



**HAL**  
open science

# Consolidation de l'information hydrologique disponible localement et régionalement pour l'estimation probabiliste du régime des crues

M. Ribatet

► **To cite this version:**

M. Ribatet. Consolidation de l'information hydrologique disponible localement et régionalement pour l'estimation probabiliste du régime des crues. Sciences de l'environnement. Doctorat INP Grenoble et PhD de l'Institut National de la Recherche Scientifique, Québec, CAN, 2007. Français. NNT : . tel-02590034

**HAL Id: tel-02590034**

**<https://hal.inrae.fr/tel-02590034v1>**

Submitted on 15 May 2020

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

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

INSTITUT NATIONAL POLYTECHNIQUE DE GRENOBLE

*N° attribué par la bibliothèque*

\_\_\_\_\_

**THÈSE EN COTUTELLE INTERNATIONALE**

pour obtenir le grade de

**DOCTEUR DE L'INP Grenoble**

et de

**PhD de l'Institut National de la Recherche Scientifique**

**Spécialité : Océan, Atmosphère, Hydrologie (INPG) et Sciences de l'Eau (INRS)**

préparée dans l'Unité de Recherche Hydrologie-Hydraulique, Cemagref (Lyon)

dans le cadre de l'École Doctorale Terre, Univers, Environnement

et de la Chaire en Hydrologie Statistique, INRS (Québec)

présentée et soutenue publiquement

par

**Mathieu Ribatet**

le 10 Décembre 2007

**Consolidation de l'information hydrologique  
disponible localement et régionalement pour  
l'estimation probabiliste du régime des crues**

***DIRECTEURS ET CO-DIRECTEUR DE THÈSE***

*Jean-Michel GRÉSILLON*

*Taha OUARDA*

*Bernard BOBÉE*

**JURY**

M. Dan ROSBJERG	Danmarks Tekniske Universitet	Président et Rapporteur
M. Philippe NAVEAU	University of Colorado / CNRS	Rapporteur
M. Jean-Michel GRÉSILLON	Cemagref	Directeur de thèse
M. Taha OUARDA	INRS	Directeur de thèse
M. Bernard BOBÉE	INRS	Co-directeur de thèse
M. Éric SAUQUET	Cemagref	Co-encadrant de thèse



## Remerciements

Mes premiers remerciements vont naturellement à ceux qui ont encadré cette thèse. C'est donc un grand merci que j'adresse à Éric, Jean-Michel et Caha. Grâce à eux, j'ai pu être un thésard épanoui car libre de prendre mes propres directions de recherche. Je leur en suis donc reconnaissant car je sais qu'un autre encadrement aurait abouti à une thèse différente, peut-être meilleure, mais dans laquelle j'aurais pris moins de plaisir. J'en profite pour remercier également les membres du jury de thèse Philippe Naveau et Dan Rosbjerg pour avoir accepté d'évaluer mon travail. J'espère avoir intégré au mieux leurs remarques au sein de ce document et dans les délais qui m'étaient accordés.

Mes deuxièmes remerciements vont droit à l'équipe de sauvetage administratif : Anne et Hélène du côté français mais aussi Suzanne pour le Québec. Elles ont su non seulement m'apporter toute l'aide dont j'avais besoin mais aussi rester stoïques devant mon manque d'organisation. Pas de jalousie toutefois, je suis sûr de n'avoir épargné aucun côté de l'Atlantique!

Merci aussi à tous ceux qui m'ont permis de décompresser et écouté mes (trop nombreux) grognements aussi durant ces trois années... Merci donc à mes potos d'il y a très très longtemps. Merci Antoine pour ton hospitalité à Roman comme à Valence, petites îles de repos que j'aime beaucoup. Merci aussi pour tes coups de fil directement dans mon bureau de Québec. Merci JJ et Dadoune pour les soirées musicales, j'oublierai jamais ce "Billy Jean" de 20 minutes - toi aussi JJ? Merci Fouca pour sortir ensemble la planche dans les grosses conditions seulement. Merci aussi pôpa Nico pour me laisser gagner au billard et être mon coach vestimentaire... Merci aussi à toutes les buses qui sont trop loin : Fleuron, Mad Rouss', Com-Com, Matt Flon et la Bess'. Merci aussi à Clément et Laure pour (entre autre) les déjeuners du mardi qui vont bien.

Merci à tous ceux que j'ai croisé de plus ou moins loin dans le cadre du travail. Merci à Nadia et Michel. Merci surtout à Pap's Bertrand pour me faire (notamment) péter les scores aux analyses de Cadarache!!! Merci à Mauro pour ces maGGnifiques moments ensemble et le si joli tour de la Gaspésie - regarde un phoque!!! Merci infiniment aussi à face de poils Palo et Valérie bien sûr car sans eux mon année à Québec aurait été nettement moins belle. Un énorme merci à Maaaaaaaaaaaaaaaaaaaaarc pour m'avoir fait partager un moment magique sur la rivière Saguenay. Merci beaucoup à toi Othmane pour ta gentillesse et ton amitié.



Merci à l'équipe so Guinness : jm, Karine, Aurélien et Stéphane. Un gros merci aussi à la Rock n' Eat family Doud, Sandra et Clotilde qui m'ont gentiment accueilli parmi eux cette dernière année. Elles sont nombreuses ces soirées et pourtant toutes différentes. Merci à Aurélie et Kamal, les plus gentils du monde. Merci au difficilement oubliable Benjamin pour la déconne du bureau du fond - félicitation aussi futur CR! Merci à toi Sandhya pour tes mille excuses. Merci aussi à Magali et les petits déjeuners chez Cora... Enfin merci à Christine qui m'a permis de ne jamais faire d'hypoglycémie au bureau ainsi qu'à Bernard pour sa bienveillance.

Bref, il y en a du monde à remercier. C'est sans doute les plus belles citations que je puisse faire, les moins forcées aussi... Pourtant, j'en oublie forcément.

Merci à tous donc et bonjour l'avenir!

# Table des matières

<b>Remerciements</b>	<b>i</b>
<b>Table des Matières</b>	<b>iv</b>
<b>Table des Figures</b>	<b>v</b>
<b>Liste des Tableaux</b>	<b>vii</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Les Besoins Opérationnels . . . . .	5
1.2 Les Champs Explorés par ces Travaux . . . . .	6
1.3 Approche de la Problématique et Plan du Mémoire . . . . .	8
<b>2 Analyse Fréquentielle Locale</b>	<b>9</b>
2.1 Les Estimateurs Statistiques . . . . .	9
2.1.1 Les Estimateurs Purement Locaux . . . . .	10
2.1.2 L'Information Historique . . . . .	11
2.2 Modélisation des Débits par les Processus . . . . .	12
2.2.1 La Méthode du GRADEX . . . . .	13
2.2.2 Les Modèles Pluie-Débit . . . . .	16
2.2.3 La Crue Maximale Probable . . . . .	17
2.3 Conclusions et Objectifs à Atteindre . . . . .	19
<b>3 Le Modèle de l'Indice de Crue</b>	<b>21</b>
3.1 Les Étapes de l'Analyse Fréquentielle Régionale . . . . .	21
3.1.1 La Régionalisation . . . . .	21
3.1.2 L'Homogénéité en Question . . . . .	23
3.1.3 L'Estimation au Site Cible . . . . .	25
3.2 Les Limites de l'Indice de Crue . . . . .	25
3.3 Conclusions et Objectifs à Atteindre . . . . .	27
<b>4 Éléments de Théorie Bayésienne</b>	<b>29</b>
4.1 Vers les Approches Bayésiennes . . . . .	29
4.2 Le Théorème de Bayes et la Loi a Priori . . . . .	29
4.3 L'Inférence Bayésienne . . . . .	30
4.4 Les Intervalles de Crédibilité . . . . .	31
<b>5 Un Modèle Bayésien pour l'Estimation Régionale des Débits</b>	<b>33</b>
5.1 L'Importance du Site Cible et de la Région . . . . .	33
5.2 Les Améliorations . . . . .	34
5.2.1 Un Assouplissement des Hypothèses de l'Indice de Crue . . . . .	34
5.2.2 Une Utilisation Différenciée de l'Information . . . . .	35
5.2.3 Des Propriétés Asymptotiques Établies . . . . .	35
5.2.4 Une Quantification des Incertitudes Rendue Possible . . . . .	36

5.3	L'Article . . . . .	36
5.4	Discussion . . . . .	48
<b>6</b>	<b>Utilisation des Sauts Réversibles pour une Prédiction Plus Robuste des Extrêmes</b>	<b>49</b>
6.1	La Nécessité d'un Nouveau Modèle Régional . . . . .	49
6.2	Les Apports du Nouveau Modèle . . . . .	51
6.3	L'Article . . . . .	52
6.4	Discussion . . . . .	65
<b>7</b>	<b>Modélisation de Tous les Excès Selon une Structure de Dépendance Extrême</b>	<b>67</b>
7.1	Une Utilisation Totale de l'Information Purement Locale . . . . .	67
7.2	La Modélisation de la Dépendance . . . . .	68
7.3	La Modélisation des Hydrogrammes Rendue Possible . . . . .	70
7.3.1	Deux Approches pour l'Estimation de la Dynamique des Crues . . . . .	70
7.3.2	Les Possibilités Offertes par le Modèle Utilisant Tous les Excès . . . . .	73
7.4	L'Article . . . . .	74
7.5	Conclusion et Critique de ce Modèle . . . . .	94
<b>8</b>	<b>Conclusions et Perspectives</b>	<b>97</b>
<b>A</b>	<b>Annexe : Éléments théoriques sur les <math>L</math>-moments</b>	<b>99</b>
<b>B</b>	<b>Annexe : Méthodes d'Ajustement de la GPD</b>	<b>101</b>
B.1	La Méthode des Moments (MOM) . . . . .	101
B.2	La Méthode des Moments Pondérés (PWU & PWB) . . . . .	102
B.3	La Méthode du Maximum de Vraisemblance (MLE) . . . . .	102
B.4	La Méthode du Likelihood Moment (LME) . . . . .	103
	<b>Index</b>	<b>104</b>
	<b>Liste des symboles</b>	<b>105</b>
	<b>Bibliographie</b>	<b>107</b>

# Table des figures

1.1	Inondation de Paris. Janvier 1910. Cartes Postales d'époque. . . . .	1
1.2	Représentation d'une densité de probabilité $f$ , de sa partie centrale et de sa queue de distribution - ici à droite. . . . .	2
1.3	Différentes variables d'étude synthétisant les caractéristiques de crue. $Q_{\text{pte}}$ : débit de pointe, $Q_{\text{moy}}$ et $Q_{\text{dep}}$ : débits moyenné et dépassé en continu sur une durée $d$ . . . . .	5
2.1	Notion de seuil de perception et des différents types d'information historique. Figure inspirée de Brázdil <i>et al.</i> [2006]. . . . .	11
2.2	Illustration de la méthode du GRADEX. . . . .	14
2.3	Schéma synthétique représentant la procédure d'estimation de la distribution des débits à l'aide de transformations pluie-débit. Figure tirée de Arnaud et Lavabre [2002]. . . . .	17
4.1	Définition d'un intervalle de crédibilité $C_\alpha(x)$ connaissant la loi <i>a posteriori</i> $\pi(\theta x)$ . . . . .	31
5.1	Assouplissement des hypothèses du modèle de Dalrymple sur les paramètres de la GPD. Ligne verticale : contraintes de l'indice de crue. Densités : contraintes du modèle proposé. . . . .	34
5.2	Précision de l'information régionale à l'aide des données au site cible. Lignes en pointillées : lois a priori. Lignes pleines : lois <i>a posteriori</i> . Lignes verticales : valeurs de références. . . . .	35
6.1	Evolution de l'hyperparamètre $p_\xi$ en fonction de la statistique $H_1$ . . . . .	52
7.1	Perte d'information : passage d'une chronique de débit à un échantillon de taille réduite. Approches classiques des maxima annuels et dépassements au dessus d'un seuil. L'Ardières à Beaujeu. . . . .	68
7.2	Différence d'utilisation de l'information entre l'approche classique POT (à gauche) et la modélisation de tous les excès (à droite). L'Ardières à Beaujeu. Losange rouge : observation exploitée lors de la procédure d'estimation. . . . .	69
7.3	Définition des 4 différentes régions du couple $(y_{i-1}, y_i)$ . . . . .	70
7.4	Illustration de la description des hydrogrammes de crue par une densité de probabilité. Panneau de gauche : hydrogamme brut. Panneau de droite : hydrogramme modifié et d'aire unité. Figure inspirée de Yue <i>et al.</i> [2002]. . . . .	71
7.5	Convergence des distributions des débits moyens associés aux durées $d_1$ et $d_2$ . . . . .	72
7.6	Construction de l'HSMF à partir des courbes QdF des débits continuellement dépassés. . . . .	73
7.7	Chronique de débits journaliers observés (à gauche) et simulés (à droite). La Sioule à St-Priest-des-Champs. Seuil à $46 \text{ m}^3 \cdot \text{s}^{-1}$ . . . . .	74
7.8	Valeurs supérieures à $46 \text{ m}^3 \cdot \text{s}^{-1}$ observées (à gauche) et simulés (à droite). La Sioule à St-Priest-des-Champs. . . . .	95



# Liste des tableaux

2.1	Synthèse sur la performance des différents estimateurs de la GPD. . . . .	11
2.2	Essai de détermination de la période de retour de la CMP. . . . .	18
6.1	Performance de chaque estimateur pour l'estimation des quantiles $Q_2$ , $Q_{10}$ et $Q_{20}$ . Nombre de régions simulées : 1000. Taille d'échantillon au site cible : 20. . . . .	50
6.2	Performance de chaque estimateur pour l'estimation des quantiles $Q_{50}$ , $Q_{100}$ et $Q_{1000}$ . Nombre de régions simulées : 1000. Taille d'échantillon au site cible : 20. . . . .	50
6.3	Performance de chaque estimateur pour l'estimation des paramètres de la GPD. Nombre de régions simulées : 1000. Taille d'échantillon au site cible : 20. . . . .	51



# 1 Introduction

**D**EPUIS quelques millénaires, l'homme a investi et aménagé le territoire dans le but de le rendre plus hospitalier et facile à vivre - selon ses critères. Le besoin qu'il a de l'eau l'a amené à s'installer notamment aux abords de fleuves ou rivières et à subir leurs caprices répétés. La notion de protection face aux risques naturels est donc présente depuis bien longtemps dans les esprits. Toutefois, mettre en oeuvre des mesures de protection contre des événements extrêmes et aléatoires n'est pas une tâche aisée puisqu'il s'agit de :

*« Prédire ou connaître le comportement de processus qui n'ont été que très peu, voire pas encore observés... »*

La société ayant énormément évolué, il en est de même de ses attentes. Les premières préoccupations étaient avant tout liées à la survie face à un environnement devenant hostile. Aujourd'hui, à cette préoccupation de survie, la société actuelle doit ajouter de nouvelles contraintes : l'industrialisation et, plus récemment, la société de l'information, ont rendu l'homme et son environnement plus vulnérables mais aussi plus conscients des risques encourus. En effet, le développement industriel nécessite de prendre en considération des événements de plus en plus extrêmes afin de ne pas tendre vers un effet multiplicatif des incidents - *i.e.* catastrophes naturelles conduisant, par exemple, à un incident industriel majeur. De même, les ouvrages de protection peuvent malgré eux jouer un rôle aggravant, comme dans le cas de rupture de digue.

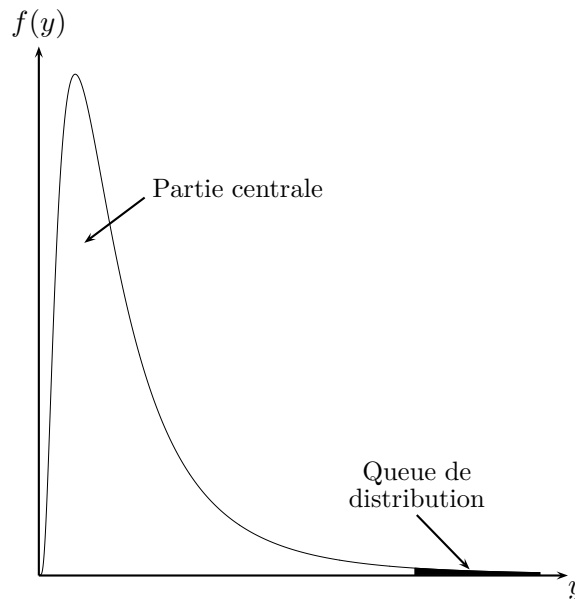
Les autorités se doivent donc de légiférer sur le niveau de protection dans le but de définir des normes de sécurité adéquates. Pour les opérationnels, ces textes ont un impact conséquent puisqu'on leur demande de construire des ouvrages dimensionnés pour le *peu probable*. Ainsi, l'article L. 512-1 du Code de l'Environnement pour la Prévention des Polluants, des Risques et des Nuisances, en vigueur actuellement sur l'ensemble du territoire français ([www.legifrance.fr](http://www.legifrance.fr)), stipule que :

*« Le demandeur fournit une étude de dangers qui précise les risques auxquels l'installation peut exposer, directement ou indirectement, les intérêts visés à l'article L. 511-1 en cas d'accident, que la cause soit interne ou externe à l'installation. Cette étude donne*



**Fig. 1.1:** Inondation de Paris. Janvier 1910. Cartes Postales d'époque.





**Fig. 1.2:** Représentation d'une densité de probabilité  $f$ , de sa partie centrale et de sa queue de distribution - ici à droite.

*lieu à une analyse de risques qui prend en compte la probabilité d'occurrence, la cinétique et la gravité des accidents potentiels selon une méthodologie qu'elle explicite. Elle définit et justifie les mesures propres à réduire la probabilité et les effets de ces accidents. »*

Comme mentionné dans la législation, une approche pour répondre à ce type de problèmes nous vient du domaine des probabilités où la notion de risque est entièrement intégrée. La Figure 1.2 représente la densité  $f$  d'une variable aléatoire  $Y$  régissant le comportement d'un processus aléatoire, *i.e.* que l'on ne peut prédire avec exactitude. Les normes de sécurité imposent de dimensionner un ouvrage pour un certain niveau de risque ; niveau de risque représenté par l'aire mise en évidence sur la Figure 1.2. D'un point de vue mathématique, le problème consiste à connaître la densité du processus aléatoire afin d'estimer cette zone de risque ; ou en d'autres termes, de modéliser la queue de distribution de la variable aléatoire  $Y$ .

