^{(72)}) \prod_{i=1}^{i_D} f_D(x_i^{(D)}; \alpha_D, \beta_D, k_D) \prod_{i=1}^{i_1} f_1(x_i^{(1)}; \alpha_1, \beta_1, k_1) \tag{26}$$

where u'_d is the derivative function of u_d, and g is the density function of the bivariate logistic distribution. Note that if $k_d=0$, $u_d(x)=\exp((x-\beta)/\alpha)$, and if $k_d \neq 0$, $u_d(x)=(1-k_d(x-\beta)/\alpha)^{-1/k_d}$.

4. Applications of the DDF models to the Marseille rainfall data series

4.1. Presentation of the series

Two data series at Marseille are available, with 67 years of hourly data (1918-2002) and 122 years of daily data (1882-2003). The daily series have been reviewed by Météo-France through the European project IMFREX for the study of climate changes, using homogeneity criteria. In both hourly and daily series, each year has less than 10% of missing values. In order to exclude problems of non-homogeneity due to seasonality, two different seasons have been considered based on the mean of monthly maximum rainfall (Kieffer Weisse, 1998) (see Figure 1). As heavy hourly and daily rainfalls occur within the September-January period, this period will be chosen as representative of extremal events. All the results presented are calculated within this period.

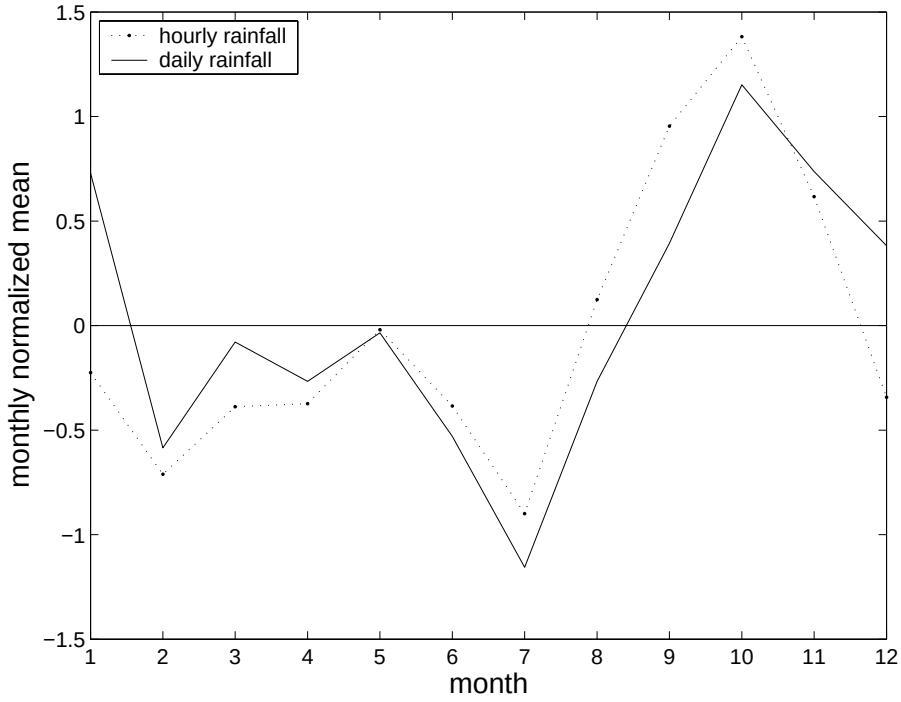


Figure 1: Seasonal fluctuation of monthly maximum rainfall: normalized and centered mean of monthly maximum rainfall for the daily and one hour rainfall data.

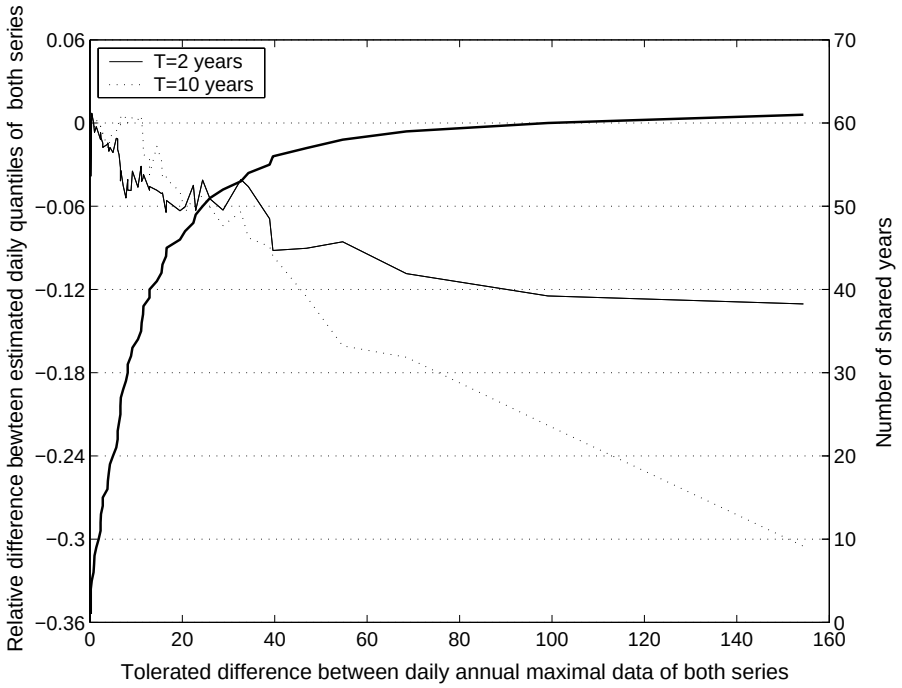


Figure 2: Selection of the tolerance threshold for the difference between daily annual maxima of both daily and hourly series. Left Y axis: relative difference between daily quantiles of both series (corresponding curves: solid and dotted lines for $T = 2$ and 10 years). Right Y axis: number of tolerated years in the hourly series, in common with the daily series (corresponding curve: bold line).