Ainsi, au problème de *dimensionnement du peu probable* équivaut un autre problème plus classique : *l'estimation d'une densité de probabilité*. Toutefois, l'analyse ne portant que sur une région à faible densité, l'étude des extrêmes diffère d'une étude statistique plus *conventionnelle*. Alors qu'une analyse statistique classique ajuste la distribution sur toutes les données ; pour une étude de risque, cette manière de procéder n'est plus pertinente. En effet, lors d'une analyse statistique classique :

1. les paramètres de la distribution sont ajustés essentiellement par les *valeurs centrales* de la distribution ;
2. la qualité prédictive du modèle sera également jugée par ces valeurs centrales ;
3. différents modèles, aux performances égales dans la partie centrale, conduiront à des extrapolations très différentes.

De même, si seuls les extrêmes sont à modéliser, pourquoi s'occuper de cette partie centrale ? Comme le montre l'exemple suivant, s'attacher à la modélisation de la partie centrale de la distribution peut conduire à des erreurs considérables sur la queue de distribution.

**Exemple 1.** Propagation des erreurs pour la loi du maximum

Soit  $Y_1, \dots, Y_n$   $n$  variables aléatoires indépendantes, identiquement distribuées de loi  $F$ .

Soit  $M_n = \max \{Y_1, \dots, Y_n\}$ . Alors

$$\Pr [M_n \leq y] = \prod_{i=1}^n \Pr (Y_i \leq y) = F(y)^n \quad (1.1)$$

Or dans la pratique,  $F$  n'est jamais connue et doit être remplacée par son estimation  $\hat{F}$ . Ainsi,  $\Pr[M_n \leq y]$  est estimée par :

$$\hat{\Pr} [M_n \leq y] = \hat{F}(y)^n \quad (1.2)$$

En remarquant que l'on peut écrire :

$$F(y) = (1 + \varepsilon(y)) \hat{F}(y), \quad \varepsilon(y) \in \mathbb{R} \quad (1.3)$$

où  $\varepsilon(y)$  est l'erreur d'estimation sur  $F(y)$ , on a :

$$\frac{\Pr [M_n \leq y]}{\hat{\Pr} [M_n \leq y]} = (1 + \varepsilon(y))^n \quad (1.4)$$

L'erreur d'estimation  $\varepsilon(y)$  commise sur  $F(y)$  est ainsi élevée à la puissance  $n$  lors du passage à la loi du maximum  $M_n$ . En particulier, pour  $n = 365$  et une erreur  $\varepsilon(y)$  de +1% sur l'estimation de  $F(y)$  entraîne une erreur de l'ordre de +3700% pour l'estimation de  $\Pr[M_n \leq y]$ .

Afin d'éviter l'effet exponentiel des erreurs illustré par l'Exemple 1, le principe d'une analyse statistique des valeurs extrêmes consiste à se concentrer uniquement sur la modélisation des valeurs les plus fortes (ou faibles) d'une variable aléatoire. En particulier, elle se distingue d'une étude statistique classique puisque :

1. les paramètres du modèle sont ajustés à partir des valeurs extrêmes uniquement ;
2. la qualité prédictive sera évaluée sur les valeurs extrêmes également.

La Théorie des Valeurs Extrêmes (**EVT**) se base principalement sur deux résultats. Ces résultats nous donnent le comportement asymptotique de la variable aléatoire  $M_n$  ou des dépassements au dessus d'un seuil  $u$ . La puissance de ces résultats provient du fait qu'il n'est pas nécessaire de connaître la loi du processus  $Y$  que l'on souhaite prédire. D'une manière générale, l'EVT est analogue au Théorème Central Limite (**TCL**) mais pour les extrêmes. Là où le TCL montre que la moyenne empirique de la variable aléatoire  $Y$  converge vers une loi Normale (indépendamment de la loi de  $Y$  et lorsque  $\mathbb{E}[Y]$  et  $\text{Var}[Y]$  existent) ; l'EVT établit des résultats analogues mais pour les valeurs extrêmes de  $Y$ .

**Théorème 1.1.** *Théorème du Type Extremal [Fisher et Tippett, 1928; Leadbetter, 1983]*  
*S'il existe des suites de constantes  $\{a_n\}_{n \geq 0}$  et  $\{b_n\}_{n \geq 0}$  telles que :*

$$\Pr \left[ \frac{M_n - b_n}{a_n} \leq y \right] \longrightarrow G(y), \quad n \rightarrow +\infty \quad (1.5)$$

où  $G$  est une fonction de répartition non dégénérée. Alors  $G$  est l'une des 3 fonctions de répartition suivantes :

$$\text{Gumbel} \quad G(y) = \exp[-\exp(-y)], \quad y \in \mathbb{R} \quad (1.6a)$$

$$\text{Fréchet} \quad G(y) = \begin{cases} 0, & y \leq 0 \\ \exp(-y^{-\alpha}), & y > 0, \alpha > 0 \end{cases} \quad (1.6b)$$

$$\text{Weibull Négative} \quad G(y) = \begin{cases} \exp(-(-y)^\alpha), & y < 0, \alpha > 0 \\ 1, & y \geq 0 \end{cases} \quad (1.6c)$$

## 1. Introduction

---

Néanmoins, il est plus pratique de reformuler ce théorème pour obtenir la convergence vers une unique loi : la loi Généralisée des Valeurs Extrêmes (**GEV**) introduite simultanément par [Von Mises \[1954\]](#) et [Jenkinson \[1955\]](#).

**Théorème 1.2.** *Théorème du Type Extremal (version GEV)*

*S'il existe des suites de constantes  $\{a_n\}_{n \geq 0}$  et  $\{b_n\}_{n \geq 0}$  telles que :*

$$\Pr \left[ \frac{M_n - b_n}{a_n} \leq y \right] \longrightarrow G(y), \quad n \rightarrow +\infty \quad (1.7)$$

*où  $G$  est une fonction de répartition non dégénérée. Alors  $G$  admet pour fonction de répartition :*

$$G(y) = \exp \left[ - \left( 1 + \xi \frac{y - \mu}{\sigma} \right)_+^{-1/\xi} \right] \quad (1.8)$$

*où  $z_+ = \max(0, z)$  et  $\sigma > 0$ .  $\mu$ ,  $\sigma$  et  $\xi$  correspondent respectivement aux paramètres de position, d'échelle et de forme.*

Ce théorème affirme donc que :

*« Quelle que soit la distribution de  $Y$ , la loi du maximum  $M_n$  converge vers une unique loi de probabilité non dégénérée : la GEV. »*

Il existe un résultat similaire pour la modélisation des dépassements au dessus d'un seuil  $u$ . Ces dépassements sont naturellement représentés par la variable aléatoire conditionnelle  $Y|Y > u$ .

**Théorème 1.3.** *Théorème des Valeurs Supérieures à un Seuil [Pickands, 1975]*

*S'il existe des suites de constantes  $\{a_n\}_{n \geq 0}$  et  $\{b_n\}_{n \geq 0}$  telles que :*

$$\Pr \left[ \frac{M_n - b_n}{a_n} \leq y \right] \longrightarrow G(y), \quad n \rightarrow +\infty \quad (1.9)$$

*où  $G$  est une GEV. Alors,*

$$\Pr [Y \leq y | Y > u] \longrightarrow H(y), \quad u \rightarrow u_f \quad (1.10)$$

*où  $u_f = \sup \{u \in \mathbb{R} : \Pr[Y \leq u] < 1\}$  et*

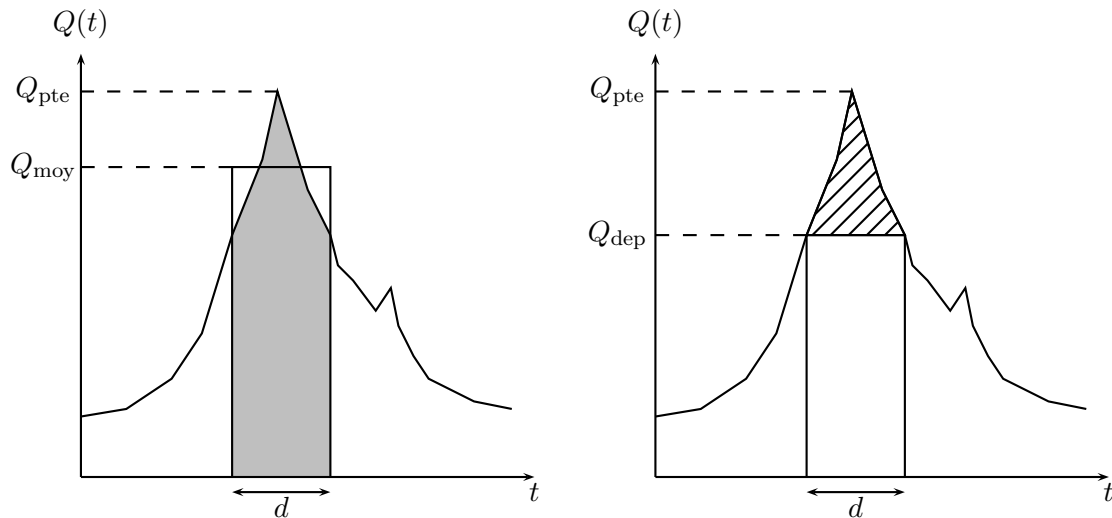
$$H(y) = 1 - \left[ 1 + \xi \frac{y - u}{\sigma} \right)_+^{-1/\xi} \quad (1.11)$$

*où  $z_+ = \max(0, z)$  et  $\sigma > 0$ .  $\sigma$  et  $\xi$  correspondent respectivement aux paramètres d'échelle et de forme. Le paramètre  $u$  sera indifféremment appelé seuil ou paramètre de position.*

De même que pour la GEV, ce théorème affirme que :

*« La loi de  $Y|Y > u$  converge vers une unique loi non dégénérée : la loi Pareto Généralisée (**GPD**). »*

Le principe d'une modélisation des valeurs extrêmes consiste donc à exploiter les théorèmes issus de l'EVT. En particulier, nous supposons que les résultats asymptotiques deviennent vrais (ou plus rigoureusement approximativement vrais) pour des niveaux finis mais suffisamment élevés. En conséquence, puisque nous ne connaissons pas la distribution  $F$  de l'Équation (1.1) et que son erreur d'estimation  $\varepsilon$  deviendra non négligeable pour la loi du maximum (*cf.* Exemple 1), nous remplaçons avantageusement la modélisation à partir des observations  $y_i$  au profit des maximums  $m_{n,i}$  de  $n$  valeurs consécutives et ajustés sur une GEV - par exemple par maximum annuels par  $m_{365,i}$ .



**Fig. 1.3:** Différentes variables d'étude synthétisant les caractéristiques de crue.  $Q_{pte}$  : débit de pointe,  $Q_{moy}$  et  $Q_{dep}$  : débits moyenné et dépassé en continu sur une durée  $d$ .

Toutefois, bien que l'EVT fournisse une théorie très générale et puissante ; sa généralité est également un inconvénient. En effet, quelle que soit la nature du processus  $Y$  que l'on souhaite modéliser (série financière, chronique de débits, ...), les mêmes théorèmes seront toujours appliqués. Du fait du nombre restreint d'observations disponibles en hydrologie pour l'ajustement du modèle, il paraît donc intéressant de spécifier l'EVT pour les variables que l'on souhaite étudier afin d'obtenir des modèles plus performants. Notamment, pour ce qui nous concerne ici, l'intégration de connaissances sur les processus générateurs de crues dans la procédure d'estimation serait une réelle avancée puisque l'on passerait d'une EVT générique à une EVT ciblée pour l'estimation des quantiles de crue.

## 1.1 Les Besoins Opérationnels

Les praticiens se doivent d'apporter des solutions efficaces en termes d'aménagements pour répondre conjointement à l'exigence toujours croissante de la société et aux textes législatifs de plus en plus stricts. Notre projet de recherche se place dans ce contexte : mettre à la disposition des opérationnels un grand éventail d'approches permettant une réponse adaptée à leurs besoins.

Lors d'un projet, le praticien sera amené à quantifier les caractéristiques de crue du bassin d'étude. Ces caractéristiques sont multiples et il est préférable de ne pas les résumer par une unique valeur. Ainsi, l'étude fréquentielle pourra porter sur les débits : (a) de pointe, (b) moyennés sur une durée  $d$  ou encore (c) dépassés en continu sur une durée  $d$ . Les débits moyennés sont une représentation du volume écoulé, alors que les débits dépassés en continu sont une mesure de l'intensité de la crue. La Figure 1.3 est une illustration de ces différentes caractéristiques de crue.

De plus, l'hydrologue s'attachera souvent à connaître la dynamique de la crue du site d'étude (*e.g.* bassin réactif ou lent) et à établir un hydrogramme de projet. Avec cet hydrogramme de projet, l'hydrologue dispose de toutes les informations nécessaires (*e.g.* estimation des débits caractéristiques, de la durée, du volume écoulé) au bon dimensionnement de l'ouvrage hydraulique.

Bien souvent, les études porteront sur des sites partiellement jaugés, *i.e.* avec une période d'enregistrement inférieure à 15 ans, augmentant ainsi la difficulté de l'analyse. En conséquence, lors de notre travail de recherche, nous nous sommes efforcé d'élaborer des modèles probabilistes adaptés aux stations dites partiellement jaugées (*i.e.* stations pour lesquelles le nombre de données disponible est restreint) et permettant d'estimer de manière robuste toutes les caractéristiques de crue

nécessaires pour l'analyse opérationnelle.

### 1.2 Les Champs Explorés par ces Travaux

Avant de rentrer dans le vif du sujet, il paraît nécessaire d'énoncer les points sensibles sur lesquels nous ne travaillerons pas mais dont nous avons connaissance. Ces éléments mis à l'écart sont : (a) la possible non stationnarité des variables d'étude et (b) les incertitudes sur les données collectées.

#### Le Problème de la Stationnarité

Dans tout notre travail de recherche, nous nous placerons dans un contexte *stationnaire*, *i.e.* le changement climatique n'influe en rien sur les variables d'étude. L'impact du changement climatique sur les variables de pluie ou de débit est, à lui seul, un sujet de recherche.

Les politiques et plus généralement la société toute entière n'ont pris conscience qu'assez récemment des répercussions possibles d'un dérèglement climatique. Ainsi, dernièrement, de nombreuses études ont été menées afin d'identifier les signes d'un changement climatique mais aussi de mieux connaître les scénarios futurs [GIECC, 2007].

Si pour les températures les scientifiques sont unanimes pour conclure quant à un réchauffement au cours du XX<sup>e</sup> siècle [Bloomfield, 1992; Hansen *et al.*, 2001; Brohan *et al.*, 2006]. Pour les précipitations et les débits, il n'existe pas de consensus à l'échelle mondiale. Les simulations issues des Modèles de Circulation Générale (GCM) montrent qu'une augmentation de la température induit une augmentation du volume et de l'intensité de l'eau précipitée. Des analyses statistiques sur des chroniques de pluies semblent confirmer ces résultats issus des GCMs [Groisman et Easterling, 1994; Karl et Knight, 1998].

La variable débit est certainement l'une des plus incertaines face à l'augmentation des gaz à effet de serre. Si les premières analyses régionales sur la non stationnarité des séries plaident pour la présence de tendances dans les chroniques de débit [Lins et Slack, 1999; Olsen *et al.*, 1999], elles ont souvent négligé l'effet de la corrélation présente entre les stations d'étude. Les dernières analyses considérant la dépendance intersite montrent que ces tendances sont nettement moins évidentes, voire inexistantes [Douglas *et al.*, 2000; Renard *et al.*, 2006b].

Toutefois, de nombreux travaux restent à engager afin de dresser un bilan fiable sur la présence de tendances au sein des chroniques de débit et leurs liens possibles avec un quelconque changement climatique. Dans cette lignée, Burn et Hag Elnur [2002] s'abstiennent de dresser toutes conclusions sur les tendances détectées sur certaines variables hydrologiques au Canada.

Les problèmes théoriques ne sont pas en reste également, notamment sur la sélection d'un test approprié pour la détection de tendances [Kundzewicz et Robson, 2004]. L'un des tests les plus utilisés pour la détection de ces tendances, *i.e.* le test non paramétrique de Mann-Kendall, comporte quelques faiblesses. Par exemple, Bayazit et Önöz [2007] posent le problème du *pré-blanchissement* des séries temporelles lors de l'application du test de Mann-Kendall. En effet, ce test étant valide pour des observations indépendantes, ce qui est loin d'être le cas pour les séries temporelles, l'étape de *pré-blanchissement* essaye de supprimer, ou du moins réduire, les biais en lien avec cette dépendance. Selon Bayazit et Önöz [2007], l'étape de *pré-blanchissement* admet les propriétés suivantes :

- « *When there is no trend, the probability of Type I error is significantly reduced, almost to the theoretically correct value ;* »
- « *When a trend exists, however, the power of the test is decreased as compared with the power before pre-whitening, so that sometimes it may cause a significant trend not to be detected.* »

La seule détection de tendances au sein d'une série temporelle n'est pas suffisante. Si les données semblent non stationnaires, il faut alors intégrer cette dérive lors de nos estimations. Ainsi, de

nouveaux cadres probabilistes ont été développés pour intégrer la non-stationnarité dans un calcul de prédétermination des pluies et des débits. Engeland *et al.* [2005] ont proposé l'utilisation d'une GEV dont les paramètres peuvent évoluer avec le temps. Renard *et al.* [2006b] ont développé une approche Bayésienne basée sur un mélange de différents modèles temporels (stationnaire, avec rupture ou tendance), pondérés par leur probabilité respective. Cette approche a ensuite été généralisée au cas d'évolutions régionalement cohérentes, en prenant en compte la dépendance spatiale entre sites [Renard *et al.*, 2006a]. Dernièrement, Cunderlik et Ouarda [2006] ont élaboré, dans un contexte régional non stationnaire, un modèle pour l'estimation des débits instantanés et moyens.

Au vu de ce constat, il paraît clair que le problème lié à la non stationnarité des données est un sujet de recherche à part entière et délicat. Plutôt que d'approcher ce problème de manière superficielle, nous avons décidé de ne pas l'aborder afin de porter nos efforts de modélisation sur l'estimation des quantiles caractéristiques de crue dans un cadre partiellement jaugé et en contexte supposé stationnaire. De plus, pour des stations partiellement jaugées (*i.e.* moins de 15 années d'enregistrement), il aurait été difficile d'identifier une quelconque tendance significative dans nos séries temporelles.

### Les Incertitudes sur les Données

Tout comme pour le possible problème de stationnarité, nous négligerons dans la suite de nos travaux les incertitudes inhérentes à l'hydrométrie. En effet, nous nous placerons dans le contexte où nos données d'entrée de modèle (*i.e.* les mesures de débits) ne comportent pas de fortes erreurs, ni de biais systématiques. De sorte que lorsque nous évoquerons les incertitudes, nous parlerons d'incertitudes sur la prédétermination de débits caractéristiques de crue.

Bien évidemment, les mesures de débit ne sont pas exemptes de sources d'erreur. Lang *et al.* [2006] dressent une liste des causes d'erreur liées aux mesures de débits. Bien que nous cherchions à modéliser les valeurs des débits extrêmes, il faut garder en mémoire que les données initiales ne sont pas des débits mais des hauteurs d'eau. La *courbe de tarage*, ajustée à partir de jaugeages *i.e.* mesures simultanées des hauteurs d'eau et des débits, est une courbe généralement donnée sous la forme :

$$Q = aH^b, \quad b \neq 1 \quad (1.12)$$

où  $Q$  est la valeur de débit,  $H$  la hauteur d'eau et  $a$  et  $b$  des coefficients à ajuster.

Trois types d'erreur impactant l'estimation des débits transités peuvent être identifiés ; celles liées à : (a) l'échantillon de jaugeages pour l'établissement de la courbe de tarage, (b) l'évolution morphologique du lit de la rivière et (c) l'instrumentation - *i.e.* le dispositif de suivi de la hauteur. La qualité d'une courbe de tarage résulte donc de la qualité de la relation hauteur/débit et des jaugeages permettant son ajustement. Concernant la première source d'erreur, Renouf *et al.* [2005], lors d'une étude sur la qualité des courbes de tarage de la banque HYDRO pour les débits de crue, établissent une remarque sans équivoque :

« Pour la majorité des stations, le débit correspondant à la hauteur maximale jaugée est de période de retour inférieure à 2 ans. Très peu de stations sont jaugées au delà de la décennale (9%) et parmi ces stations, les grands bassins sont les mieux renseignés [...] »

Pour la thématique qui nous concerne, *i.e.* les débits extrêmes à l'échelle de la France, il manque donc cruellement de jaugeages permettant d'établir une relation fine hauteur/débit pour la gamme des forts débits estimés ; et les valeurs des débits enregistrés proviennent donc bien souvent d'une extrapolation de la courbe de tarage.

Quant aux erreurs de type (b), Lang *et al.* [2006] soulignent le problème de la pertinence des mesures hydrométriques en période de crue puisque les fortes valeurs de débits peuvent entraîner : des phases de transport solide modifiant le niveau de fond du lit ou des perturbations hydrauliques

au droit des capteurs altérant la qualité des mesures. De même, l'évolution de la rivière à l'échelle de plusieurs années (*e.g.* envasement progressif, incision du lit, aménagements hydrauliques), remet en question l'existence d'une unique courbe de tarage sur toute la période d'instrumentation et demande une mise à jour régulière de la relation hauteur/débit ainsi que de la définition des périodes de validité de telles relations.

Enfin, pour le troisième type d'erreur et même si les incertitudes associées sont moindres, les appareils hydrométriques (*e.g.* capteurs piézorésistifs, bulle à bulle, ultra-sons, radars) peuvent eux aussi introduire un biais. Par exemple, le taux de matières en suspension, la réactivité du capteur sont des facteurs pouvant altérer la qualité des données.

### 1.3 Approche de la Problématique et Plan du Mémoire

L'hydrologie statistique est souvent critiquée, notamment en France, comme étant peu rigoureuse car elle semble contradictoire avec les statistiques classiques pour lesquelles un jeu de données important est nécessaire. Le changement climatique est également une nouvelle source de critique puisqu'elle contredit la fameuse hypothèse *identiquement distribué*. Le changement climatique, s'il contredit en effet l'hypothèse de stationnarité, n'est pas incompatible avec les approches statistiques. Nombreux sont les auteurs qui ont développé des modèles autorisant une dérive temporelle sur les paramètres du modèle [Douglas *et al.*, 2000; Burn et Hag Elnur, 2002; Kundzewicz et Robson, 2004; Renard *et al.*, 2006a; Cunderlik et Ouarda, 2006].

En fait, l'hydrologie statistique a bel et bien sa place puisqu'elle permet une estimation rigoureuse des plus forts quantiles. D'ailleurs, les nombreux développements autour de l'EVT depuis les 30 dernières années [Leadbetter, 1983; Resnick, 1987; Tawn, 1988; Smith *et al.*, 1997; Ledford et Tawn, 1996; Coles et Pauli, 2002; Falk et Michel, 2006] ont permis d'asseoir plus encore sa légitimité. Nos travaux se placent donc dans cette lignée.

Ce mémoire de thèse est organisé de la manière suivante. Le Chapitre 2 présente les approches les plus fréquemment rencontrées pour l'analyse fréquentielle des quantiles de crue à partir des données disponibles au site d'étude. Le Chapitre 3 est consacré à l'estimation régionale des quantiles de crue et plus particulièrement au modèle de l'indice de crue. Le Chapitre 4 conclut la revue de littérature et les aspects théoriques par une brève introduction aux statistiques Bayésiennes.

Les Chapitres 5, 6 et 7 présentent les trois articles déjà publiés ou en cours de révision. Ainsi, le Chapitre 5 présente un modèle Bayésien pour l'estimation régionale des quantiles de crue. Le Chapitre 6 présente une extension du modèle précédent permettant une estimation plus précise pour les quantiles les plus extrêmes. Enfin, le Chapitre 7 présente un modèle innovant permettant l'analyse fréquentielle des débits extrêmes mais également d'estimer la dynamique des crues.

Le Chapitre 8 dresse les conclusions sur les travaux engagés et propose quelques pistes pour améliorer les modèles introduits durant cette thèse. Enfin, quelques résultats théoriques mais néanmoins pratiques sont reportés en Annexes A et B.

# Analyse Fréquentielle Locale

# 2

DEPUIS les premiers travaux de Gumbel [1960], les lois de Gumbel et Exponentielle sont couramment utilisées pour la modélisation des maxima annuels ou des événements supérieurs à un seuil. Elles présentent l'avantage d'être simples à appliquer avec un ajustement (notamment par maximum de vraisemblance) possédant une solution analytique. Ces deux lois sont actuellement largement utilisées dans les bureaux d'étude. La banque de données hydrométriques HYDRO ([www.hydro.eaufrance.fr](http://www.hydro.eaufrance.fr)), dans un souci d'uniformisation, a également considéré la loi de Gumbel comme loi de référence pour l'estimation des quantiles de crue sur tout le réseau hydrométrique français. On comprend ainsi pourquoi les opérationnels, tout comme les enseignants, restent attachés à ces deux lois.

Il est intéressant de noter qu'à la fois la loi Exponentielle et celle de Gumbel sont des cas particuliers des lois GPD et GEV respectivement - *i.e.* avec un paramètre de forme  $\xi = 0$ . Néanmoins, depuis peu, plusieurs auteurs soulignent le risque de la modélisation par la loi Gumbel/Exponentielle [Koutsoyiannis *et al.*, 1998; Coles et Pericchi, 2003; Muller, 2006]. En effet, comme ce sont des cas particuliers de la GEV/GPD, ces modèles supposent que le paramètre de forme  $\xi$  est nul. Notamment, la modélisation des extrêmes par des lois de Gumbel et Exponentielle néglige considérablement les incertitudes sur ce paramètre de forme  $\xi$  en imposant une valeur n'étant probablement pas la bonne - et même presque sûrement fautive au sens mathématique. Il a été montré que l'hypothèse  $\xi = 0$  pouvait sous-estimer dramatiquement les plus forts quantiles [Koutsoyannis, 2004; Coles et Pericchi, 2003]. Ainsi, à l'heure actuelle, l'usage systématique de ces deux lois pour l'estimation des quantiles extrêmes est déconseillé dans la littérature scientifique tout en restant pourtant bien ancré dans le monde opérationnel.

D'après la formulation des lois GEV et GPD, les extrapolations sont essentiellement dirigées par le paramètre de forme  $\xi$ . Malheureusement, ce paramètre est également le plus difficile à estimer des trois. Nous voyons donc apparaître l'un des plus épineux problèmes de l'analyse des valeurs extrêmes : obtenir des extrapolations fiables. Une des solutions les plus rencontrées est d'incorporer de l'information supplémentaire dans la procédure d'estimation afin de consolider les estimations. Notre travail se situe précisément dans cette perspective.

Avant de présenter les nouveaux développements, nous présentons ici quelques modèles (ou classes de modèles) en provenance de l'hydrologie ou des statistiques théoriques et permettant d'obtenir des estimations plus robustes. En ce sens, une vision objective sur ces modèles sera livrée afin de mettre en valeur les objectifs que nous désirons atteindre.

## 2.1 Les Estimateurs Statistiques

Cette section présente plusieurs estimateurs issus des dernières avancées en statistiques théoriques. L'objectif, purement mathématique, est de proposer de nouveaux estimateurs plus performants en termes de convergence et/ou de propriétés asymptotiques dans le but de mieux estimer les quantiles extrêmes. Deux types d'approche particulières ont retenu notre attention : (a) les



estimateurs purement locaux et (b) ceux faisant intervenir l'information historique.

### 2.1.1 Les Estimateurs Purement Locaux

La famille des estimateurs **purement locaux** profite des derniers résultats théoriques pour l'estimation des débits caractéristiques. Ils se distinguent des autres approches présentées plus loin du fait qu'ils ne font appel qu'aux données contemporaines de débits disponibles au site d'étude, *i.e.* issues des enregistrements continus.

Les estimateurs les plus fréquemment utilisés en hydrologie statistique sont le maximum de vraisemblance (**MLE**), la méthode des moments (**MOM**) et des moments pondérés (**PWM**). De nombreux auteurs ont étudié par simulation la performance de ces estimateurs pour la GPD afin d'établir un classement des meilleures options. Ainsi, [Hosking et Wallis \[1987\]](#) ont montré que l'estimation par MLE est moins performante que par les méthodes MOM et PWM lorsque la taille de l'échantillon est inférieure à 500. De plus, lors de leur étude, ils ont montré que l'approche par PWM est plus performante que la méthode MOM lorsque le paramètre de forme  $\xi$  est strictement positif (*i.e.* distribution non bornée supérieurement) et plus particulièrement lorsque  $\xi \approx 0.2$ . En revanche, lorsque la distribution est bornée supérieurement (*i.e.*  $\xi < 0$ ), la méthode MOM est à préconiser.

[Coles et Dixon \[1999\]](#) montrent que la supériorité des PWM par rapport au MLE est illusoire. En effet, par définition des PWM, il est supposé implicitement que le paramètre de forme vérifie  $\xi < 1$  et contraint ainsi la distribution à avoir une moyenne finie. Ils montrent ainsi qu'en introduisant la même connaissance au sein d'une vraisemblance pénalisée, l'estimation par maximum de vraisemblance pénalisée (**MPLE**) était tout aussi performante que les PWM pour les petits échantillons. Récemment, [Diebolt \*et al.\* \[2004\]](#) ont introduit les moments pondérés généralisés afin de supprimer la restriction au domaine  $\xi < 1$  et d'obtenir la normalité asymptotique au delà de  $\xi < 0.5$ . Cette généralisation prend un réel intérêt dans le cadre de données très extrêmes comme en assurance ; mais perd un peu de son importance pour des applications hydrologiques pour lesquelles  $\xi < 0.5$ .

[Davison et Smith \[1990\]](#) ont montré également que les estimateurs PWM étaient plus sensibles à une contamination que par le MLE. Par contamination, nous entendons qu'une certaine proportion de nos données n'est plus identiquement distribuée avec la partie restante de l'échantillon.

Certains auteurs proposent donc des estimateurs plus robustes à une contamination des observations. Par exemple, [Dupuis \[1998\]](#) et [Peng et Welsh \[2001\]](#) proposent respectivement l'estimateur Optimal Bias Robust Estimator (**OBRE**) et Medians Estimator (**ME**). [Juárez et Schucany \[2004\]](#) proposent le Minimum Density Power Divergence Estimator (**MDPDE**) et le comparent au MLE et aux estimateurs de [Dupuis](#) et [Peng et Welsh](#). Ils ont montré qu'en l'absence de contamination, l'estimateur MLE est le plus efficace. En revanche, lorsque la contamination devenait importante, les estimateurs OBRE et MDPDE sont plus performants. L'estimateur ME, en revanche, a montré une performance trop faible.

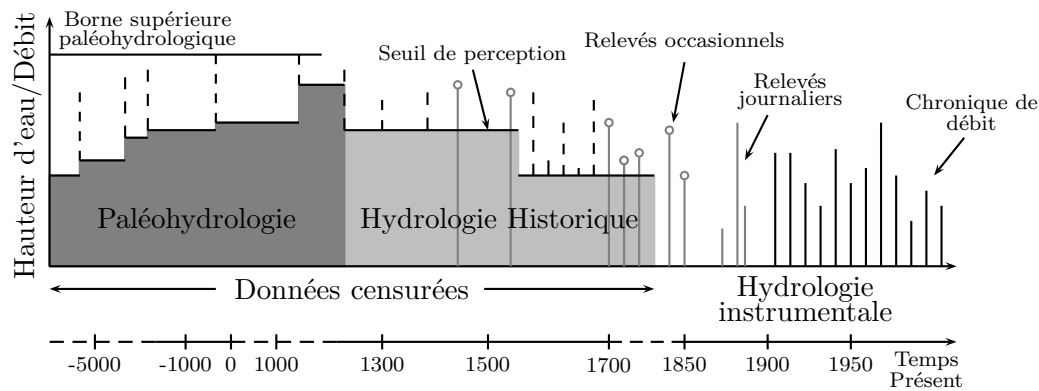
Enfin, certains estimateurs utilisés classiquement (*i.e.* MLE, MOM, PWM) comportent des problèmes d'existence ou numériques. Par exemple, [Hosking et Wallis \[1987\]](#) établissent que la log-vraisemblance tendra toujours vers l'infini lorsque  $\xi < 1$  et que  $-\sigma/\xi$  tend vers le maximum de l'échantillon. De plus, lorsque  $\xi < -1$ , l'estimateur MLE n'existe plus puisque la vraisemblance n'admet pas de maximum local. Lorsque  $\xi > 0.5$ , l'estimateur ME n'existe pas puisque la variance est alors infinie. De même, lorsque  $\xi > 1$ , les PWM n'existent plus asymptotiquement. De plus avec les PWM et lorsque  $\xi < 0$ , certaines observations peuvent ne plus appartenir au domaine de définition de la loi (*i.e.*  $x \geq -\sigma/\xi$ ) [[Zhang, 2007](#)].

Afin d'éviter ces problèmes d'existence et numériques, [Zhang \[2007\]](#) propose un estimateur nommé Likelihood Moment Estimator (**LME**). Ce nouvel estimateur a l'avantage d'être basé sur une optimisation numérique simple évitant de nombreux problèmes d'instabilité et conservant de bonnes qualités de convergence.

**Tab. 2.1:** Synthèse sur la performance des différents estimateurs de la GPD.

Estimateur	Zone de Performance	Existence	Robuste
<i>MLE</i>	Taille > 500	$\xi \geq -1$	Oui <sup>†</sup>
<i>MOM</i>	$\xi \approx 0$	$\xi \leq 0.5$	Non
<i>PWM</i>	$\xi \approx 0.2$	$\xi \leq 1$	Non
<i>LME</i>	—	$r < 0.5, r \neq 0$	Non
<i>MPLE</i>	—	$-1 \leq \xi \leq 1$	Oui <sup>†</sup>
<i>MDPD</i>	—	$\xi < 1 + \alpha^{-1}, \text{ si } \xi > 0$	Oui
<i>ME</i>	—	$\xi \leq 0.5$	Oui
<i>OBRE</i>	—	—	Oui

<sup>†</sup>Robuste lorsque la contamination n'est pas trop importante.



**Fig. 2.1:** Notion de seuil de perception et des différents types d'information historique. Figure inspirée de Brázdil et al. [2006].

Le Tableau 2.1 regroupe les principales caractéristiques des estimateurs introduits lors de cette section. Néanmoins, pour toutes les études de performance, la qualité des estimateurs est jugée en fonction de la vraie valeur du paramètre de forme  $\xi$ . Bien entendu, cette valeur ne sera jamais connue en pratique ; de sorte qu'il est difficile de faire un réel classement des meilleurs estimateurs pour le praticien. De plus, les estimateurs les plus robustes face à une contamination présenteront souvent une vitesse de convergence plus lente. Dès lors, pour des échantillons de taille restreinte (ce qui est souvent le cas en hydrologie), ces estimateurs sont d'une utilité limitée. Néanmoins, les travaux sur de nouveaux estimateurs issus de l'EVT restent très actifs et méritent intérêt.

### 2.1.2 L'Information Historique

Une autre approche pour consolider l'extrapolation aux plus forts quantiles est de s'intéresser aux crues historiques. Plusieurs auteurs font appel à ce type d'information supplémentaire [Parent et Bernier, 2003; Naulet et al., 2005; Reis Jr. et Stedinger, 2005]. L'information dont l'hydrologue dispose alors est dite *censurée*. En effet, ces données historiques ne sont *visibles* que si elles sont au delà d'un seuil de perception. Par exemple, une crue historique est relatée dans les documents anciens selon l'importance des dégâts, de la présence d'observateurs sur les lieux ou encore de l'importance de l'événement météorologique. Nous voyons donc que cette information diffère d'une chronique de débit classique puisque les valeurs sont présentes uniquement lors de circonstances particulières ; alors que la chronique classique relève les données de façon régulière. La Figure 2.1 illustre les différents types d'information dont l'hydrologue peut disposer.

Toutefois, l'incorporation de données historiques est une tâche très complexe et qui peut laisser perplexe. En effet, les paysages, l'aménagement du territoire ainsi que la végétation ont énormé-

ment évolué durant les derniers siècles, impactant probablement les processus de transformation pluie-débit. Puisque les relevés historiques consistent souvent en des hauteurs d'eaux, la transformation hauteur-débit n'est pas évidente et présente des incertitudes non négligeables. De plus, dans un contexte de fluctuation climatique (n'oublions pas que des données remontant jusqu'à la préhistoire peuvent être considérées), les données d'entrées sont forcément entachées d'incertitudes considérables.

Enfin, l'incorporation dans un modèle statistique rigoureux des différents types d'information (actuelle, historique, préhistorique) n'est pas triviale. Supposons que nous disposons d'une information historique disponible sur  $H$  années avec des observations notées  $y_j$ ,  $j = 1, \dots, r$ . Alors, il existe deux types d'approche conduisant à deux vraisemblances bien différentes [Parent et Bernier, 2003] : l'approche où l'information est censurée par un seuil et celle censurée par le nombre de dépassements.