The consistency between the two data series has been tested, comparing their annual maximum values, and retaining the years of the hourly series where the absolute value of the difference between daily annual maxima is less than some tolerance threshold. This tolerance threshold is calculated by the following method: for any tolerance threshold between 1 mm and around 160 mm, two hourly and daily sub-series are defined. Both sub-series contain only those years whose difference between rainfall maximum amounts is lower than the tolerance. The optimum tolerance threshold is a compromise between the number of selected years in the sub-series (bold curve in Figure 2) and the relative difference between daily quantiles estimated for both sub-series (solid and dotted lines Figure 2, for return periods 2 and 10 years). The chosen threshold is 19 mm, which corresponds to a relative difference less than 6%, and 45 validated years in the hourly series.

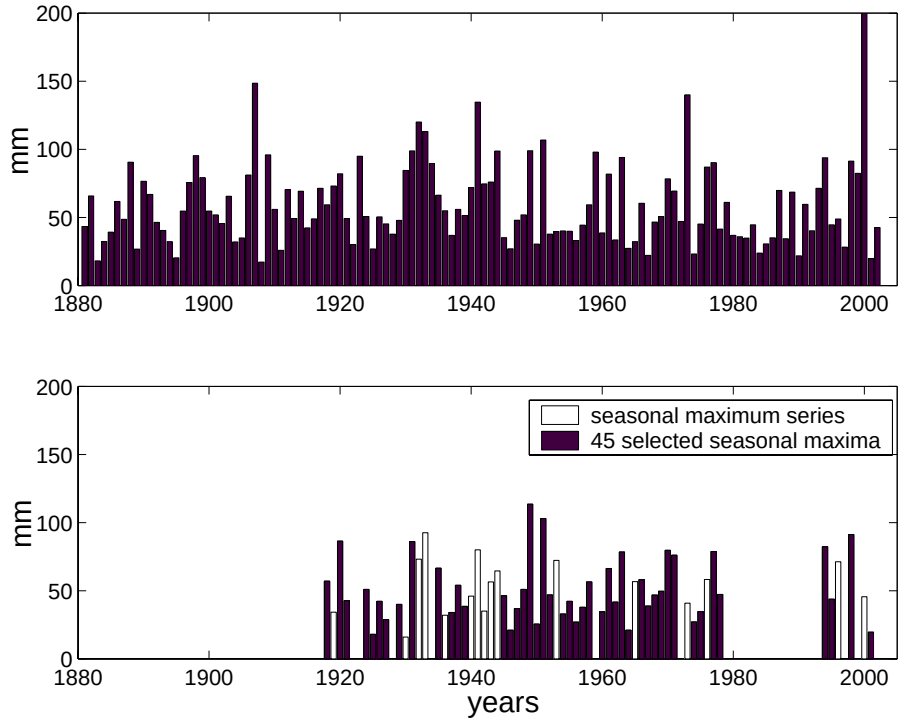


Figure 3: Top: daily seasonal maximum rainfall amount of the daily Marseille series (122 years). Bottom: daily seasonal maximum rainfall depth calculated on the hourly series (67 years). Black bars represent the 45 selected years of hourly series, after check of the consistency between series.

A comparison between the extreme values of the two data sets (daily and hourly values on Figure 3) shows that the hourly recording rain gauge has some difficulties in recording very extreme rainfall events. For example, the extreme events in 2000 (200 mm), 1973 (140 mm) and 1932 (120 mm) are missing values in the hourly rainfall series.

Some aspects of stationarity of the 122 years series have been checked. Firstly, a likelihood ratio test has been applied between a GEV distribution with a temporal trend in scale and position parameters, and a GEV distribution with fixed parameters. Both distributions are nested models $M_s \subset M_t$, where M_s, M_t are the stationary and trend models. Therefore the deviance statistic:

$$D=2(l_t(M_t)-l_s(M_s)) \qquad (27)$$

is χ_2^2 distributed, where $l_s(M_s)$, and $l_t(M_t)$ are the maximized log-likelihoods for models M_s and M_t respectively. The stationarity hypothesis is rejected at the level

of significance α if $D>\chi_{2,\alpha}^2$, where $\chi_{2,\alpha}^2$ is the (1-α) quantile of the χ_2^2 distribution. As the computed statistic $D = 0.186$, the stationary model is not rejected at a level larger than α=25%. Secondly, stationarity has been tested on different annual variables: mean annual rainfall of wet days (with more than 1 mm precipitation), annual maximum, annual ratio of zero rainfall, annual ratio of values above upper thresholds. The non-parametric Mann-Kendall test (Mann, 1945; Kendall, 1975) is used to detect monotonic trends in series of independent data. The Mann-Kendall statistic S computes, for every ordered pair of the studied series, the number of pairs where the first value is larger than the second, and the number of pairs where the first value is lower than the second, and calculates the difference between the two quantities $S = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} sign(X_j - X_i)$. Let a variable Z be defined:

$$Z = \begin{cases} (S-1)/\sqrt{Var(S)} & S > 0 \\ 0 \quad \text{if} & S = 0 \\ (S+1)/\sqrt{Var(S)} & S < 0 \end{cases} \qquad (28)$$

$$\text{with } Var(S) = n(n-1)(2n+5)/18$$

If $n \geq 10$, Z follows approximatively a standard normal distribution. No significant trend was detected for the four variables by the Mann-Kendall test, for a level α=10%.

4.2. Dependence between rainfall depths

The correlation coefficient between 1 hourly and 72 hourly maximum rainfall is equal to 0.41. This value is quite high, therefore the independence hypotheses is only justified by physical reasons, as the corresponding rainfall processes are considered to be different (Kieffer Weisse, 1998). The correlation coefficient is 0.56 between 1 hour and 24 hours, and 0.90 between 24 hours and 72 hours, justifying the bivariate distribution in likelihood L_3 between 24 hourly and 72 hourly rainfalls.