### Information Censurée par un Seuil

Dans cette approche, il est supposé qu'il existe un seuil de perception  $u^*$  et que les données historiques sont connues puisque ce sont les **seules** à dépasser ce seuil  $u^*$ . Pour ce modèle, la vraisemblance pour les données historiques s'écrit alors :

$$\ell(y; \lambda, \sigma, \xi) = \frac{(\lambda H)^r \exp(-\lambda H)}{r!} \sigma^{-r} \exp \left\{ -\frac{1+\xi}{\xi} \sum_{j=1}^r \log \left( 1 + \xi \frac{y_j - u^*}{\sigma} \right) \right\} \quad (2.1)$$

où  $\lambda = \Pr[Y \leq u^*]$  et  $\sigma$  et  $\xi$  sont les paramètres d'échelle et de forme de la GPD respectivement.

### Information Censurée par le Nombre de Dépassement

Pour cette approche, nous supposons que les données historiques correspondent aux  $r$  **plus fortes** valeurs sur la période  $H$ . La vraisemblance devient alors :

$$\ell(y; \lambda, \sigma, \xi) = (\lambda H)^r \sigma^{-r} \exp \{ -uH [1 - G(y_m, u, \sigma, \xi)] \} \exp \left[ (1 + \xi) \frac{y_j - u}{\sigma} \right] \quad (2.2)$$

où  $u$  est un seuil suffisamment élevé,  $y_m = \min(y_j : j = 1, \dots, r)$ ,  $\sigma$  et  $\xi$  sont les paramètres d'échelle et de forme de la GEV,  $\lambda = \Pr[Y \leq u]$  et  $G$  est la fonction de répartition d'une GEV.

Pour ce type d'approches, les incertitudes sur les quantiles estimés sont disponibles. Toutefois, ces incertitudes sont très sensibles à la qualité ainsi qu'à la méthode d'incorporation des données historiques. Finalement, l'incorporation de données historiques est séduisante mais elle est confrontée à de nombreuses et fortes incertitudes tant au niveau du modèle en lui même que des données recueillies.

## 2.2 Modélisation des Débits à Partir de la Connaissance des Processus

D'une manière générale, les méthodes de ce type veulent rendre compte des connaissances disponibles en termes de processus dans les approches stochastiques pour améliorer la robustesse des ajustements probabilistes sur les débits. Elles s'appuient, pour ce qui est des crues, sur une information pluviométrique souvent plus accessible (couverture spatiale et temporelle plus importante) et plus précises que celle liée aux stations de jaugeage.

Nous pouvons citer les travaux de [Gottschalk et al. \[1997\]](#) qui s'appuient sur la courbe de récession pour consolider les valeurs des débits d'étiage. En crue, des exemples sont proposés par

Rulli et Rosso [2002] et De Michele et Rosso [2002]. Ils font appel à la distribution statistique des pluies et à une transformation pluie-débit simplifiée pour construire/extrapoler la distribution des débits. Dans ce dernier cas, les paramètres de la loi GEV modélisant les débits sont fonction des paramètres de la GEV pour les pluies extrêmes modulés par les capacités de rétention des sols et par le routage des crues.

Pour illustrer nos propos, nous avons retenu trois méthodes particulières : (a) le GRADEX qui est la méthode de référence en France pour le dimensionnement des grands barrages français [Comité Français des Grands Barrages, 1994], (b) les approches de type pluie-débit et notamment le modèle Shypre [Arnaud, 1997] et (c) la Crue Maximale Probable, la méthode de référence anglo-saxonne pour le dimensionnement des évacuateurs de crue des grands barrages [Dumas, 2006].

### 2.2.1 La Méthode du GRADEX

En France, deux méthodes dédiées aux crues et aux hypothèses relativement proches, ont vu le jour : la méthode SPEED - Système Probabiliste d'Etudes par Evénements Discrets - [Cayla, 1993] et la méthode du GRADEX [Guillot et Duband, 1967]. La méthode du GRADEX est actuellement la plus répandue et a pour objectif de définir des crues de projet de période de retour supérieure à 1000 ans.

Le GRADEX développé à la D.T.G. d'EDF permet l'utilisation des données de pluie en complément de celles des débits afin d'affiner l'ajustement aux plus forts quantiles. Ce modèle repose sur trois hypothèses [Paquet *et al.*, 2006] :

#### HYPOTHÈSE 1

*« La quantité de pluie tombant sur le bassin, au pas de temps retenu comme pertinent, voit, asymptotiquement pour les valeurs élevées, ses probabilités de dépassement décroître exponentiellement ; »*

#### HYPOTHÈSE 2

*« Le déficit d'écoulement, différence entre le volume d'eau précipité et le volume d'eau écoulé à l'exutoire du bassin pendant un pas de temps équivalent, n'augmente plus au-delà d'une certaine valeur de pluie ; »*

#### HYPOTHÈSE 3

*« Quelle que soit sa durée de retour, le débit de pointe de crue se déduit du débit moyen (toujours sur un pas de temps de même longueur) grâce à un coefficient multiplicatif moyen caractéristique du bassin, coefficient que l'on peut estimer à partir des hydrogrammes de crues réelles enregistrées sur le bassin. »*

En pratique (*cf.* Figure 2.2),

1. on adopte la distribution des débits jusqu'à une période de retour critique  $T_{\text{crit}}$  ;
2. on prolonge la fonction quantile à l'aide du gradex des pluies de bassin.

Dans un contexte opérationnel, la période de retour critique  $T_{\text{crit}}$  qui est un paramètre du modèle, ne sera jamais estimée mais imposée, après expertise de l'hydrologue, à une valeur comprise entre 10 et 50 ans selon la réactivité du bassin d'étude.

D'après Margoum *et al.* [1995], le modèle du GRADEX n'est pas performant pour le domaine des crues rares (*i.e.*  $10 < T < 1000$  ans) puisqu'ils soulignent une :

*« [...] surestimation des débits, introduite par la cassure brutale dans la distribution des débits au seuil d'extrapolation  $T_{\text{crit}}$ . »* même si elle *« [...] devient faible en valeur relative pour les grandes périodes de retour, et surtout n'est pas dommageable lorsqu'il s'agit d'études liées à la sécurité des ouvrages. »*

Margoum *et al.* [1995] ont donc amélioré la méthode du GRADEX en intégrant un modèle paramétrique autorisant un raccord progressif entre la distribution des débits et celle des pluies ; permettant ainsi une estimation plus fiable des crues rares.

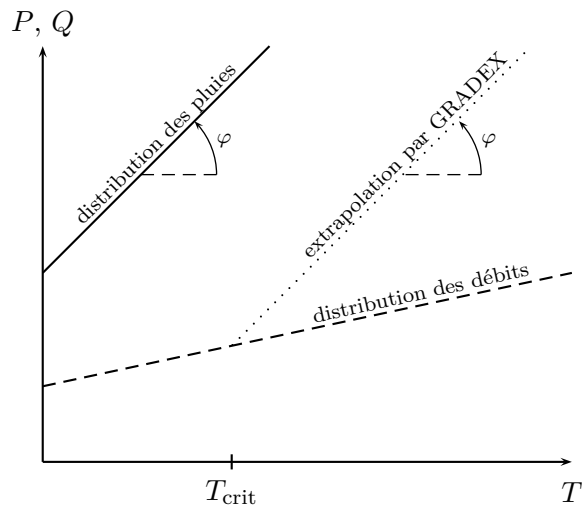


Fig. 2.2: Illustration de la méthode du GRADEX.

Bien que le GRADEX repose sur des hypothèses hydrologiques raisonnables, il reste un modèle discuté [Reed, 1994] avec des applications bien souvent cantonnées à la France. Quelques applications par des hydrologues non français ont tout de même été effectuées - *e.g.* en Autriche [Merz *et al.*, 1999], Colombie [Mosquera-Machado et Ahmad, 2007] et USA [Naghattini *et al.*, 1996]. Toutefois, comme le remarque [Reed, 1994], il est difficile d'apprécier les performances du GRADEX puisque ce modèle s'attache à l'estimation des crues extrêmes ( $T \geq 1000$  ans) pour lesquelles il est difficile, voire impossible, de prouver les bonnes/mauvaises prédictions. Bien que le GRADEX constitue une méthode robuste, simple d'utilisation et physiquement fondé, son emploi n'est pas sans danger. Ainsi, trois remarques peuvent être émises sur la structure même du modèle.

### Attachement à la Décroissance Exponentielle

Le GRADEX suppose que la distribution des pluies et des débits est à « décroissance exponentielle simple »- au moins dans sa branche asymptotique. En particulier, le choix des lois de Gumbel/Exponentielle néglige considérablement les incertitudes sur le paramètre de forme  $\xi$  puisque, pour ces lois,  $\xi$  est supposé parfaitement connu et nul. Cette « décroissance exponentielle » est problématique puisque nombre d'auteurs soulignent son incapacité à modéliser les plus forts extrêmes [Muller, 2006; Coles et Pericchi, 2003; Koutsoyiannis *et al.*, 1998].

L'amélioration du GRADEX proposé par Margoum *et al.* [1995] n'échappe pas non plus à cette décroissance exponentielle. L'extension proposée par Margoum *et al.* permet simplement d'éviter une cassure non physique mais ne change rien au comportement asymptotique des débits. En effet, par construction du modèle, asymptotiquement la distribution des débits converge vers la distribution (supposée exponentielle) des pluies. Notamment, la distribution des débits obtenue par le modèle de Margoum *et al.* sera toujours bornée supérieurement par issue de la méthode du GRADEX initiale.

Paquet *et al.* [2006] montrent sur 411 longues séries françaises de précipitations journalières que, pour des modèles à décroissance exponentielle,

*« la conclusion ne fait pas de doute : il y a beaucoup plus de valeurs étiquetées centennales ou millennales que ce qu'on devrait trouver en moyenne. »*

Pour autant, Paquet *et al.* [2006] ne rejettent pas le modèle exponentiel. L'hypothèse avancée par les auteurs est que « les horsains pourraient eux-mêmes suivre une loi exponentielle. », mais différente, conservant ainsi la première hypothèse du modèle du GRADEX. Toutefois, il paraît décevant de ne pas mentionner la souplesse des lois GEV/GPD pour contourner la faiblesse des modèles à « décroissance exponentielle simple » et de se tourner vers des hypothèses plus complexes

comme la mixture de lois. Enfin, une réelle étude de performance des approches de types GRADEX et de celles en provenance de l'EVT serait nécessaire. Cette étude doit d'ailleurs être réalisée dans le cadre d'une thèse au sein d'EDF en 2008.

### Notion de Vraisemblance et de Parcimonie

La force du GRADEX réside dans sa simplicité et sa construction géométrique de la distribution des débits. Toutefois, signalons que le modèle perd, au profit de la simplicité, toute notion de vraisemblance. Pourtant cette vraisemblance existe et la densité de probabilité du GRADEX s'écrit sous la forme suivante :

$$f(y) = \begin{cases} \sigma_1^{-1} \exp\left(\frac{u_1 - y}{\sigma_1}\right), & y \leq u_1 + \sigma_1 \log(\rho T_{\text{crit}}) \\ \sigma_2^{-1} \exp\left(\frac{u_2 - y}{\sigma_2}\right), & \text{sinon} \end{cases} \quad (2.3)$$

où  $u_1, \sigma_1$  sont les paramètres de la loi Exponentielle estimés à partir des débits observés,  $\rho$  le nombre moyen de dépassements du seuil  $u_1$  par an,  $\sigma_2$  le paramètre d'échelle de la loi Exponentielle relatives aux pluies (généralement appelé *gradex des pluies*) et  $u_2$  un paramètre de position permettant un prolongement adéquat entre la distribution des débits observés et celle des pluies. En particulier, afin de conserver une distribution de probabilité,  $u_2$  doit être défini selon la contrainte de continuité suivante :

$$1 - \exp\left[\frac{u_2 - u_1 - \sigma_1 \log(\rho T_{\text{crit}})}{\sigma_2}\right] = 1 - \frac{1}{\rho T_{\text{crit}}} \quad (2.4)$$

En conséquence,  $u_2$  est donné par

$$u_2 = u_1 + (\sigma_1 - \sigma_2) \log(\rho T_{\text{crit}}) \quad (2.5)$$

Ainsi,

$$\begin{aligned} \int_{\mathbb{R}} f(y) dy &= \int_{u_1}^{u_1 + \sigma_1 \log(\rho T_{\text{crit}})} \sigma_1^{-1} \exp\left(\frac{u_1 - y}{\sigma_1}\right) dy + \int_{u_1 + \sigma_1 \log(\rho T_{\text{crit}})}^{+\infty} \sigma_2^{-1} \exp\left(\frac{u_2 - y}{\sigma_2}\right) dy \\ &= 1 - \exp\left[\frac{u_1 - u_1 - \sigma_1 \log(\rho T_{\text{crit}})}{\sigma_1}\right] + \exp\left[\frac{u_2 - u_1 - \sigma_1 \log(\rho T_{\text{crit}})}{\sigma_2}\right] \\ &= 1 - \frac{1}{\rho T_{\text{crit}}} + \exp\left[\frac{u_1 + (\sigma_1 - \sigma_2) \log(\rho T_{\text{crit}}) - u_1 - \sigma_1 \log(\rho T_{\text{crit}})}{\sigma_2}\right] \\ &= 1 - \frac{1}{\rho T_{\text{crit}}} + \frac{1}{\rho T_{\text{crit}}} \\ &= 1 \end{aligned}$$

On voit donc que  $f$  est bien une densité de probabilité dès lors que  $u_2$  satisfait la relation (2.5). En revanche, si cette contrainte n'est pas respectée, la notion de probabilité n'existe plus.

Par ces calculs, nous ne souhaitons pas remettre en cause la validité théorique du GRADEX mais montrer qu'il est possible d'ajuster ce modèle par des techniques statistiques conventionnelles (maximum de vraisemblance notamment) et ainsi d'obtenir les incertitudes, jusqu'alors inaccessibles, sur les estimations. Soulignons également que si la contrainte (2.4) n'est qu'approximativement respectée, cela pourrait avoir des conséquences non négligeables sur l'estimation d'événements ayant des fréquences au dépassement de l'ordre de  $10^{-3}$  ou  $10^{-4}$ .

Enfin, à partir de la vraisemblance du GRADEX (cf. Équation(2.3)), nous voyons que ce modèle dispose de quatre paramètres à ajuster :  $u_1, \sigma_1, \sigma_2$  et  $T_{\text{crit}}$ . Notamment en terme de parcimonie, ce modèle sera donc plus pénalisé que les modélisations par les lois GEV/GPD.

### Quelques Choix Difficiles

Le GRADEX comporte également quelques points difficiles à mettre en oeuvre. En particulier, pour le calcul du débit de pointe, un choix pertinent pour le pas de temps  $d$  définissant le débit moyen n'est pas aisé puisqu'une durée trop grande conduit à des erreurs d'estimation importantes en sortie de modèle (puisque le coefficient multiplicatif de l'hypothèse 3 est alors très dispersé); alors qu'une durée trop petite n'aurait plus de sens physique.

De même, la détermination du point de rupture  $T_{\text{crit}}$  est également un choix particulièrement difficile et fortement dépendant du site d'étude.

Le SCHADEX [Paquet *et al.*, 2006] présente la toute dernière évolution des approches de type GRADEX et conduit à l'estimation de plusieurs paramètres d'échelle de pluie d'une même saison selon le type de circulation atmosphérique considéré. Le SCHADEX fait intervenir le modèle hydrologique MORDOR [Garçon, 1996] afin de simuler la réponse du bassin à des pluies de tous ordres. Toutefois, ni le GRADEX, ni aucune de ses extensions, n'est ajusté comme un modèle statistique classique - par exemple par maximum de vraisemblance, le point de rupture  $T_{\text{crit}}$  est un paramètre à estimer. Dès lors les incertitudes sur les quantiles prédits ne sont pas disponibles bien que d'importance réelle.

### 2.2.2 Les Modèles Pluie-Débit

Les approches pluie-débit cherchent à modéliser numériquement l'évolution des débits à partir d'une chronique de pluie alimentant le bassin. Ces approches se distinguent donc de celles présentées plus haut puisqu'elles concentrent l'effort de modélisation sur les processus générateurs des débits et non les débits eux-mêmes.

Les premières approches pluie-débit étaient de type *design storm i.e.* les débits sont estimés à partir d'une transformation pluie-débit pour un événement pluvieux synthétique précis : la *design storm*. La période de retour du débit de pointe obtenu est alors supposée égale à celle de l'événement pluvieux - bien que cette hypothèse reste très critiquée [Larson et Reich, 1972; Dickinson *et al.*, 1992].

Une alternative à l'approche par *design storm*, introduite par Eagleson [1972] et désormais bien établie [Rahman *et al.*, 2002; Loukas, 2002], consiste à obtenir une simulation continue des débits à partir de chroniques de pluie (stochastiques ou enregistrées), puis de faire une analyse fréquentielle sur cette série temporelle simulée de débits. La Figure 2.3 représente de manière synthétique les différentes étapes d'estimation des quantiles de crue par les approches de type pluie-débit. La première étape (partie gauche de la figure) consiste à caler les paramètres du simulateur de pluies et ceux du modèle pluie-débit à partir des mesures de précipitations et de débits disponible. Lors de la deuxième étape (partie droite de la figure), un grand nombre de scénarios pluvieux sont simulés et les débits associés sont déduits du modèle hydrologique. La chronique de débits simulés résultante est alors exploitée afin de consolider l'estimation de la distribution des débits les plus extrêmes.

Par construction des approches pluie-débit, l'estimation des quantiles de crue est fortement dépendante de la qualité prédictive des modèles employés. Plusieurs auteurs mentionnent la difficulté de représentativité des modèles hydrologiques ou de simulation stochastique des pluies. Ainsi, Moughamian *et al.* [1987] et Cadavid *et al.* [1991] soulignent la difficulté d'estimation des paramètres du modèle de générations des pluies. Ils constatent également qu'une analyse statistique des valeurs extrêmes s'est révélée plus efficace sur leurs cas d'étude. Plus récemment, Goel *et al.* [2000] ont analysé cinq bassins versants (quatre en Inde et un aux USA). Ils ont montré que les mêmes difficultés soulignées par Moughamian *et al.* et Cadavid *et al.* étaient toujours d'actualité puisque, pour leur étude, les débits de pointe ont toujours été sous-estimés - notamment pour les plus fortes périodes de retour.

Les modèles hydrologiques modélisant la réponse du bassin versant ne sont pas non plus exempt de problèmes. Beven [2001] dresse un très instructif état des lieux des modèles hydrologiques. Il



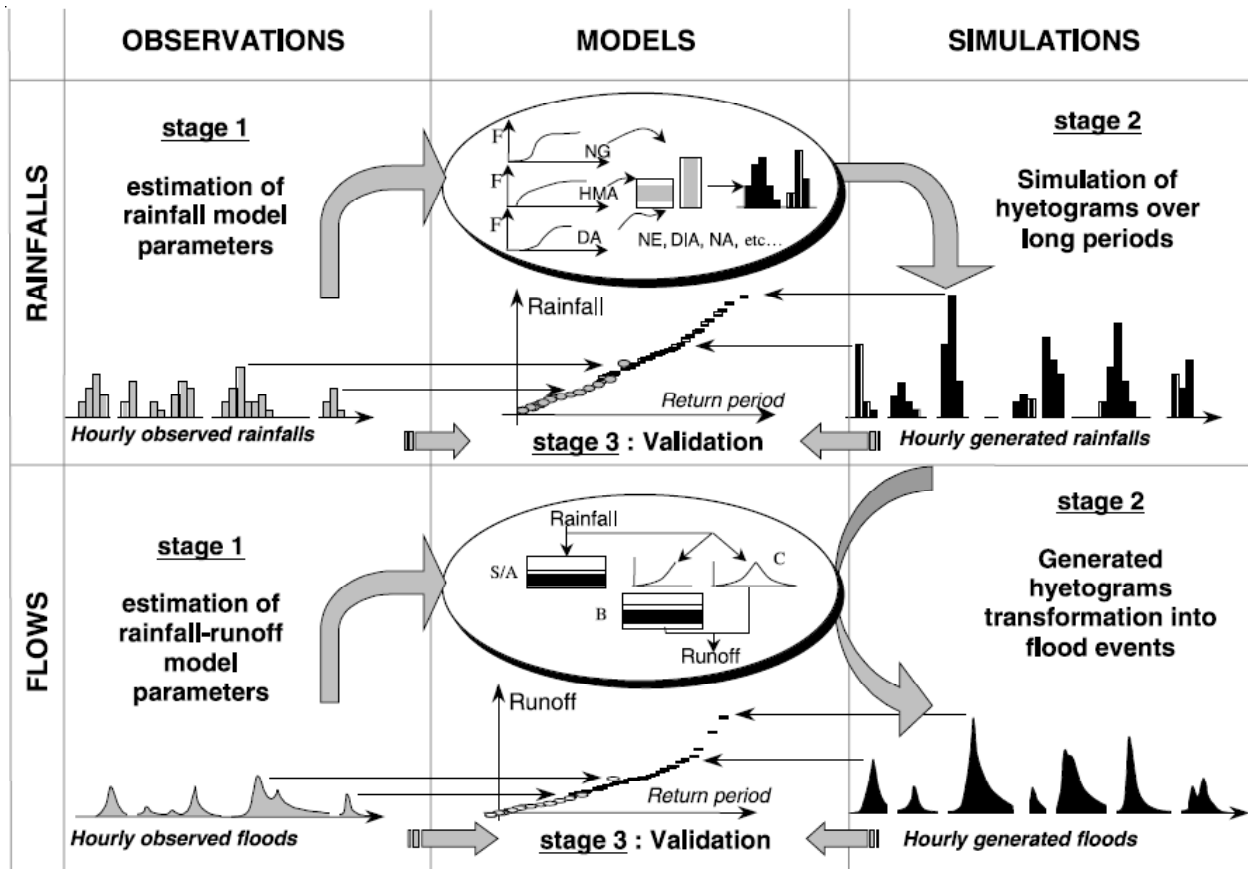


Fig. 2.3: Schéma synthétique représentant la procédure d'estimation de la distribution des débits à l'aide de transformations pluie-débit. Figure tirée de [Arnaud et Lavabre \[2002\]](#).

aborde notamment les problèmes : (a) de non linéarité existant entre les pluies et les débits, (b) d'échelles pour lesquelles sont disponibles les données nécessaires à la modélisation, (c) d'unicité *i.e.* il n'existera jamais un modèle conceptuel universel adapté à tous les bassins versants, (d) d'*équi-finalité* *i.e.* il n'existe pas un jeu optimal de paramètres mais tout un ensemble et (e) de la connaissance des incertitudes engendrées par le modèle.

Parmi les faiblesses mentionnées par [Beven](#), la plupart concernent un grand nombre de modèles. En revanche, le problème des incertitudes est spécifique à ce type d'approches. En effet, que ce soit pour la génération stochastique de pluies ou pour la réponse du bassin versant, l'estimation des quantiles de crue est fortement dépendante des modèles employés. Pourtant la qualité de ces modèles est difficilement accessible et le praticien sera souvent contraint à caler au mieux et de manière subjective les paramètres du modèle suivant les données disponibles [[Raines et Valdes, 1993](#)].

Finalement, le praticien dispose donc de débits simulés dont il ne peut parfaitement connaître le biais et les incertitudes inhérents à l'utilisation de ces modèles. Enfin, il est illusoire de croire qu'il suffirait d'obtenir une chronique simulée de débits de longueur très importante pour avoir des estimations qui seraient apparemment plus précises.

### 2.2.3 La Crue Maximale Probable

Cette approche se distingue des précédentes puisqu'au lieu d'obtenir la distribution des débits de pointe et les périodes de retour associées, ce modèle apporte une unique valeur correspondant au débit de la **Crue Maximale Probable (CMP)** selon la terminologie employée. Ces valeurs



## 2. Analyse Fréquentielle Locale

**Tab. 2.2:** Essai de détermination de la période de retour de la CMP.  $R$  : rapport entre la CMP et  $Q_{10000}$  estimé par le GRADEX.  $T_{\text{CMP}}$  : estimation de la période de retour de la CMP par le GRADEX. Tableau tiré du [Comité Français des Grands Barrages \[1994\]](#).

Barrage	Pays	$R$	$T_{\text{CMP}}$ (an)
Serre de la Fare 1	France	1.22	26000
Péligre (mini)	Haïti	1.08	29000
La Touche Poupard	France	1.35	37000
Péligre (moyen)	Haïti	1.24	47000
Khalilkan	Irak	1.25	60000
Bakurman	Irak	1.26	65000
Péligre (maxi)	Haïti	1.37	68000
Mazeau	France	1.51	86000
Charpal	France	1.69	140000
Serre de la Fare 2	France	1.84	250000
Serre de la Fare 3	France	2.11	550000
Sir	Turquie	1.91	640000
Trézence	France	2.37	750000
Roppe	France	2.03	820000
Bairon	France	2.27	1040000
Veurdre	France	2.64	1250000
Panaltdalan	Philippines	2.15	1400000

sont utiles pour définir la crue de conception de grands ouvrages hydrauliques et/ou dont la rupture peut causer des pertes économiques importantes ou même des pertes humaines [[Dumas, 2006](#)].

La CMP est souvent considérée comme un modèle très pénalisant. En effet, ce modèle conceptuel estime la CMP en superposant des événements pluvieux extrêmes sur un bassin déjà saturé ; les valeurs de débits étant obtenues à partir d'un modèle de transformation pluie-débit. Dans cette logique, le [Committee on Safety Criteria for Dams \[1985\]](#) définit la CMP (ou PMF en anglais) selon la terminologie suivante :

*« The PMF is the flood that may be expected from the most severe combination of critical meteorological and hydrologic conditions that are reasonably possible in the region. »*

[Lave et Balvanoyos \[1998\]](#) soulignent que la CMP est par définition subjective puisque les définitions de « plus sévère combinaison d'événements météorologiques et hydrologiques extrêmes au sein d'une région » et « raisonnablement possibles » différeront bien souvent d'un expert à l'autre.

La CMP est souvent présentée comme une crue dont la probabilité de dépassement est théoriquement nulle. Toutefois, la notion de *probabilité de dépassement nulle* est un abus de langage puisque ce modèle n'est aucunement probabiliste. Cette terminologie vient du fait qu'il est quasiment impossible qu'un événement pluvieux catastrophique survienne sur un bassin déjà entièrement saturé. En conséquence, il est impossible d'associer une période de retour à cette CMP - ni théoriquement puisque les probabilités n'existent pas, ni conceptuellement puisque  $T = \infty$ .

Ce type d'approche n'est pas exempt d'incertitudes présentes lors : (a) de l'analyse de pluie ; (b) de l'expertise météorologique et (c) de l'estimation des paramètres du modèle de transformation pluie-débit. [Dumas \[2006\]](#) affirme que le dernier est générateur des plus fortes incertitudes :

*« [...] surtout en ce qui concerne l'estimation du taux de ruissellement en fonction du degré d'humidité du sol puisque le rendement marginal des précipitations lorsqu'on passe des crues courantes aux crues les plus rares, comme c'est le cas lors de la simulation des CMP, n'a pu être ajusté et demeure théorique. »*

Enfin, le [Comité Français des Grands Barrages \[1994\]](#), dans un état de l'art des méthodes de

type GRADEX, estime via le GRADEX les périodes de retour associées aux CMP. Ces périodes de retour estimées sont reportées au sein du Tableau 2.2. Selon les auteurs,

« *De tels résultats ne doivent pas surprendre.[...] Il ne s'agit pas alors d'un constat de désaccord entre les deux méthodes [...]* »

Toutefois, sur des estimations aussi incertaines, force est de constater que nous ne pouvons juger à partir de ce retour d'expérience de la pertinence des modèles : ni pour la CMP, ni pour le GRADEX.

## 2.3 Conclusions et Objectifs à Atteindre

Dans ce chapitre, nous avons rappelé quelques approches existantes pour l'analyse fréquentielle des débits. Ces approches sont essentiellement des approches statistiques et intègrent parfois des connaissances hydrologiques. Notre travail se situe également dans cette perspective. Nous cherchons à intégrer les notions physiques dans les modèles statistiques en prenant soin de ne pas *détruire* la rigueur statistique du modèle. De plus, nous viserons à intégrer de l'information essentielle en laquelle nous avons suffisamment confiance. Et si toutefois cette information supplémentaire était incertaine, il ne faudrait aucunement lui affecter le même poids que celle jugée plus fiable. De même, il serait très pertinent d'avoir un modèle statistique permettant l'estimation probabiliste de la dynamique de crue des bassins.

Ainsi, au vu de ce bref état de l'art sur l'estimation locale des débits caractéristiques de crue, nous proposons de définir un nouveau modèle permettant d'utiliser au mieux l'information *fiable* disponible au site d'étude en réduisant au maximum la perte d'information inhérente à l'analyse fréquentielle des seuls extrêmes. Afin d'obtenir non seulement les quantiles de crue mais aussi les incertitudes associées, nous nous devons de proposer un modèle statistique rigoureux. Ce modèle sera introduit lors du Chapitre 7.

La connaissance des processus régissant le site d'étude sera également utilisée. En ce sens, nous nous placerons dans un cadre d'estimation régionale des débits caractéristiques. Le chapitre suivant se veut d'ailleurs une introduction à l'analyse fréquentielle régionale.



# 3 Le Modèle de l'Indice de Crue

DANS ce chapitre, nous allons présenter des estimateurs dits **régionaux**, c'est à dire qu'ils feront appel à des données en provenance d'autres sites que celui étudié. L'objectif est, tout comme pour le chapitre précédent, d'obtenir des estimations de quantiles de crue plus fiables. Ainsi, les modèles régionaux partagent un point commun avec certains des modèles évoqués précédemment (GRADEX, information historique) : celui d'intégrer de l'information supplémentaire pour rendre les estimations plus performantes.

## 3.1 Les Étapes de l'Analyse Fréquentielle Régionale

Le premier estimateur régional fut proposé par Dalrymple [1960] sous le nom d'**indice de crue** et reste encore le modèle le plus utilisé pour l'estimation régionale des débits. Sa simplicité de mise en oeuvre et sa robustesse en font un modèle incontournable et appliqué dans de nombreuses régions du globe. Parmi toutes les applications, nous pouvons citer celles de Kjeldsen et Jones [2007] pour le Royaume-Uni, Merz et Blöschl [2005] pour l'Autriche, Kjeldsen *et al.* [2002] pour l'Afrique du Sud, Madsen et Rosbjerg [1997] pour la Nouvelle-Zélande et Javelle *et al.* [2002] au Canada.

L'idée du modèle de l'indice de crue, basée sur des propriétés d'invariance d'échelle des distributions des débits, est simple :

*« Au sein d'une région dite homogène, toutes les distributions sont, à un facteur d'échelle près, identiques. Ce facteur d'échelle appelé indice de crue étant dépendant du site considéré. L'unique distribution sous-jacente de la région homogène est appelée loi régionale. »*

Trois étapes distinctes permettent d'obtenir une estimation régionale des quantiles de crue :

1. La construction d'une région homogène : c'est l'étape de « régionalisation » ;
2. La validation de la pertinence de la région : c'est l'étape de « test d'homogénéité » ;
3. L'estimation de la loi régionale puis de la distribution des débits au site cible : c'est l'étape « d'estimation ».

Le modèle de l'indice de crue n'est pas le seul estimateur régional proposé dans la littérature [Coles et Tawn, 1996; Northrop, 2004] ; toutefois étant le plus utilisé, il sera le point de départ de nos développements.

### 3.1.1 La Régionalisation

Le but de cette étape consiste à obtenir un ensemble de stations, la région homogène, ayant un comportement statistique similaire. En parcourant la littérature scientifique sur le sujet, nous avons pu identifier les approches suivantes :

#### Critères Géographiques

Les sites sont regroupés selon la distance euclidienne qui les séparent, ou plus simplement par leur appartenance à une même région administrative [Beable et McKerchar, 1981]. Une analyse plus fine des liaisons entre stations montre sans ambiguïté que la proximité géographique n'induit pas toujours une proximité hydrologique. Cette approche est désormais fortement déconseillée car elle n'assure en rien l'homogénéité entre les stations [Robson et Reed, 1999; Ouarda *et al.*, 2001].

#### Considérations Hydrométéorologiques

Pour cette approche, la construction du groupe de stations réputées similaires est effectuée selon différents critères hydrométéorologiques. Ainsi, Schaefer [1990] a regroupé les sites en considérant la pluie annuelle moyenne. Gingras et Adamowski [1993] ont construit les régions homogènes en examinant les occurrences des plus forts événements de chaque année. De Michele et Rosso [2002] ont proposé une démarche similaire sur la base d'un indice de saisonnalité du régime hydrologique. Toutefois, Cunderlik *et al.* [2004] soulignent que l'identification de la saisonnalité est une tâche subjective, laborieuse et parfois non concluante. Ils proposent donc une nouvelle méthode entièrement objective pour l'obtention d'un indice de saisonnalité en comparant les occurrences observées avec celles dérivées d'un modèle non saisonnier - *i.e.* dont les occurrences suivent une loi uniforme sur l'année.

#### Partition Statistique

A partir d'un échantillon étendu de bassins, les stations sont regroupées selon un algorithme de partition statistique. A chaque itération de l'algorithme de partition, les sites sont placés récursivement dans l'un des deux ensembles proposés selon la valeur d'une statistique déterminée. Cette statistique peut être le rapport des vraisemblances [Wiltshire, 1985], une mesure de la dispersion du coefficient de variation [Wiltshire, 1986] ou des  $L-CV$  et  $L-Skewness$  [Pearson, 1991]. Bien que facile d'utilisation, cette approche a deux inconvénients. Le premier réside dans la procédure dichotomique de partitionnement. En effet, un choix optimal à chaque étape n'assure en rien une partition finale optimale. De plus, les statistiques considérées peuvent poser problème lorsqu'il existe une dépendance entre plusieurs sites de la région.

#### Classification Statistique

De la même manière que pour le partitionnement statistique, on peut faire appel aux techniques usuelles de classification. Il s'agit de collecter les caractéristiques descriptives pour chaque site et de regrouper les sites en fonction de leur similarité vis à vis de ces caractéristiques. Ces données peuvent être des descripteurs physiographiques du bassin versant (altitude, nature des sols, ...), statistiques du site (moyenne, médiane) ou même les deux. Kjeldsen *et al.* [2002] ont regroupé les sites selon la latitude, la longitude, la surface du bassin versant, la pluie annuelle moyenne, la pente moyenne du bassin versant, l'altitude. Onibon *et al.* [2004] utilisent une classification hiérarchique sur les pluies annuelles moyennes. Robson et Reed [1999] utilisent les variables suivantes : la superficie du bassin versant, la pluie interannuelle, un indice de débit de base (Base Flow Index (**BFI**) [Tallaksen et Van Lanen, 2004]) quantifiant la contribution souterraine dans l'écoulement total. Les régions homogènes sont alors construites en minimisant une distance particulière dans un espace dont les axes sont les variables explicatives.

#### Analyse Factorielle

Ces méthodes sont basées sur des techniques géométriques permettant de réduire la dimension du problème tout en dégradant le moins possible l'information initiale. Pour les études qui nous

concernent, la dimension du problème sera donnée par le nombre de variables physiographiques et météorologiques dont le praticien dispose pour son étude. Ainsi, il s'agira de construire un espace de dimension inférieure dont les bases sont des combinaisons linéaires de ces variables. En hydrologie, les techniques factorielles les plus utilisées sont l'analyse en composantes principales [Chokmani et Ouarda, 2004] et l'analyse canonique des corrélations [Ribeiro-Correa *et al.*, 1995; Ouarda *et al.*, 2001]. L'analyse canonique des corrélations généralise l'analyse en composante principale puisqu'elle permet de décrire la relation de dépendance entre deux ensembles de variables. En particulier, elle permet de faire le lien direct entre les variables hydrométriques et physiographiques/climatiques.

### Synthèse

Il n'existe pas de méthodes universelles pour établir ces régions homogènes. Chaque méthode présente avantages et inconvénients. Il s'agit donc pour le praticien de vérifier les limites et la pertinence de chacune d'entre elles compte tenu des données mises à disposition. Enfin, ce sera bien souvent le test d'homogénéité qui sera le plus discriminant quant au choix de la méthodologie employée. La lecture des différentes références montre également que le choix des variables explicatives n'est pas automatique et évolue d'une région à l'autre. De ce fait, nous nous gardons bien de privilégier une méthode précise ainsi que certaines variables physiographiques, météorologiques et hydrologiques pour constituer une région homogène.

### 3.1.2 L'Homogénéité en Question

Supposons un instant que nous disposions d'une région homogène. Il paraît essentiel de tester la pertinence de cette région. En effet, dans la majorité des cas, les groupes de stations réputées homogènes auront été établis à l'aide de techniques purement objectives et ne faisant pas appel à la connaissance des processus hydrologiques générant la distribution des débits. Comme l'estimation des débits sera dépendante de la région établie, cette étape de vérification de l'homogénéité est primordiale.

Pourtant, à ce jour, il n'existe que très peu de tests évaluant le degré d'homogénéité d'une région. Nous pouvons distinguer deux grandes familles pour ces tests ; les tests basés sur : (a) les valeurs de débits et (b) un indice de saisonnalité. Pour la première famille, nous pouvons citer la statistique d'Hosking et Wallis [1993], le test de Wiltshire [1986] ou encore des tests basés sur les statistiques d'Anderson-Darling et de Durbin et Knott. Pour les tests d'homogénéité basés sur un indice de saisonnalité, nous pouvons citer les travaux de [Cunderlik et Burn, 2006a,b; Ouarda *et al.*, 2006; Bayliss et Jones, 1993] utilisant les statistiques circulaires afin d'identifier la saisonnalité des crues.