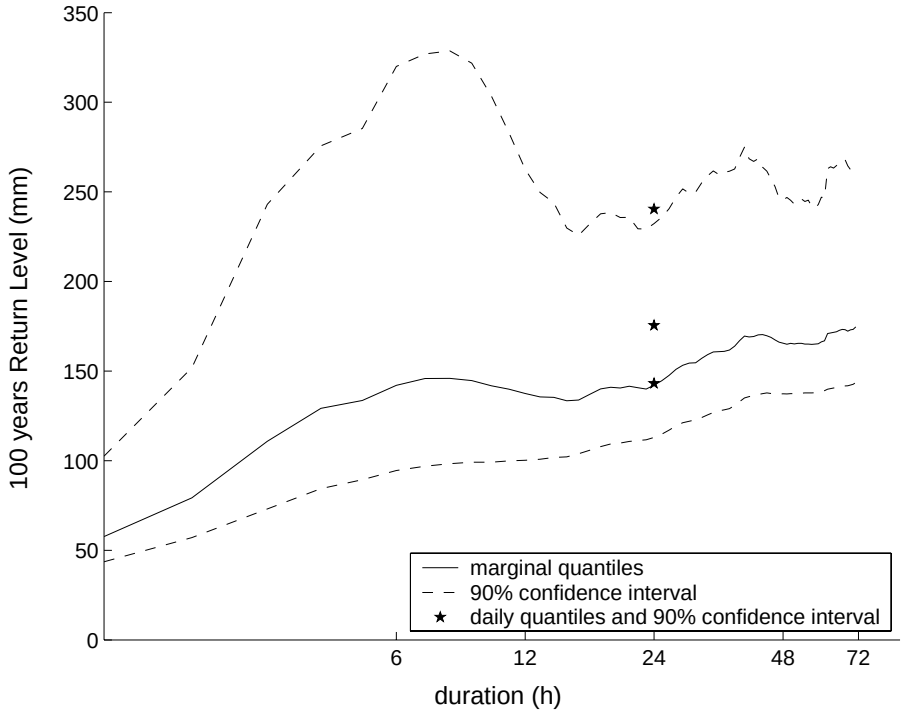


Figure 4: Marginal estimates and 90% confidence intervals (estimated by a Bayesian analysis) of $H_D(100)$ (stars) and $H_d(100)$ for d=1 h to 72 h.

From Figure 4, it is seen that the estimated value for $H_6(100)$ is too large to meet the required marginal ordering constraint with $H_{12}(100)$. This is due to the marginal estimator $\hat{k}_{12}$, which is larger than $\hat{k}_6$. $H_{24}(100)$, $H_{48}(100)$, $H_{72}(100)$ are too small to meet the ordering constraints with H_D. This is explained by the fact that the hourly series contains only 45 years, whereas the daily series is 122 years long, and the largest rainfall values are not included within the hourly series.

The logistic bivariate extreme distribution is fitted to the bivariate rainfall data (24 h, 72 h). The Pickands dependence function of this particular distribution compares well with its non-parametric estimators by Pickands (1981, 1989) and Capéraà et al. (1997) (see Figure 5).

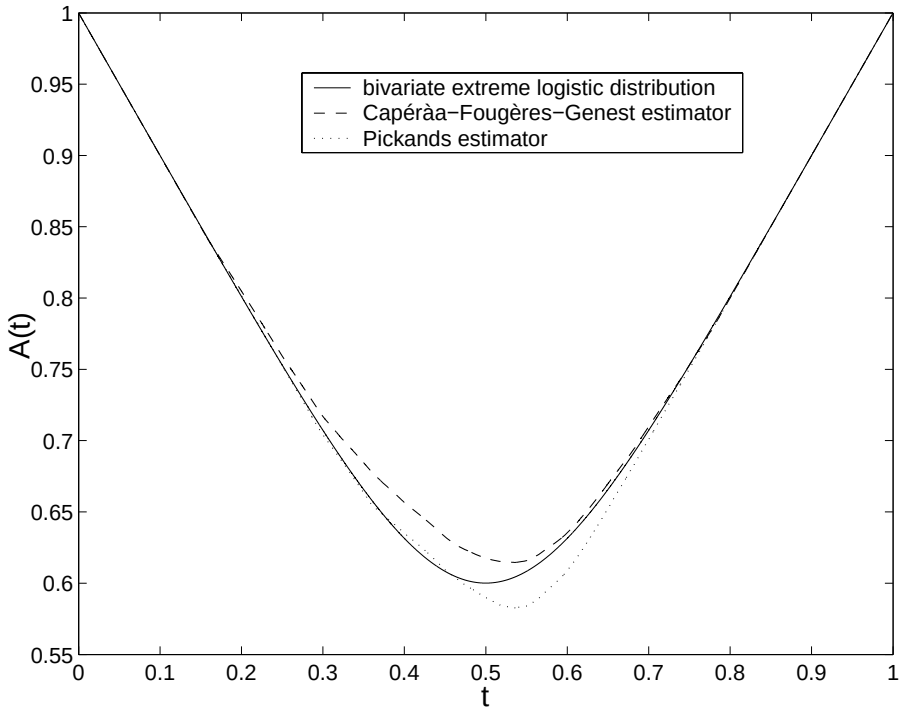


Figure 5: Estimation of the Pickands dependence function $A(t)$: comparison between non-parametric Pickands and Capéraà *et al.* estimations and logistic estimation.

Moreover, a likelihood ratio test has been applied between three nested models, fitted to the bivariate data (24 h, 72 h): independence case, logistic bivariate distribution and a logistic asymmetric bivariate distribution, whose bivariate distribution is:

$$G(x,y)=P(H_{24}\le x,H_{72}\le y)=\exp\left(-(1-\Psi_1)x^{-1}-(1-\Psi_2)y^{-1}-\left\{(\Psi_1 x^{-1})^{1/\Phi}+(\Psi_2 y^{-1})^{1/\Phi}\right\}^{\Phi}\right) \quad (29)$$