Viglione *et al.* [2007] ont récemment réalisé une étude de performance pour quelques uns de ces tests. Bien que leurs cas d'étude s'éloignent quelque peu du cadre opérationnel (*i.e.* données simulées, sites mutuellement indépendants, loi des débits extrêmes parfaitement connue), ils montrent que le test d'Hosking et Wallis [1993] et une version bootstrap de la statistique d'Anderson-Darling conduisaient à des détections plus fines des régions hétérogènes. Toutefois, le test d'Hosking et Wallis étant de loin le plus utilisé des deux et à ce jour non remis en cause, nous ne présenterons dans la suite que ce dernier. Enfin, Chebana et Ouarda [2007] ont généralisé le test d'Hosking et Wallis au cadre multivarié offrant de nouvelles perspectives, notamment pour la modélisation régionale jointe des débits de pointe et de leurs volumes par exemple.

L'idée du test d'Hosking et Wallis est relativement simple et basée sur des simulations numériques. Ces simulations permettent de quantifier la variabilité d'échantillonnage d'une région qui serait effectivement issue d'une unique loi régionale ; et ainsi de la comparer à la variabilité observée sur notre région d'étude. Le principe de ce test est décrit par l'Algorithme 1.

Il paraît nécessaire de faire quelques remarques sur ce test. Premièrement, chaque débit extrême de la région sera simulé à partir d'une loi Kappa à 4 paramètres (Algorithme 1, étape 9) dont la

### 3. Le Modèle de l'Indice de Crue

---

#### Algorithme 1 : Test d'Homogénéité [Hosking et Wallis, 1997]

---

**Entrées** : Un groupe de stations supposées homogènes  
**Sorties** : Une statistique  $H$  caractérisant le degré d'homogénéité  
 /\*  $n_{\text{sites}}$  : Nombre de sites dans la région d'étude \*/  
 /\*  $ech(i)$  : Echantillon pour le site  $i$  \*/  
 /\*  $C_i$  : Indice de crue pour le site  $i$  \*/  
 /\*  $T(\cdot)$  : Une statistique définie par les équations (3.2a) à (3.2c) \*/  
 /\*  $n_{\text{sim}}$  : Nombre de simulation de régions stochastiques \*/  
 /\*  $w_i$  : Taille de l'échantillon au site  $i$  \*/  
**1 pour**  $i = 1$  à  $n_{\text{sites}}$  **faire**  
**2** |  $ech(i) \leftarrow ech(i)/C_i$  ;  
**3** | Calculer les 4 premiers  $L$ -moments  $\ell_1^{(i)}, \tau^{(i)}, \tau_3^{(i)}$  et  $\tau_4^{(i)}$  pour  $ech(i)$  ;  
**4 fin**  
**5** Calculer les  $L$ -moments régionaux :

$$\ell_1^R \leftarrow \frac{\sum_i w_i \ell_1^{(i)}}{\sum_i w_i}, \quad \tau^R \leftarrow \frac{\sum_i w_i \tau^{(i)}}{\sum_i w_i}, \quad \tau_3^R \leftarrow \frac{\sum_i w_i \tau_3^{(i)}}{\sum_i w_i}, \quad \tau_4^R \leftarrow \frac{\sum_i w_i \tau_4^{(i)}}{\sum_i w_i}$$

**6** Calculer  $T(\text{region})$  pour la région d'étude ;  
**7** Ajuster une loi Kappa à 4 paramètres à partir de  $\ell_1^R, \tau^R, \tau_3^R$  et  $\tau_4^R$  ;  
**8 pour**  $j = 1$  à  $n_{\text{sim}}$  **faire**  
**9** | Simuler une région stochastique selon la loi Kappa ;  
**10** | Calculer  $T(j)$  pour la région simulée ;  
**11 fin**  
**12** Calculer

$$H \leftarrow \frac{T(\text{region}) - m}{\sigma}, \quad \text{où } m = \frac{1}{n_{\text{sim}}} \sum_{j=1}^{n_{\text{sim}}} T(j) \quad \text{et} \quad \sigma = \sqrt{\frac{1}{n_{\text{sim}} - 1} \sum_{j=1}^{n_{\text{sim}}} (T(j) - m)^2}$$

**13** Conclure quant à l'homogénéité de la région d'étude selon la valeur de  $H$ .

---

fonction de répartition :

$$F(y) = \begin{cases} \left[ 1 + \kappa \left( 1 + \xi \frac{y-u}{\sigma} \right)^{-1/\xi} \right]^{-1/\kappa}, & \xi \neq 0, \kappa \neq 0 \\ \exp \left[ - \left( 1 + \xi \frac{y-u}{\sigma} \right)^{-1/\xi} \right], & \xi \neq 0, \kappa = 0 \\ \left[ 1 + \kappa \exp \left( - \frac{y-u}{\sigma} \right) \right]^{-1/\kappa}, & \xi = 0, \kappa \neq 0 \\ \exp \left[ - \exp \left( - \frac{y-u}{\sigma} \right) \right], & \xi = 0, \kappa = 0 \end{cases} \quad (3.1)$$

Le choix de cette loi Kappa n'est pas illogique puisque la GEV (e.g.,  $\kappa = 0$ ), la GPD (e.g.,  $\kappa = -1$ ) ainsi que la loi Logistique Généralisée (e.g.,  $\kappa = 1$ ) en sont des cas particuliers. Deuxièmement, ce test nécessite l'utilisation d'une statistique  $T$  bien définie - Algorithme 1, étapes 6 et 10. Pour ce faire, Hosking et Wallis [1997] ont proposé trois statistiques particulières :

$$T_1 = \sqrt{\frac{\sum_{i=1}^N n_i (\tau^{(i)} - \tau^R)^2}{\sum_{i=1}^N n_i}} \quad (3.2a)$$

$$T_2 = \sqrt{\frac{\sum_{i=1}^N n_i \sqrt{(\tau^{(i)} - \tau^R)^2 + (\tau_3^{(i)} - \tau_3^R)^2}}{\sum_{i=1}^N n_i}} \quad (3.2b)$$

$$T_3 = \sqrt{\frac{\sum_{i=1}^N n_i \sqrt{(\tau_4^{(i)} - \tau_4^R)^2 + (\tau_3^{(i)} - \tau_3^R)^2}}{\sum_{i=1}^N n_i}} \quad (3.2c)$$

où  $\tau^{(i)}$ ,  $\tau_3^{(i)}$  et  $\tau_4^{(i)}$  sont respectivement les 2<sup>e</sup>, 3<sup>e</sup> et 4<sup>e</sup> rapports des  $L$ -moments pour le site  $i$ ,  $\tau^R$ ,  $\tau_3^R$  et  $\tau_4^R$  correspondent aux moyennes régionales des 2<sup>e</sup>, 3<sup>e</sup> et 4<sup>e</sup> rapports des  $L$ -moments pondérés par la taille de l'échantillon.

Finalement, Hosking et Wallis [1997] montrent que la statistique  $T_1$  a un plus fort pouvoir discriminant que les statistiques  $T_2$  et  $T_3$ . Ils recommandent donc son utilisation et définissent la classification suivante en fonction de la valeur de  $H$  :

- $H \leq 0$  : certains sites semblent être fortement corrélés
- $0 < H < 1$  : la région peut être considérée comme probablement homogène
- $1 \leq H < 2$  : la région est probablement hétérogène
- $H \geq 2$  : la région est certainement hétérogène

### 3.1.3 L'Estimation au Site Cible

Une fois le regroupement de bassins réputés homogènes effectué et testé, il faut essayer d'utiliser au mieux l'information disponible sur ces bassins, afin d'en déduire la distribution des quantiles de crue au site d'étude. Dans le modèle introduit par Dalrymple [1960], il est supposé que toutes les fonctions de répartition inverses de chaque site de la région homogène sont identiques à une constante multiplicative près. Ainsi, les fonctions de répartition inverses des différents sites une fois normalisées se confondent toutes en une unique fonction de répartition inverse régionale :

$$Q^{(S)} = C^{(S)}Q^{(R)} \quad (3.3)$$

où  $Q^{(S)}$  est la fonction de répartition inverse du site d'étude,  $Q^{(R)}$  est la fonction de répartition inverse régionale et  $C^{(S)}$  est l'indice de crue relatif au site d'étude.

Ainsi, pour obtenir la distribution des débits au site cible, il faut avoir une estimation de son indice de crue  $C^{(S)}$  et de la loi régionale  $F_R$  - où  $Q^{(R)} = F_R^{-1}$ . La plupart du temps, l'indice de crue sera défini comme la médiane empirique obtenue à partir de l'échantillon du site cible [Robson et Reed, 1999]. D'autres auteurs préconisent l'utilisation de la moyenne empirique [Hosking et Wallis, 1997] ou encore pour la convention interne au Cemagref, le quantile de débit instantané de période de retour 10 ans [Galéa et Prudhomme, 1997]. Toutefois, la médiane empirique paraît être le choix le plus avantageux puisque c'est *l'estimateur empirique* le plus robuste des trois. Ce résultat intuitif est d'ailleurs conforté par une étude menée par Viglione *et al.* [2007]. La procédure d'estimation de la loi régionale  $F_R$  est décrite au sein de l'Algorithme 2.

Dans un contexte de site non jaugé, l'estimation de l'indice de crue  $C^{(S)}$  au site d'étude sera souvent obtenue à partir de régressions linéaires [Robson et Reed, 1999; Kjeldsen et Jones, 2007] ou même non-linéaires [Pandey et Nguyen, 1999] sur les variables physiographiques du bassin, par krigeage ou encore réseau de neurones [Dawson *et al.*, 2006; Muttiah *et al.*, 1997]. Grover *et al.* [2002] analysent la performance de plusieurs modèles d'estimation de l'indice de crue sur des bassins versants de l'Ontario (Canada). Pour leur application, ils ont montré que l'estimation par des méthodes non-linéaires et une fonction objectif adimensionnelle conduisait aux meilleures estimations. De plus, la mise en plan géostatistique des résidus améliore bien souvent les résultats - notamment dans le cadre d'une régression linéaire simple.

## 3.2 Les Limites de l'Indice de Crue

L'indice de crue est le modèle d'estimation régional le plus utilisé à ce jour en hydrologie. Il paraît donc important de connaître les raisons d'un tel succès. Il est évident que l'hypothèse d'invariance



### 3. Le Modèle de l'Indice de Crue

---

#### Algorithme 2 : Ajustement de la loi régionale par la méthode de l'indice de crue

---

**Entrées** : Un groupe de stations supposées homogènes  
**Sorties** : Une statistique  $H$  caractérisant le degré d'homogénéité  
 /\*  $n_{\text{sites}}$  : Nombre de sites dans la région d'étude \*/  
 /\*  $ech(i)$  : Echantillon pour le site  $i$  \*/  
 /\*  $C_i$  : indice de crue pour le site  $i$  \*/  
 /\*  $w_i$  : Taille de l'échantillon au site  $i$  \*/  
 1 Choisir la loi régionale *i.e.* GPD ou GEV ;  
 2 **pour**  $i = 1$  à  $n_{\text{sites}}$  **faire**  
 3      $ech(i) \leftarrow ech(i)/C_i$  ;  
 4     Calculer les 3 premiers  $L$ -moments  $\ell_1^{(i)}$ ,  $\tau^{(i)}$  et  $\tau_3^{(i)}$  pour  $ech(i)$  ;  
 5 **fin**  
 6 Calculer les  $L$ -moments régionaux :

$$\ell_1^R \leftarrow \frac{\sum_i w_i \ell_1^{(i)}}{\sum_i w_i}, \quad \tau^R \leftarrow \frac{\sum_i w_i \tau^{(i)}}{\sum_i w_i}, \quad \tau_3^R \leftarrow \frac{\sum_i w_i \tau_3^{(i)}}{\sum_i w_i}$$

7 Ajuster la loi régionale à partir de  $\ell_1^R$ ,  $\tau^R$  et  $\tau_3^R$ .

---

d'échelle sur une région homogène est séduisante. Qui plus est, elle a l'avantage d'être rapidement abordable pour les services opérationnels qui n'y voient pas un modèle compliqué, mais basé sur des hypothèses physiques relativement raisonnables.

De plus, par construction, l'estimateur régional découlant du concept de l'indice de crue est très robuste. En effet, l'un des paramètres de la GPD le plus difficile à estimer est sans aucun doute le paramètre de forme  $\xi$ . Avec ce modèle,  $\xi$  est estimé comme la moyenne de tous les paramètres de forme de la région. Comme ce paramètre contrôle essentiellement le comportement des plus forts quantiles (*i.e.*  $T > 50$  ans), ce modèle donne l'impression d'être particulièrement adapté à la modélisation des plus forts extrêmes.

C'est sans doute pour ces deux raisons que la méthodologie proposée par [Hosking et Wallis](#) a remporté un franc succès.

Néanmoins, ce modèle est progressivement remis en cause du fait des hypothèses sous-jacentes bien trop fortes [[Katz et al., 2002](#); [Robinson et Sivapalan, 1997](#); [Gupta et al., 1994](#)]. En particulier, selon ce modèle, le coefficient de variation doit être constant sur toute la région. Pourtant, [Gupta et al. \[1994\]](#) montrent sur une application nord américaine que cette variable semble croître pour des tailles de bassin versant faibles et décroître pour des tailles plus grandes.

De plus, pour les lois de probabilité prédites par l'EVT, l'Équation (3.3) implique des relations particulières sur les paramètres des lois de chaque site. En effet, si l'Équation (3.3) est vérifiée et que l'on suppose que la distribution des quantiles de crue suit une GPD ou GEV, alors nous avons les relations :

$$\frac{u^{(S)}}{\sigma^{(S)}} = \frac{u^{(R)}}{\sigma^{(R)}} = K \quad \text{et} \quad \xi^{(S)} = \xi^{(R)}, \quad \forall S \quad (3.4)$$

où  $u^{(S)}$ ,  $\sigma^{(S)}$  et  $\xi^{(S)}$  correspondent aux paramètres de position, d'échelle et de forme du site  $S$ ,  $u^{(R)}$ ,  $\sigma^{(R)}$  et  $\xi^{(R)}$  aux paramètres de position, d'échelle et de forme de la loi régionale et où  $K$  est une constante réelle.

De plus, selon la méthodologie d'[Hosking et Wallis](#), le modèle ne dispose pas de vraisemblance *globale* ou *régionale*. C'est à dire qu'il existe bel et bien une vraisemblance pour chaque site puisque l'on ajuste les paramètres de la GPD/GEV pour chacun d'entre eux. Néanmoins, elle ne se généralise pas à la région toute entière puisque chaque site est ajusté un à un et non globalement. Finalement, il est impossible d'obtenir les incertitudes sur les estimations obtenues.

Le principe de l'indice de crue repose donc sur une invariance d'échelle sur les fonctions de répartition inverses de chaque site de la région. Cette hypothèse est extrêmement forte et ne pourra bien entendu jamais être corroborée par les données. Ainsi, l'étape de construction d'une région

homogène consiste en fait à satisfaire au mieux cette hypothèse. Toutefois, les répercussions peuvent être importantes, notamment en termes de biais si les données s'éloignent trop des hypothèses du modèle.

Enfin, l'information disponible au sein de la région est très mal exploitée. D'une part, les sites pouvant être dépendants, cette indépendance n'est nullement prise en compte. D'autre part, chaque observation d'un site quelconque de la région a le même poids, y compris celles du site cible. Cette manière de procéder est loin d'être optimale puisque, aussi homogène que soit la région d'étude, les données les plus pertinentes pour étudier le site cible sont bien évidemment celles du site cible.

### 3.3 Conclusions et Objectifs à Atteindre

Dans ce chapitre, nous avons introduit les bases de l'estimation régionale des quantiles de crue, notamment en présentant le modèle de référence : l'indice de crue. Ce modèle n'est néanmoins pas sans défaut et nous proposerons de palier quelques unes de ses lacunes. Comme notre travail de recherche n'a pas pour but l'amélioration des techniques de construction des régions dites homogènes, l'influence du degré d'homogénéité de la région sur l'estimation des quantiles de crue sera également analysée.

En conséquence, les Chapitres 5 et 6 proposeront deux nouveaux modèles Bayésiens régionaux. Notamment, nous apporterons un soin tout particulier à intégrer l'information régionale dans un modèle statistique rigoureux. Ce modèle statistique fera la distinction entre l'information disponible au site d'étude et celle en provenance des autres sites de la région. De plus, comme nous le verrons avec le deuxième modèle, nous proposerons une alternative pour obtenir une estimation plus robuste du paramètre de forme  $\xi$  afin d'obtenir des estimations plus efficaces pour les quantiles les plus extrêmes.



# Éléments de Théorie Bayésienne

C E chapitre se veut une introduction succincte à la théorie Bayésienne. Deux de nos modèles développés faisant appel à cette théorie, nous essaierons de mettre en évidence les différences fondamentales existant entre les statistiques classiques et Bayésiennes.

*Remarque.* Convention Typologique

Jusqu'à présent, nous avons utilisé la notation  $f$  pour définir une densité de probabilité. Pour les statistiques Bayésiennes,  $\pi$  est souvent utilisé pour représenter la densité d'une loi de probabilité. Nous respectons cette convention ; de sorte que la présence de  $\pi$  nous rappellera que nous sommes dans un procédure d'estimation Bayésienne.

## 4.1 Vers les Approches Bayésiennes

Le but d'une analyse statistique est bien souvent de modéliser le comportement d'un processus stochastique. Dans notre travail de recherche, le processus à modéliser correspond à des caractéristiques de crue. Ce processus sera supposé gouverner par une variable aléatoire, elle même régie par une densité de probabilité  $\pi$ .

En pratique,  $\pi$  ne sera que très rarement connue et un choix arbitraire devra être fait. Ainsi, la variable aléatoire représentant le processus sera supposée appartenir à une famille paramétrique définie par  $\{\pi(x; \theta) : \theta \in \Theta\}$ ,  $\theta$  étant le vecteur des paramètres de la densité  $\pi$  et  $\Theta$  l'espace de définition pour  $\theta$ .

Lors d'une analyse statistique classique, le processus sera modélisé en ajustant une valeur adéquate pour le vecteur des paramètres  $\theta$ , par exemple par maximum de vraisemblance. Ainsi, les statistiques classiques supposent donc qu'il existe une densité  $\pi$  et une valeur fixe  $\theta^*$  modélisant parfaitement le processus.