$$with\ \Psi_1,\Psi_2\in[0,1]$$

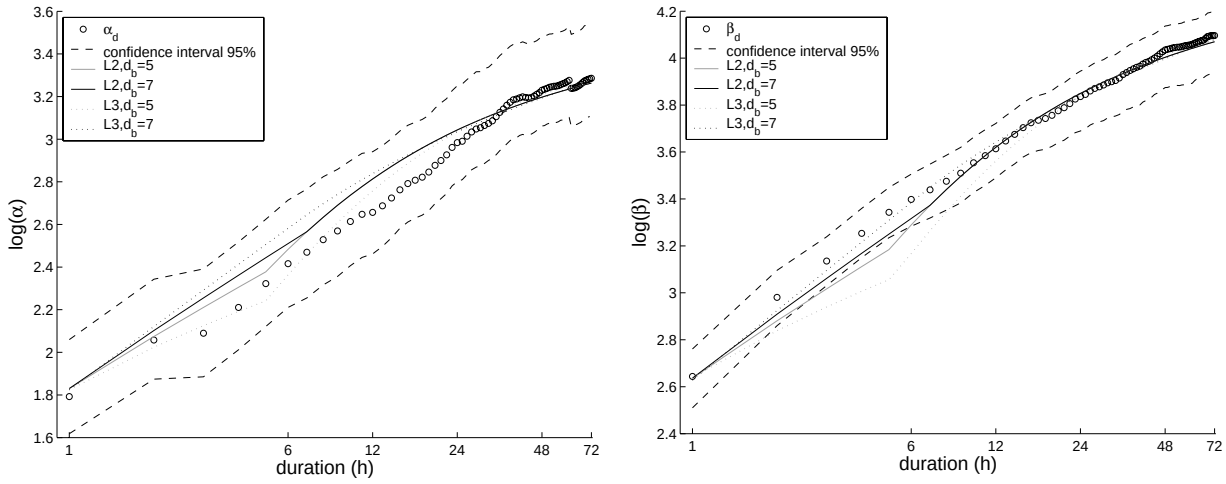
The logistic asymmetric bivariate distribution is a more general model than the logistic symmetric one: the two variables are exchangeable in the symmetric case, but not in the asymmetric case. No significant difference was detected at the 10% level between the logistic and asymmetric logistic models ($\hat{\psi}_1=1,\hat{\psi}_2=1$), but the logistic model was better than the independence model, with a significant ratio test (the p-value is lower than 0.1%). The dependence parameter Φ of the logistic distribution was estimated to be 0.24 by likelihood maximization, implying a high

level of dependence between 24 hourly and 72 hourly rainfall data. The effect of the bivariate logistic distribution is to change the shape parameter estimation: marginally, $\hat{k}_{24}$ =0, $\hat{k}_{72}$ =0.04, but with the bivariate logistic distribution, applied on 24 and 72 hourly data and without the constraint k_{24}=k_{72}:

$\hat{k}_{24}$ =-0.14, $\hat{k}_{72}$ =-0.12, which is close to the daily marginal estimator: $\hat{k}_D$ =–0.13.

4.3. Choice of duration ranges

The durations are separated into two ranges. The boundary duration d_b between short and long durations is added as an extra parameter, in likelihood definitions L_1, L_2, and L_3. Then the boundary duration is estimated by likelihood maximization, under the constraints (21). The maximum likelihood estimator of d_b is 5.6, with likelihood L_1, whereas likelihoods L_2, L_3 are not discriminant for d_b, and give equal maximum likelihoods with d_b=5, 6 or 7 hours, while other parameters change slightly.



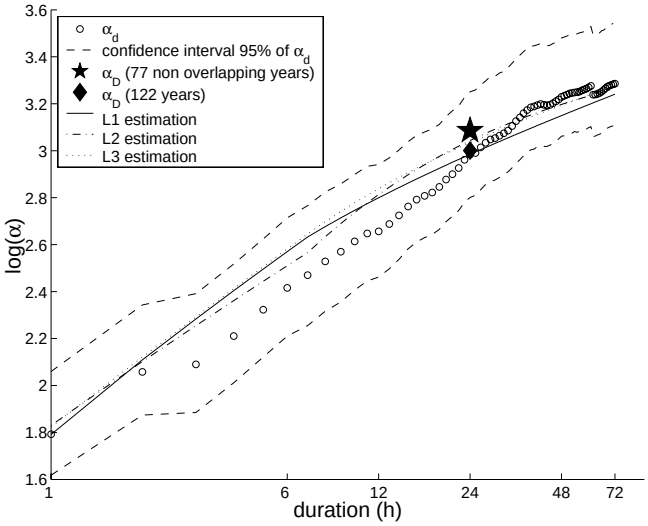
(a) (b)

Figure 6: Choice of the duration d_b of the boundary between short and long durations. Comparison under L_2 and L_3 maximizations and constraints (21), with two splits (d_b = 5 or 7 hours), of: (a) α_d estimates; (b) β_d estimates.

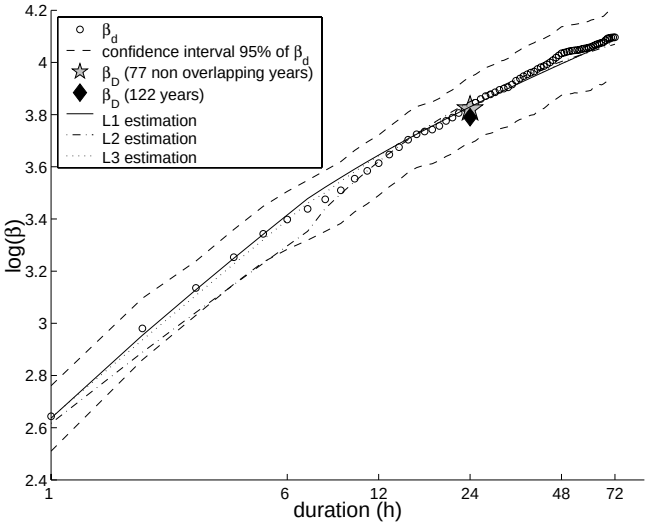
Figure 6 shows that d_b=7 hours gives a better fit of the parameters, with estimations inside the 95% confidence for α_d and β_d. Estimations with d_b =5 hours

are closed to the marginal estimate $\hat{\alpha}_d$, but outside the 95% confidence interval of $\hat{\beta}_d$. Results with d_b =6 hours are intermediate, but outside the 95% confidence interval of $\hat{\beta}_d$. The parameter k_d is not used for the choice of d_b, since k_d is constant in the model (k_d=k_D), and d_b does not affect the parameter k_D of the daily data. The chosen value for d_b is therefore 7 hours.

(a)



(b)



(c)

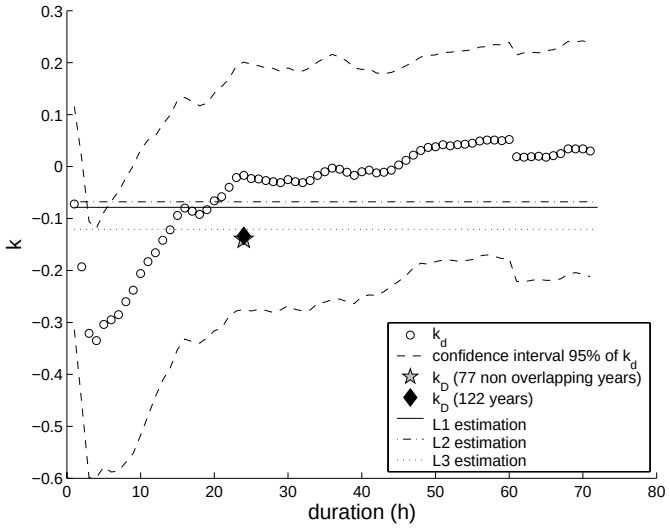


Figure 7: Comparison of α_d, β_d, k_d estimates under L_1, L_2 and L_3 maximizations and constraints (21): (a) α_d estimates; (b) β_d estimates; (c) k_d estimates.

Figure 7 shows the good fit of the maximum likelihood estimations under constraints (21), for L_1, L_2, L_3. The estimations of $\hat{\alpha}_d, \hat{\beta}_d$ with likelihoods L_1, L_2 and L_3 are close to the marginally estimated parameters. The marginal estimates of $\hat{k}_d$ are approximatively constant after about 24 hours, and present a minimum for the 4 hours duration. The variability of $\hat{k}_d$ is due to the sampling sensitivity of the estimator, and to the fact that the maximum rainfall in six hours (103 mm) is very close to the maximum rainfall in 15 hours (104 mm).

4.4. Comparison between the three likelihood definitions

The results are presented after the run of 80 000 MCMC simulations. The parameters were computed on the last 40 000 iterations, thus allowing 40 000 burning iterations. The convergence of the MCMC algorithm is assessed by the R statistic (Gelman et al., 1997) calculated for each parameter in the second half of the burning iterations. Eight parallel sequences of Metropolis algorithm have been considered, with a random starting point, sampled in the prior distribution. As the computed ratio R is very close to one, the convergence of the MCMC simulations can be accepted.

The estimated parameters are presented in the Table 2, with the median of the 40 000 last simulated parameters. The 90% confidence intervals are calculated by sorting each marginal simulated parameter, and excluding the values lower and larger than the 5% and 95% empirical quantiles. The comparison between the median and the middle of the 90% confidence interval shows that the posterior distribution of (α_D, β_D, k_D) is symmetric, but this was not the case for the other parameters.

	L_1	L_2	L_3
α_D	44.06(41.07,46.99)	44.82(40.92,48.65)	44.47(40.44,48.39)
β_D	20.05(18.45,21.83)	21.38(19.16,23.87)	21.23(18.89,23.89)
k_D	-0.083(-0.158,-.014)	-0.076(-0.183,0.015)	-0.131(-0.236,-0.035)
Θ	1.17(1.02,1.42)	1.18(1.02,1.48)	1.19(1.02,1.54)
θ	4.09(-1.03,11.88)	5.61(-1.33,21.66)	9.22(-1.42,32.96)
η	0.89(0.69,1.12)	0.96(0.74,1.25)	0.96(0.75,1.41)
α_s	10.15(4.41,16.89)	10.23(4.38,21.24)	9.36(4.27,18.51)
θ_s	0.94(-0.50,2.08)	0.88(-0.55,2.95)	0.68(-0.55,2.45)
Φ			0.265(0.212,0.335)

Table 2: Estimated parameters and 90% confidence intervals.

The MCMC estimates of the parameters α_d, β_d, k_d, for d between 1 h and 72 h (not graphically shown) are similar to those obtained with the maximum likelihood (presented in Figure 7). In the three cases, the k_d estimations are negative, implying unbounded quantiles, when the return period becomes infinite.

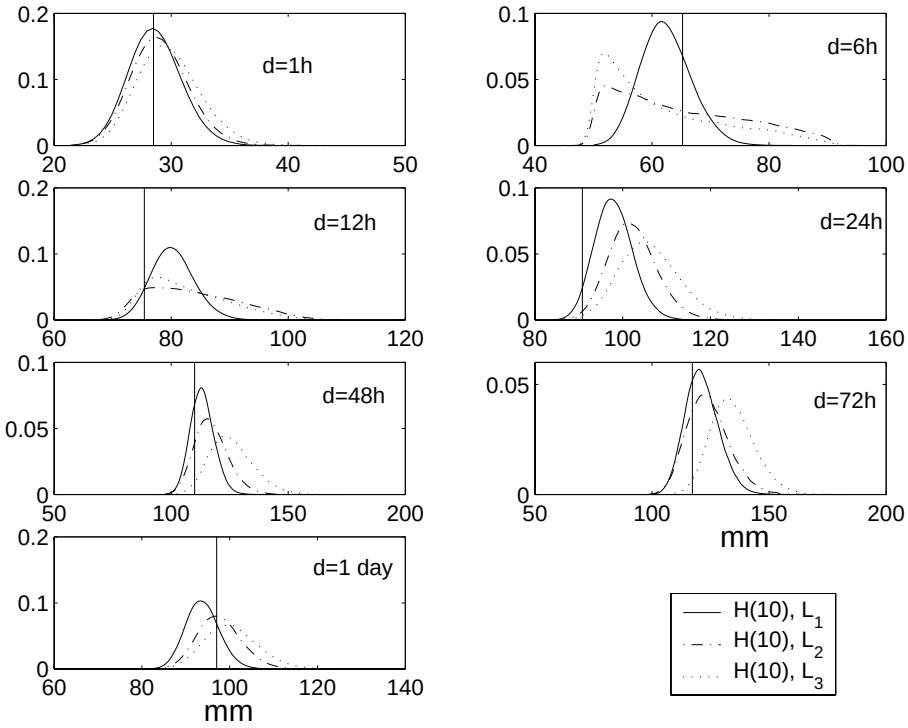


Figure 8: Posterior distributions of the quantiles $H_d(10)$ for d=1 h, 6 h, 12 h, 24 h, 48 h, 72 h and d=1 day (with the last 40 000 simulations of the MCMC algorithm). The vertical lines are the marginally estimated quantiles $H_d(10)$.