C'est sur ce dernier point que les statistiques Bayésiennes diffèrent. En effet, cette branche de la statistique suppose qu'il n'existe pas de valeur  $\theta^*$  idéale mais que le vecteur des paramètres du modèle  $\theta$  est lui même une variable aléatoire. Ainsi,  $\theta$  admet une densité de probabilité ; en particulier avec des *valeurs plus probables*.

## 4.2 Le Théorème de Bayes et la Loi a Priori

Les statistiques Bayésiennes supposent qu'il est possible d'avoir une *intuition* sur  $\theta$  sans pour autant faire appel aux données disponibles. Ces intuitions peuvent alors être traduites en termes de loi de probabilité ; ce sont les lois **a priori**. Ces lois *a priori* sont définies par des paramètres, appelés **hyperparamètres**, devant être estimés sans avoir recours aux données disponibles. Toute la théorie Bayésienne repose sur un résultat simple permettant de passer d'une loi *a priori* sur les paramètres à une loi dite **a posteriori** sur les paramètres sachant les données.

**Théorème 4.1.** *Théorème de Bayes [Bayes, 1763]*

$$\pi(\theta|x) = \frac{\pi(\theta)\pi(x;\theta)}{\int_{\Theta}\pi(\theta)\pi(x;\theta)d\theta} \quad (4.1)$$

où  $\pi(\theta|x)$  correspond à la loi *a posteriori*,  $\pi(\theta)$  la loi *a priori* et  $\pi(x;\theta)$  la vraisemblance.

*In fine*, nous n'obtenons plus seulement une estimation ponctuelle des paramètres  $\theta$  mais leurs distributions toutes entières  $\pi(\theta|x)$ . Ainsi, il n'est plus nécessaire de considérer des résultats asymptotiques pour obtenir, par exemple, la variance des paramètres. De même, la prédiction est entièrement scindée dans la théorie Bayésienne. En effet, si l'on appelle  $Z$  une future observation ayant pour distribution  $\pi(z|\theta)$  et  $\pi(\theta|x)$  la loi *a posteriori* obtenue à partir de nos observations  $x$  alors nous pouvons obtenir la distribution de  $Z$  sachant toutes nos observations.

**Définition 4.1.** Loi prédictive

$$\pi(z|x) = \int_{\Theta}\pi(z|\theta)\pi(\theta|x)d\theta \quad (4.2)$$

*Remarque.* Notons que le modèle prédictif Bayésien introduit deux types d'incertitudes, les incertitudes engendrées par l'estimation du modèle  $\pi(\theta|x)$  et celles engendrées par la variabilité des observations futures  $\pi(z|\theta)$ .

### 4.3 L'Inférence Bayésienne

Une différence fondamentale entre les statistiques classiques et Bayésienne est de considérer les paramètres estimés comme des variables aléatoires. Bien que la connaissance de la distribution des paramètres au travers de la loi *a posteriori* soit importante, il est parfois nécessaire de résumer la distribution *a posteriori* afin d'obtenir, comme pour l'inférence classique, un jeu de paramètres optimaux.

La détermination de ces paramètres optimaux n'est pas unique et dépend essentiellement de la définition apportée au mot *optimal*. Ce critère d'optimalité peut être représenté par une fonction perte  $\mathcal{P}(\theta^*, \hat{\theta})$  quantifiant le *coût à payer* en estimant  $\theta^*$  par  $\hat{\theta}$ . On retrouve souvent les mêmes fonctions pertes mais il n'est pas insensé d'adopter un choix spécifique au problème rencontré. Néanmoins, les plus couramment utilisées sont les fonctions :

Perte Quadratique :

$$\mathcal{P}(\theta^*, \hat{\theta}) = (\theta^* - \hat{\theta})^2 \quad (4.3)$$

Perte Absolue :

$$\mathcal{P}(\theta^*, \hat{\theta}) = |\theta^* - \hat{\theta}| \quad (4.4)$$

Perte 0-1 :

$$\mathcal{P}(\theta^*, \hat{\theta}) = \begin{cases} 0, & \text{si } |\theta^* - \hat{\theta}| \leq \varepsilon \\ 1, & \text{si } |\theta^* - \hat{\theta}| > \varepsilon \end{cases} \quad (4.5)$$

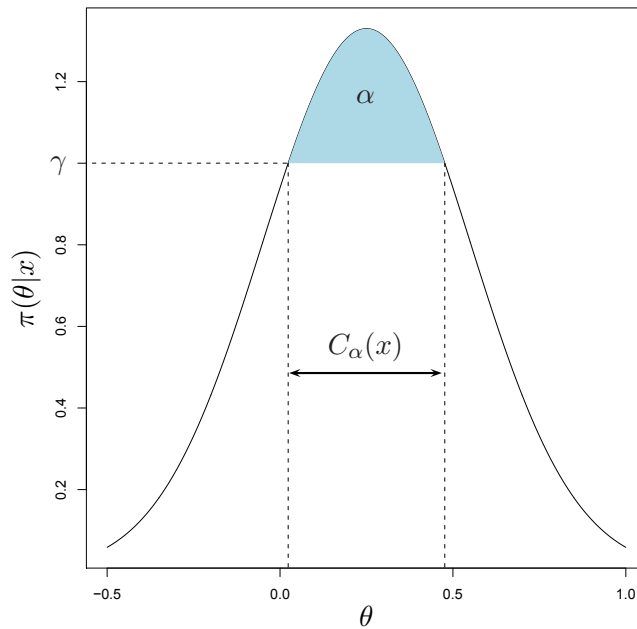
où  $\varepsilon$  quantifie la précision de l'estimation.

Perte linéaire :

Pour  $g$  et  $h$  fonctions connues,

$$\mathcal{P}(\theta^*, \hat{\theta}) = \begin{cases} g(\theta^* - \hat{\theta}), & \text{si } \hat{\theta} \geq \theta^* \\ h(\hat{\theta} - \theta^*), & \text{si } \hat{\theta} < \theta^* \end{cases} \quad (4.6)$$

Le jeu optimal de paramètres correspond à celui qui minimise la fonction perte.



**Fig. 4.1:** Définition d'un intervalle de crédibilité  $C_\alpha(x)$  connaissant la loi a posteriori  $\pi(\theta|x)$ .

*Remarque.* Les fonctions pertes énoncées plus haut correspondent à des estimateurs que nous connaissons bien en statistique classique puisque :

- la fonction perte quadratique est minimisée par l'espérance de la loi *a posteriori* ;
- la fonction perte absolue est minimisée par la médiane de la loi *a posteriori* ;
- la fonction perte 0–1 est minimisée par le mode de la loi *a posteriori* dès lors que  $\varepsilon$  est suffisamment petit.

En pratique, le jeu optimal de paramètres sera donc estimé à partir de la moyenne, la médiane ou encore le mode de la loi *a posteriori*. Notons que la loi *a posteriori* peut ne pas être connue analytiquement. De sorte que la loi *a posteriori* ne sera connue qu'à partir d'un échantillon étendu distribué selon cette dernière - à l'aide des algorithmes de type MCMC ou Gibbs. Pour ces cas spécifiques, le jeu optimal de paramètres sera donc estimé en prenant la moyenne, médiane ou mode empiriques obtenu à partir de cet échantillon étendu.

## 4.4 Les Intervalles de Crédibilité

L'idée des **intervalles de crédibilité** est de proposer un équivalent aux intervalles de confiance des statistiques classiques. Puisque dans l'approche Bayésienne, les paramètres à estimer sont considérés comme des variables aléatoires, cette notion d'intervalle de crédibilité est toute naturelle.

### Définition 4.2. Intervalle de Crédibilité

Un intervalle de crédibilité  $C_\alpha(x)$  associé à la probabilité  $\alpha$ , et sachant les observations  $x$  est de la forme :

$$C_\alpha(x) = \{\theta : \pi(\theta|x) \geq \gamma\} \quad (4.7)$$

où  $\gamma$  est défini tel que

$$\int_{C_\alpha(x)} \pi(\theta|x) d\theta = \alpha \quad (4.8)$$

La Figure 4.1 illustre la construction d'un tel intervalle de crédibilité. D'une manière un peu moins formelle, un intervalle de crédibilité  $C_\alpha(x)$  pour un niveau de confiance  $\alpha$  est défini tel que la densité *a posteriori*  $\pi(\theta|x)$  est maximale tout en imposant  $\Pr[\theta \in C_\alpha(x)] = \alpha$ .

Il est important de souligner les points communs et les différences entre les intervalles de confiance et de crédibilité. Ces deux approches vérifient, sous l'hypothèse que le modèle est le bon,  $\Pr[\theta \in I] = \alpha$ , où  $I$  est soit un intervalle de confiance, soit de crédibilité associé au même niveau de confiance  $\alpha$ . Ces deux intervalles renseignent donc sur la précision de notre estimation.

En revanche, l'intervalle de crédibilité est dépendant de la loi *a priori*  $\pi(\theta)$ ; de sorte que deux lois *a priori* différentes conduiront à deux intervalles de crédibilité différents. C'est l'une des plus fortes controverses des statistiques Bayésiennes : l'analyse est dépendante de la loi *a priori*. La définition de la loi *a priori* doit donc être établie avec attention et honnêteté lors de l'exploitation de l'information dont nous disposons *a priori*.

En effet, si nous intégrons plus d'information *a priori* qu'il n'est raisonnable, les intervalles de crédibilité seront bien plus resserrés. De plus, les estimations elles-mêmes pourraient être complètement fausses si jamais l'information *a priori* n'était pas pertinente avec les données disponibles.

Au contraire, avec une loi *a priori* non informative, et d'après le Théorème 4.1, la loi *a posteriori* sera (approximativement) identique à la vraisemblance et nous nous ramenons aux statistiques classiques.

# 5 Un Modèle Bayésien pour l'Estimation Régionale des Débits

DANS ce chapitre, nous introduisons les idées qui nous ont encouragés à développer un nouveau modèle régional pour l'estimation des quantiles de crue. Comme énoncé au sein du Chapitre 3, le concept de l'indice de crue sera notre point de départ et méthode de référence pour l'estimation régionale des débits caractéristiques de crue. Toutefois, nous devons relâcher les hypothèses trop contraignantes d'invariance sur les fonctions de répartition. De même, nous porterons attention à faire la distinction entre l'information au site cible et celle des autres sites de la région. Enfin, le tout sera intégré au sein d'un modèle statistique bien fondé.

## 5.1 L'Importance du Site Cible et de la Région

Partant du constat que l'information disponible au sein de toute la région homogène est mal exploitée par le concept de l'indice de crue (*cf.* Section 3.2), nous proposons une nouvelle approche. En effet, l'information la plus précieuse, aussi réduite soit elle, est celle dont l'hydrologue dispose au site d'intérêt. L'information en provenance des autres sites réputés similaires sera alors considérée comme une *pseudo information* sur le site cible. En particulier, afin d'accorder un poids plus important aux données du site cible, cette pseudo information sera perçue comme une intuition sur le comportement du site d'intérêt.

Pour faire le distinguo entre ces deux types d'information, la théorie Bayésienne nous est utile. Pour cela, rappelons le Théorème de Bayes [Bayes, 1763] :

$$\pi(\theta|x) = \frac{\pi(\theta)\pi(x;\theta)}{\int_{\Theta}\pi(\theta)\pi(x;\theta)d\theta} \quad (5.1)$$

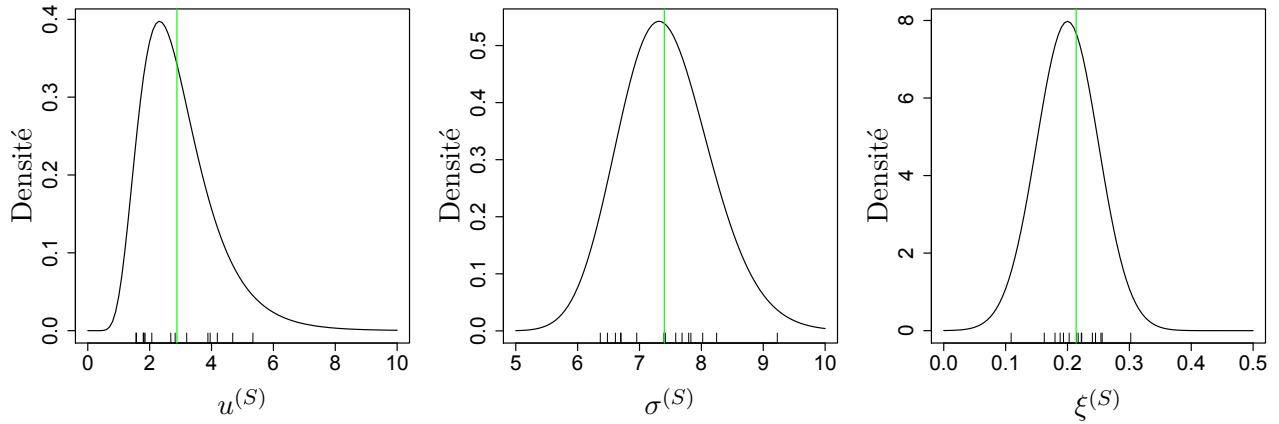
où  $x$  est le vecteur des données disponibles,  $\Theta$  l'espace de définition des paramètres de la loi,  $\pi(\theta|x)$  la loi *a posteriori*,  $\pi(\theta)$  la loi *a priori* et  $\pi(x;\theta)$  la vraisemblance.

La loi *a priori*  $\pi(\theta)$  est souvent associée à un jugement d'expert, c'est à dire une connaissance préalable que l'on pourrait avoir sur le processus que l'on souhaite modéliser avant d'avoir recours aux données disponibles. La loi *a posteriori*  $\pi(\theta|x)$  correspond donc à la connaissance dite d'expert et celle apportée par les données.

Dans un contexte d'estimation régionale et plus particulièrement pour le concept de l'indice de crue, nous proposons que la connaissance *a priori* soit définie à partir de l'information en provenance des sites de la région homogène - le site cible étant exclu. En effet, sous l'hypothèse de l'indice de crue, les sites de la région ont un comportement similaire et il est donc logique de considérer ces sites comme une intuition sur le comportement du site d'étude.

Les données disponibles au site d'étude seront quant à elles intégrées de manière plus conventionnelle au sein de la fonction de vraisemblance  $\pi(x;\theta)$ . Ainsi, l'information disponible au sein de la région homogène est utilisée de manière différente et notamment l'information au site cible peut jouer un rôle prépondérant.





**Fig. 5.1:** Assouplissement des hypothèses du modèle de Dalrymple sur les paramètres de la GPD. Ligne verticale : contraintes de l'indice de crue. Densités : contraintes du modèle proposé.

## 5.2 Les Améliorations

Le modèle que nous proposons [Ribatet *et al.*, 2007b] permet non seulement d'obtenir des estimations plus robustes sur les quantiles de crue mais aussi de relâcher les hypothèses de l'indice de crue. De plus, ce modèle possède des propriétés théoriques intéressantes que nous présentons dans les paragraphes suivants et qui en sont les points forts.

### 5.2.1 Un Assouplissement des Hypothèses de l'Indice de Crue

Rappelons que le modèle de l'indice de crue impose une relation déterministe sur les paramètres de la GPD pour tous les sites de la région homogène, c'est à dire que :

$$u^{(S)} = K u^{(R)}, \quad \sigma^{(S)} = K \sigma^{(R)}, \quad \xi^{(S)} = \xi^{(R)}, \quad \forall S \quad (5.2)$$

où  $u^{(S)}, \sigma^{(S)}$  et  $\xi^{(S)}$  correspondent aux paramètres de position, d'échelle et de forme du site  $S$ ,  $u^{(R)}, \sigma^{(R)}$  et  $\xi^{(R)}$  aux paramètres de position, d'échelle et de forme de la loi régionale et  $K$  est une constante réelle.

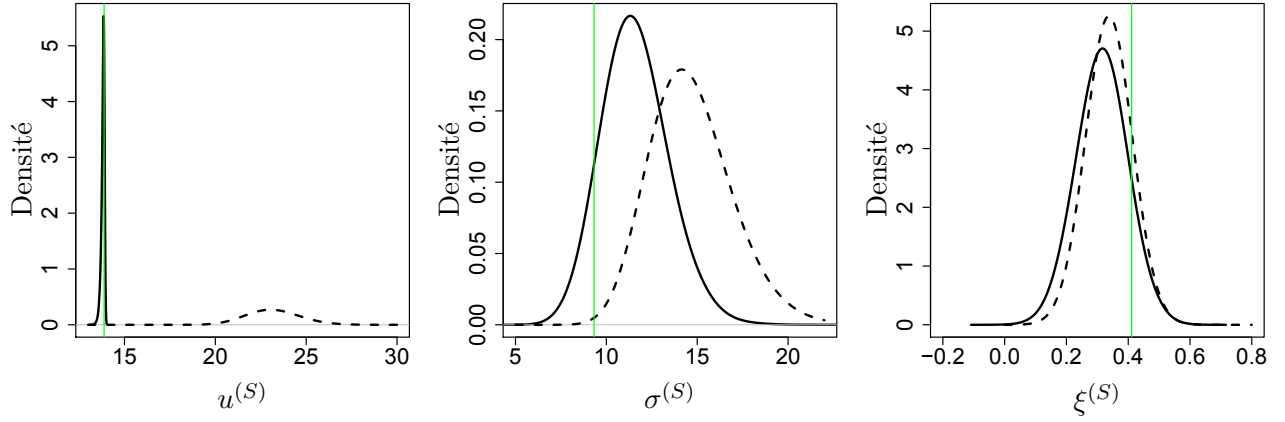
Pour notre modèle, les relations (5.2) ne sont plus imposées mais relaxées de telle sorte que les paramètres  $u^{(S)}, \sigma^{(S)}$  et  $\xi^{(S)}$  sont désormais considérés comme des variables aléatoires et vérifient :

$$\Pr \left[ u^{(S)} \in (a_1, b_1) \right] = p_u, \quad \Pr \left[ \sigma^{(S)} \in (a_2, b_2) \right] = p_\sigma, \quad \Pr \left[ \xi^{(S)} \in (a_3, b_3) \right] = p_\xi, \quad \forall S \quad (5.3)$$

où les  $a_i$  et  $b_i$  sont des réels,  $i = 1, \dots, 3$  et  $p_u, p_\sigma$  et  $p_\xi \in [0, 1]$ .

Ainsi, nous avons assoupli les hypothèses de l'indice de crue qui imposent une contrainte entièrement déterministe sur les paramètres de la GPD au profit d'une contrainte sur la variabilité de ces paramètres.