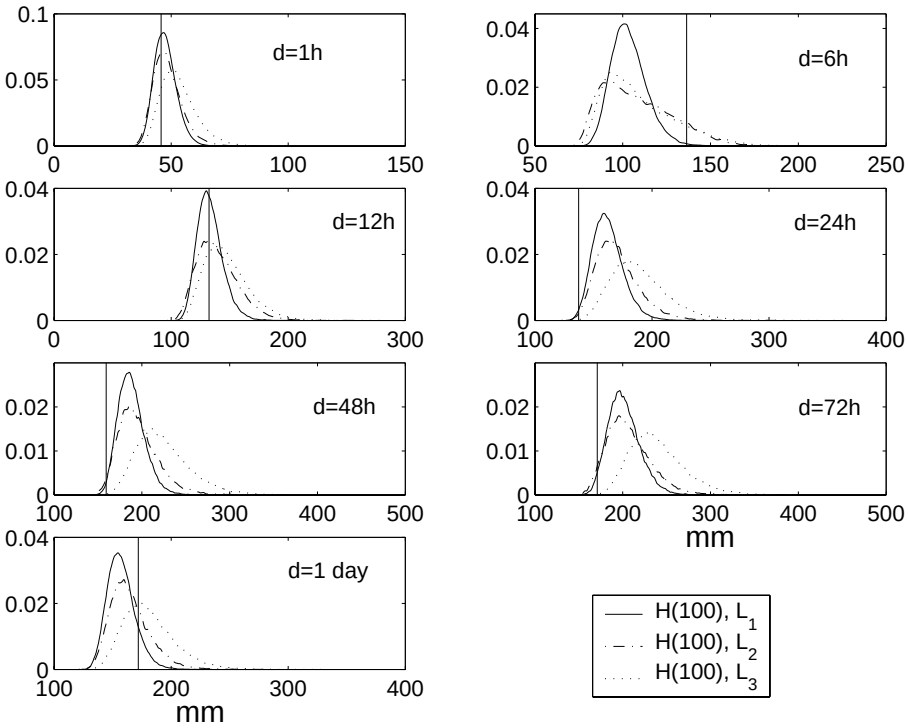


Figure 9: Posterior distributions of the quantiles H_d (100) for d=1 h, 6 h, 12 h, 24 h, 48 h, 72 h and d=1 day (with the last 40 000 simulations of the MCMC algorithm). The vertical lines are the marginally estimated quantiles H_d (100).

The posterior distributions of the quantiles $H_d(T)$ are presented in Figure 8 and Figure 9 for the return periods T = 10 and 100 years. The shape of the distributions is generally skewed, and the supports become larger with the successive likelihood definitions (L_1 to L_3). Namely, less data are used in the L_2 definition than in the L_1 one, and the L_3 definition includes an extra dependence parameter Φ.

As the daily series contains a large number of extreme values (200 mm, 148 mm, 140 mm, 138 mm, etc.) and as the hourly series does not contain the most extreme values, the estimated long duration quantiles are significantly larger than the marginal estimates (vertical lines in Figure 8 and Figure 9), especially when fewer hourly data are included in the estimation procedure (L_2, L_3). This is due to the link between long durations and daily rainfalls, by Θ in equation (11). For the same reason, daily quantiles are slightly lower than their marginal estimates. Moreover, the short durations are linked to the long durations only by k_D, and by continuity hypotheses of the parameters α_d, β_d at the boundary between short and long durations (cf. equations (13) to (15)). Thus daily data produce less effect on the short duration estimations. The six hours quantiles estimations are lower than their marginal estimations, because of the linkage between 1 hourly and 6 hourly rainfalls. The six hourly marginal quantiles are high because of the presence in the six hours series of the maximum rainfall falling in 15 hours, not present in the one hourly series.

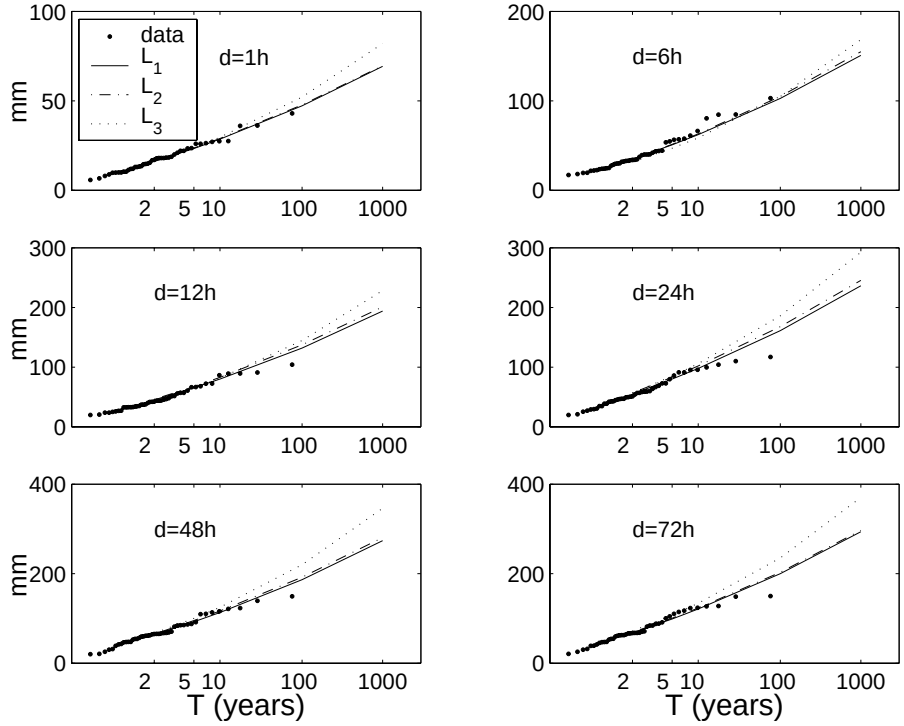


Figure 10: Median of the simulated distributions of maximum rainfall amount H_d, d=1 h to 72 h (with the last 40 000 MCMC simulations).

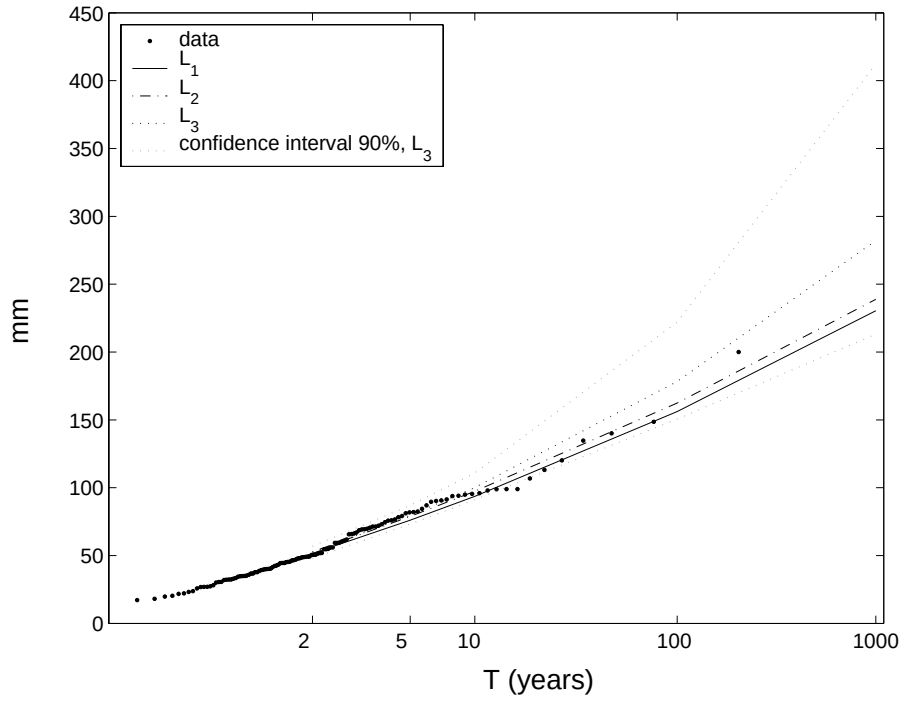


Figure 11: Median of the simulated distribution of maximum daily rainfall amount (with the last 40 000 MCMC simulations).

Figure 10 and Figure 11 show the DDF curves: the median quantiles are quite similar with the L_1 and L_2 likelihood definitions, and are larger with L_3, due to the k_D value in this last case (cf. Table 2). Thus the logistic dependence model has a significant influence on the estimations, with a dependence parameter Φ equal to

0.26: its 90% confidence interval does not contain 1, which corresponds to the independent case.

5. Conclusions

Based on hypotheses of GEV or Gumbel distribution for the seasonal maximum rainfall distribution, the DDF models agree with a GEV distribution, with negative shape parameter. The quantiles of any duration between 1 hour and 72 hours, and any return period between 2 years and 1000 years have been estimated by the proposed empirical model, associated with three different likelihood definitions. The durations have been separated into short (less than 7 hours) and long (above 7 hours) durations. The proposed model has eight parameters, or nine if the 24 hours and 72 hours rainfalls are modeled by a bivariate logistic distribution. The likelihood choice formulates hypotheses of independence or dependence between data, and needs a choice of the most representative and non-redundant data. Both independence hypotheses and the bivariate distribution have been used to define likelihoods. In a future work, multivariate distributions with more than two dimensions would provide the model with more information.

The daily rainfall data have been included into the DDF study, through the extremal indexes. Then estimates have been improved since daily series are often longer than hourly series. Namely, the daily Marseille series contains measurements that have not been recorded by the hourly rain gauge, and particularly extreme values. The rainfall intensities on durations 12 hours to 72 hours are linked together and with daily rainfalls, by the extremal indexes of the daily and 24 hourly series, and by the same shape parameter. The quantiles estimated by the models are thus significantly larger than those marginally estimated, in the case of the long durations, proving the important effect of adding the daily data into the model.

A significant difference between independence hypotheses and bivariate logistic distribution has been shown in the case of the 24 hourly and 72 hourly rainfall data. The dependence case, treated with the bivariate logistic distribution, gives a stronger negative scale parameter, close to the parameter $\hat{k}_D$ estimated on the

daily maximal data. The dependence parameter estimated in this case is about 0.26 with a 90% confidence interval equal to [0.21, 0.33], showing the strong dependence between these two durations. Moreover, the likelihood ratio test between models with or without dependence shows that the bivariate logistic distribution is significantly valid relative to the more general asymmetric bivariate logistic distribution, and significantly better than the model under independence hypotheses.

These results have been allowed by the Bayesian framework, which gives a method for defining the posterior distribution of parameters, and includes the prior knowledge on the parameters and the physical behavior of rainfall. Estimations and confidence intervals of parameters have been calculated through a two-step MCMC algorithm. The posterior distributions are generally far from normal and reproduce the heavy tail of the quantiles, proving the usefulness of the Bayesian approach instead of a maximum likelihood estimation of confidence intervals, based on the asymptotic normality of the estimators. Bayesian and maximum likelihood estimations of the medians are quite similar.

Acknowledgement

Météo-France is gratefully acknowledged for providing the Marseille rainfall series: the daily rainfall long series has been reviewed by Météo-France within the European IMFREX project, and the 1 hourly rainfall series belongs to the Météo-France network and has been communicated by Cemagref Aix-en-Provence within a national project of rainfall quantile mapping.

References

Ancona-Navarrete, M. A. and J. A. Tawn (2000). A comparison of methods for estimating the extremal index. Extremes 3(1): 5-38.

Beirlant, J., Y. Goegebeur, J. Segers and J. Teugels (2004). Statistics of Extremes, Theory and Applications.

Bernard, M. M. (1932). Formulas for rainfall intensities of long durations. Trans. ASCE 96: 592-624.

Borga, M., C. Vezzani and G. Dalla Fontana (2005). Regional rainfall depth-duration-frequency equations for an alpine region. Natural Hazards 36(1-2): 221-235.

Burlando, P. and R. Rosso (1996). Scaling and multiscaling models of depth-duration-frequency curves for storm precipitation. Journal of Hydrology 187(1-2): 45-64.

Capéraà, P., A. L. Fougères and C. Genest (1997). A non-parameteric estimation procedure for bivariate extreme value copulas. Biometrika 84: 567-577.

Chow, V. T., D. R. Maidment and L. W. Mays (1988). Applied hydrology.

Coles, S. (2001). An introduction to statistical modeling of extreme values. London.

Coles, S. and L. Pericchi (2003). Anticipating catastrophes through extreme value modelling. Journal of the Royal Statistical Society. Series C: Applied Statistics 52(4): 405-416.

de Lima, M. I. P. and J. Grasman (1999). Multifractal analysis of 15-min and daily rainfall from a semi-arid region in Portugal. Journal of Hydrology 220(1-2): 1-11.

Garcia-Bartual, R. and M. Schneider (2001). Estimating maximum expected short-duration rainfall intensities from extreme convective storms. Physics and Chemistry of the Earth, Part B: Hydrology, Oceans and Atmosphere 26(9): 675-681.

Gelman, A., J. B. Carlin, H. S. Stren and D. B. Rubin (1997). Bayesian Data Analysis. London.

Gumbel, E. J. (1958). Statistics of Extremes. New York, Columbia University Press.

Hershfield, D. M. (1961). Rainfall frequency atlas of the United States for durations from 30 minutes to 24 hours and return periods from 1 to 100 years. U. S. D. o. C. Weather Bureau Technical Paper 40. Washington D.C.

Kendall MG. (1975). Rank correlation methods. London: Griffin.

Kieffer Weisse, A. (1998). Etude des précipitations exceptionnelles de pas de temps court en relief accidenté (Alpes françaises). Méthode de cartographie des précipitations extrêmes. Ph. D. Mécanique des milieux géophysiques et Environnement. Grenoble, Institut National Polytechnique de Grenoble.

Koutsoyiannis, D., D. Kozonis and A. Manetas (1998). A comprehensive study of rainfall intensity-duration-frequency relationships. Journal of Hydrology 206(1-2): 118-135.

Leadbetter, M. R. (1983). Extremes and local dependence in stationary sequences. Zeit. Wahrscheinl. -theorie 65: 291.

Mann, HB. (1945). Nonparametric tests against trend. Econometrica. 13:245-259.

Nadarajah, S., C. W. Anderson and J. A. Tawn (1998). Ordered multivariate extremes. J.R. Statistic. Soc. B 60(2): 473-496.

Pickands, J. (1981). Multivariate extreme value distributions. Bulletin of the International Statistical Institute, Proceedings of the 43rd Session, Buenos Aires.

Pickands, J. (1989). Multivariate negative exponential and extreme value distributions. Extreme Value Theory: Proceedings, Oberwolfach.

Renard, B., V. Garreta and M. Lang (2006). An empirical comparison of MCMC methods used in bayesian inference. Application for regional trend detection. Water Resources Research, submitted.

Robinson, M. E. and J. A. Tawn (2000). Extremal analysis of processes sampled at different frequencies. Journal of the Royal Statistical Society Series B-Statistical Methodology 62: 117-135.

van Montfort, M. A. J. (1997). Concomitants of the Hershfield factor. Journal of Hydrology 194(1-4): 357-365.

Veneziano, D. and P. Furcolo (2002). Multifractality of rainfall and scaling of intensity-duration-frequency curves. Water Resources Research 38(12): 421-4212.

Weiss, L. L. (1964). Ratio of true to fixed-interval maximum rainfall. Journal of the Hydraulic Division of ASCE 90: 77-82.

Comportement asymptotique de la distribution des pluies extrêmes en France

Résumé :

Le comportement des valeurs extrêmes de pluie en France a été analysé au travers de variables locales telles que les maxima annuels ou saisonniers de pluies mesurées sur différents pas de temps entre l'heure et la journée, les valeurs supérieures à un seuil élevé, ou la série temporelle de succession d'averses. Différents modèles, issus de la théorie des valeurs extrêmes uni-variée et bi-variée ou de générateurs stochastiques de pluie, ont été présentés pour étudier le comportement asymptotique de ces variables aléatoires. Dans le cas des séries temporelles d'averses, la persistance dans le temps des valeurs fortes a été modélisée à l'aide d'un processus Markovien. Les incertitudes associées aux différents modèles ont également été analysées, avec des méthodes bayésiennes ou fréquentielles. Nous avons pu valider nos modèles avec de longues séries de mesures pluviométriques, avec des chroniques de pluies horaires et avec des chroniques d'événements pluvieux décrits par des averses fournis par Météo-France et le Cemagref. Dans de nombreux cas, nous avons en particulier noté que la distribution des extrêmes est non bornée, et de queue plus lourde qu'une loi Gumbel ou exponentielle.

Asymptotic behavior of extreme rainfall distribution in France

Abstract :

Rainfall extremes are analyzed by local variables such as annual or seasonal maximum of rainfall for various durations between one hour and one day, the excess over a high threshold, or the temporal series of storm events. Models from univariate or bivariate extreme value analysis or stochastic rainfall generators have been presented to describe the asymptotic behavior of these variables. In the case of temporal storm events series, the temporal persistence has been modeled through a Markovian process. The models uncertainties have been analyzed too, with Bayesian or frequency analysis methods. Models have been validated with long series of daily rainfall, hourly series and storm events series, provided by Météo-France and Cemagref. In particular, it has been often noted that the distribution of the extremes has no bound and a heavier tail than the Gumbel or exponential distribution.

Mots Clés : Théorie des valeurs extrêmes, Maxima annuels, Dépassements de seuil, Générateur stochastique de pluie, Processus Markovien, Analyse bayésienne, Dépendance temporelle.

Discipline : Statistiques appliquées à l'hydrologie

Cemagref, Laboratoire Hydrologie-Hydraulique, 3 bis, quai Chauveau, CP 220 69336 Lyon Cedex 09