La Figure 5.1 en est une illustration. Là où le modèle de l'indice de crue impose une unique valeur pour les paramètres de la GPD (représentée par les lignes verticales), notre modèle propose *a priori* des régions à plus ou moins forte densité pour ces paramètres (représentées par les marginales de la loi *a priori*). Nous remarquons de plus que les valeurs imposées par le modèle de l'indice de crue se retrouvent dans des régions à forte probabilité. Néanmoins, elles ne correspondent pas forcément avec les modes des marginales de la loi *a priori* - puisque le site cible est omis lors de la définition de la loi *a priori*.



**Fig. 5.2:** Précision de l'information régionale à l'aide des données au site cible. Lignes en pointillées : lois *a priori*. Lignes pleines : lois *a posteriori*. Lignes verticales : valeurs de références.

### 5.2.2 Une Utilisation Différenciée de l'Information

Un de nos objectifs était de proposer un modèle faisant la distinction entre les données du site d'étude et celles des autres sites de la région. En particulier, l'information des sites autres que le site cible est intégrée au sein de la loi *a priori* - cf. densités de la Figure 5.1. Les données disponibles au site d'étude intégrées dans la vraisemblance viennent préciser cette information régionale .

Toutefois, il est également possible que l'information régionale ne soit pas pertinente pour le site d'étude ; auquel cas, l'information du site cible pourra contredire la loi *a priori*. Notamment, il est possible que l'information régionale soit adéquate pour un des paramètres de la GPD et très mauvaise pour d'autres.

La Figure 5.2 en est une illustration. Pour les paramètres de position  $u^{(S)}$  et d'échelle  $\sigma^{(S)}$ , les marges de la loi *a posteriori* sont bien plus resserrées et proches des valeurs de références. En revanche, pour le paramètre de forme  $\xi^{(S)}$ , la marge de la loi *a posteriori* correspondante est plus lâche que celle de la loi *a priori* reflétant une incertitude plus grande sur ce paramètre.

### 5.2.3 Des Propriétés Asymptotiques Établies

Contrairement à l'estimateur de l'indice de crue, la théorie Bayésienne nous montre que cet estimateur Bayésien est **consistant** ; c'est à dire qu'il converge presque sûrement vers la valeur théorique. Il possède également des propriétés de Normalité asymptotique et d'indépendance asymptotique envers la loi *a priori* - ou plutôt dans notre contexte vis à vis de la configuration de la région homogène. Ces propriétés sont des éléments statistiques essentiels pour tout estimateur et valident théoriquement l'utilisation de notre approche.

#### **Théorème 5.1.** *Consistance*

Soient  $X$  une v.a. ayant pour paramètre  $\theta_{\text{True}}$  et  $\mathbf{x}(n) = (x_1, \dots, x_n)$  des réalisations indépendantes de  $X$ . Supposons de plus que :

$$\Pr [X \in V] > 0, \quad \forall V \in \mathcal{V}(\theta_{\text{True}}), \quad X \sim \pi(\theta) \quad (5.4)$$

où  $\mathcal{V}(\theta_{\text{True}})$  représente l'ensemble des voisinages de  $\theta_{\text{True}}$  et  $\pi(\theta)$  la loi *a priori*.

Alors,

$$\Pr [\theta = \theta_{\text{True}}] = 1, \quad n \longrightarrow \infty, \quad \theta \sim \pi(\theta|\mathbf{x}(n)) \quad (5.5)$$

Autrement dit, ce théorème démontre la convergence vers la vraie valeur du paramètre dès lors que la densité *a priori* n'est pas nulle en cette valeur. Dans notre contexte d'estimation régionale, cela

signifie que l'estimateur converge vers la valeur théorique dès lors que la région supposée homogène n'est pas complètement incompatible avec le site d'étude.

Le théorème suivant montre un résultat encore plus important.

**Théorème 5.2.** *Normalité asymptotique*

Soient  $X$  une v.a. ayant pour paramètre  $\theta_{\text{True}}$  et  $\mathbf{x}(n) = (x_1, \dots, x_n)$  des réalisations indépendantes de  $X$ . Supposons de plus que :

$$\Pr[X \in V] > 0, \quad \forall V \in \mathcal{V}(\theta_{\text{True}}), \quad X \sim \pi(\theta) \quad (5.6)$$

Alors,

$$\theta|\mathbf{x}(n) \longrightarrow \mathcal{N}\left(\theta_{\text{True}}, I_n(\theta_{\text{True}})^{-1}\right), \quad n \rightarrow \infty \quad (5.7)$$

où  $I_n(\theta_{\text{True}})^{-1}$  est l'inverse de la matrice d'information de Fisher.

Ce théorème affirme que si nous disposons de suffisamment de données, alors la distribution *a posteriori* devient indépendante de la loi *a priori* et converge vers la même loi asymptotique que celle du maximum de vraisemblance - *i.e.* une loi Normale de moyenne la valeur théorique et de matrice de variance-covariance égale à l'inverse de la matrice d'information de Fisher. Autrement dit, d'un point de vue plus pratique, deux lois *a priori* différentes conduisent asymptotiquement vers la même densité *a posteriori*. Dans le contexte d'estimation régionale, le Théorème 5.2 établit qu'indépendamment de la région homogène considérée et de la loi *a priori*  $\pi(\theta)$ , l'estimateur converge vers la distribution théorique.

### 5.2.4 Une Quantification des Incertitudes Rendue Possible

Puisque nous nous plaçons dans le cadre Bayésien, les quantiles estimés mais aussi leurs incertitudes sont entièrement disponibles au travers d'intervalles de crédibilité - *cf.* Section 4.4. Contrairement aux statistiques classiques, nous n'obtenons pas une estimation ponctuelle des paramètres du modèle ou d'un quantile cible mais la densité toute entière. Il est donc immédiat d'obtenir l'estimation d'un quantile ainsi qu'un intervalle de crédibilité associé.

## 5.3 L'Article

# A Regional Bayesian POT Model for Flood Frequency Analysis

Mathieu Ribatet<sup>1,2</sup>, Eric Sauquet<sup>2</sup>, Jean-Michel Grésillon<sup>2</sup>, Taha B.M.J. Ouarda<sup>1</sup>

<sup>1</sup> INRS-ETE, University of Québec, 490, de la Couronne Québec, Qc, G1K 9A9, CANADA

<sup>2</sup> Cemagref, 3 bis quai Chauveau CP 220, 69336 Lyon Cedex 09, FRANCE

Published online: 9 august 2006

**Abstract.** Flood Frequency Analysis is usually based on the fitting of an extreme value distribution to the series of local streamflow. However, when the local data series is short, frequency analysis results become unreliable. Regional frequency analysis is a convenient way to reduce the estimation uncertainty. In this work, we propose a regional Bayesian model for short record length sites. This model is less restrictive than the Index Flood model while preserving the formalism of “homogeneous regions”. Performance of the proposed model is assessed on a set of gauging stations in France. The accuracy of quantile estimates as a function of homogeneousness level of the pooling group is also analysed. Results indicate that the regional Bayesian model outperforms the Index Flood model and local estimators. Furthermore, it seems that working with relatively large and homogeneous regions may lead to more accurate results than working with smaller and highly homogeneous regions.

**Key words.** Regional Frequency Analysis – Bayesian Inference – Index Flood – *L-moments* – Markov Chain Monte Carlo

## 1 Introduction

Flood frequency analysis is essential in preliminary studies to define the design flood. Methods for estimating design flow usually consist of fitting one of the distributions given by the extreme value theory to a sample of flood events. If modelling exceedance over a threshold is of interest, a theoretical justification (Fisher and Tippett, 1928; Balkema and Haan, 1974; Pickands, 1975) exists for the use of the Generalized Pareto distribution (GP).

$$F(x) = 1 - \left(1 + \frac{\xi(x - \mu)}{\sigma}\right)^{-1/\xi} \quad (1)$$

Correspondence to: ribatet@lyon.cemagref.fr

where  $1 + \xi(x - \mu)/\sigma > 0$ ,  $\sigma > 0$ .  $\mu$ ,  $\sigma$  and  $\xi$  are the location, scale and shape parameters. This distribution is defined for  $\xi \neq 0$ , and can be derived by continuity in the case  $\xi = 0$ , corresponding to the Exponential case:

$$F(x) = 1 - \exp\left(-\frac{x - \mu}{\sigma}\right) \quad (2)$$

A comprehensive review of the Extreme Value Theory is given by Embrechts et al. (1997) and Coles (2001).

However, frequency analysis can lead to unreliable flood quantiles when little data is available at the site of interest. A convenient way to improve estimates of flood statistics is to incorporate data from other gauged locations in the estimation procedures. This approach is widely applied in hydrology and is known as Regional Flood Frequency Analysis (**RFFA**). One of the most popular and simple approaches privileged by engineers is the Index Flood method (Dalrymple, 1960). The standard procedure allows: a) the delineation of homogeneous regions, i.e. a set of sites which behave - hydrologically and/or statistically - in the same way; b) the derivation of a regional flood frequency distribution; c) and the estimation of the parameters and quantiles at the site of interest.

Regions are collections of gauged basins with similar site characteristics related to the flood magnitude. The pooled stations are not necessarily in the proximity of the site of interest. Forming homogeneous regions can be achieved in various ways. Regions have first been geographically established. More recent work promoted the use of geographically non-contiguous regions (Burn, 1990; GREHYS, 1996). Recent research has defined the concept of “region of influence” (Acreman and Wiltshire, 1989). Other techniques can be used such as Artificial Neural Networks to identify groups of stations (Hall et al., 2002).

The Index Flood model assumes that flood distributions at all sites within a region are identical, up to a scale factor. The Index Flood approach is not exempt from critics as its application requires strong assumptions. One major implicit assumption, noticed by Gupta et al. (1994), is that the coefficient of variation of peak

flows is to be constant across the region. This fundamental property seems not to be verified in practice (Robinson and Sivapalan, 1997) and not physically justified (Katz et al., 2002).

The assumptions of the Index Flood model need often to be relaxed to suit the observations. For this purpose, Gabriele and Arnell (1991) proposed a hierarchical approach to RFFA. The skewness is still supposed to be constant on the whole region, but the coefficient of variation and the mean annual flood can vary slowly from one subregion to another. However, the two authors underlined the practical difficulty to delineate these subregions.

In the Index Flood model, each observation from any site within the region have the same weight. However, it seems not optimal as, obviously, the most precious information come from the target site. Indeed target data - even short - are the only one which are “really” distributed as the target site.

We suggest here to carry out a Bayesian approach that encompasses the classical Index Flood model and uses the whole data in a more efficient manner. In summary, the proposed Bayesian approach differs from the Index Flood model as it: a) uses the at-site information in a more efficient way since this approach distinguishes the target site data and the regional data; b) does not impose a purely deterministic relationship between sites within the region.

The main goal of this article is to test the efficiency and robustness of the developed regional Bayesian model when dealing with short record length series. For this purpose, classical frequency analysis *i.e.* local and traditional RFFA will be compared to the suggested regional Bayesian approach. Section 2 presents a brief summary of the classical Index Flood model. Relevant theoretical aspects of Bayesian theory are introduced and applied to flood modelling in a RFFA context in Section 3. Section 4 describes the data set used to illustrate the method. Section 5 describes the procedure used to elicit the prior distribution. Section 6 outlines the weaknesses and strengths of each approach on a typical homogeneous region. Finally section 7 presents an analysis of the effect of homogeneity level on quantile estimation.

### 2 The Index Flood model

The Index Flood method states that flood frequency distributions within a particular region are supposed to be identical when divided by a scale factor - namely the Index Flood. Mathematically, this assumption is expressed as:

$$Q^{(S)} = C^{(S)}Q^{(R)} \quad (3)$$

where  $Q^{(S)}$  is the quantile function at site  $S$ ,  $C^{(S)}$  the Index Flood at site  $S$  and  $Q^{(R)}$  the regional quantile function *i.e.* the dimensionless quantile function valid across the homogeneous region.

Equation (3) is the core of the model and leads to strong constraints concerning at-site distribution parameters. Consequently, the shape parameter is the same

throughout the homogeneous region, whereas the location and scale parameters have simple scaling behaviour - see Appendix A.

Equation (3) is supposed to be satisfied if all sites are hydrologically and/or statistically similar. Therefore, one of the main aspects of this approach is to identify a homogeneous region which includes the target site.

Similarity in basin characteristics is necessary but not sufficient to ensure the homogeneity of the region regarding the statistics of the flood peaks. Hosking and Wallis (1993, 1997) suggested a heterogeneity measure  $H_1$  to assess if a region is “acceptably homogeneous” ( $H_1 < 1$ ), “probably heterogeneous” ( $1 \leq H_1 < 2$ ) or “definitively heterogeneous” ( $H_1 \geq 2$ ). Note the case  $H_1 \leq 0$  seems to detect correlations between sites within the region.

Once the region satisfies the homogeneity test of Hosking and Wallis (1993, 1997), the regional flood frequency distribution and the related at-site distribution is computed in a classical way. That is, by fitting the regional distribution to the weighted mean of sample  $L$ -moments. Details for computing heterogeneity statistics, regional flood frequency and at-site distribution can be found in (Hosking and Wallis, 1993, 1997).

By definition of the Index Flood model, it can be seen that any realisations of each samples have the same weight. Giving equal weights to all site observations is debatable since the most relevant information is certainly the target site one. Relevance of the target site information is obvious as this is the only one which is “really” distributed as the target site. Thus, in this approach the available information is not efficiently used.

### 3 Regional Bayesian model

The Bayesian concepts have already been applied with success to the regional frequency analysis of extreme rainfalls (Coles and Pericchi, 2003) and floods (Madsen and Rosbjerg, 1997; Northrop, 2004). Regional information is not used to build a regional distribution but to specify a kind of “suspicion” about the target site distribution. This is easily achieved in the Bayesian framework through the so called prior distribution.

The main goal of Bayesian inference is to compute the posterior distribution. The posterior distribution  $\pi(\theta|x)$  is given by the Bayes Theorem:

$$\pi(\theta|x) = \frac{\pi(\theta)\pi(x;\theta)}{\int_{\Theta} \pi(\theta)\pi(x;\theta)d\theta} \propto \pi(\theta)\pi(x;\theta) \quad (4)$$

where  $\theta$  is the vector of parameters of the distribution to be fitted,  $\Theta$  is the space parameter.  $\pi(x;\theta)$  is the likelihood function,  $x$  is the vector of observation and  $\pi(\theta)$  is the prior distribution.

In theory, the posterior distribution is entirely known but is often insolvable - because of the integral. One of the solutions is to fix a prior model which leads to an analytical - or semi-analytical - posterior distribution and which allows the posterior distribution to be computed more easily (Parent and Bernier, 2003). Nevertheless,

the most convenient way is to implement Markov Chain Monte Carlo (**MCMC**) techniques to sample the posterior distribution. This approach avoids using a purely artificial prior model with no theoretical and/or physical justifications.

For our application, the likelihood function corresponds to the GP distribution as peaks over a threshold are of interest. From Eq. (4), if the prior distribution is known, posterior distribution can be computed - up to a constant. The next section describes how to define the prior distribution.

### 3.1 Prior Distribution

The prior model is usually a multivariate distribution which must represent beliefs about the distribution of the parameters *i.e.*  $\mu$ ,  $\sigma$  and  $\xi$  prior to having any information about the data.

As the proposed model is fully parametric, the prior distribution  $\pi(\theta)$  is a multivariate distribution entirely defined by its hyper parameters. In our case study, the marginal prior distributions were supposed to be independent lognormal for both location and scale parameters and normal for the shape parameter. Thus,

$$\pi(\theta) \propto J \exp [(\theta' - \gamma)^T \Sigma^{-1} (\theta' - \gamma)] \quad (5)$$

where  $\gamma$ ,  $\Sigma$  are hyper parameters,  $\theta' = (\log \mu, \log \sigma, \xi)$  and  $J$  is the Jacobian of the transformation from  $\theta'$  to  $\theta$ , namely  $J = 1/\mu\sigma$ .  $\gamma = (\gamma_1, \gamma_2, \gamma_3)$  is the mean vector,  $\Sigma$  is the covariance matrix. As marginal priors are supposed to be independent,  $\Sigma$  is a 3-3 diagonal matrix with diagonal elements  $d_1, d_2, d_3$ .

### 3.2 Estimation of the hyper parameters

Hyper parameters are defined through the Index Flood concept. Consider all sites of a region except the target site - say the  $j$ -th site. A set of pseudo target site parameters can be computed:

$$\tilde{\mu}^{(i)} = \mu_*^{(i)} C^{(j)} \quad (6)$$

$$\tilde{\sigma}^{(i)} = \sigma_*^{(i)} C^{(j)} \quad (7)$$

$$\tilde{\xi}^{(i)} = \xi_*^{(i)} \quad (8)$$

for all  $i \neq j$ , where  $C^{(j)}$  is the at-site Index Flood and  $\mu_*^{(i)}, \sigma_*^{(i)}, \xi_*^{(i)}$  are respectively the location, scale and shape at-site parameter estimates from rescaled sample.

Under the hypothesis of the Index Flood model, pseudo parameters  $(\tilde{\mu}^{(i)}, \tilde{\sigma}^{(i)}, \tilde{\xi}^{(i)})$  for  $i \neq j$  are expected to be similar to the target site distribution parameters. Note that, information from the target site sample is not used to elicit the prior distribution. Thus,  $C^{(j)}$  in equations (6) and (7) must be estimated without use of the  $j$ -th site sample.

From these pseudo parameters, hyper parameters can be computed:

$$\gamma_1 = \frac{1}{N-1} \sum_{i \neq j} \log \tilde{\mu}^{(i)}, \quad d_1 = \frac{1}{N-1} \sum_{i \neq j} Var [\log \tilde{\mu}^{(i)}] \quad (9)$$

$$\gamma_2 = \frac{1}{N-1} \sum_{i \neq j} \log \tilde{\sigma}^{(i)}, \quad d_2 = \frac{1}{N-1} \sum_{i \neq j} Var [\log \tilde{\sigma}^{(i)}] \quad (10)$$

$$\gamma_3 = \frac{1}{N-1} \sum_{i \neq j} \tilde{\xi}^{(i)}, \quad d_3 = \frac{1}{N-2} \sum_{i \neq j} (\tilde{\xi}^{(i)} - \gamma_3)^2 \quad (11)$$

It is important at this step to incorporate the uncertainties on the elicitation of the prior distribution. Indeed, it may avoid problems related to misleading information resulting from a region not so homogeneous and moderating a ‘‘suspicion’’ that may be too true.

For this purpose, two types of uncertainties are taken into account: the one from parameter estimation, and the other one from target site Index Flood estimation. Thus, hyper parameters  $\gamma_1$  and  $\gamma_2$  are estimated differently than  $\gamma_3$  as pseudo parameters for location and scale parameters depends on the target site Index Flood. Under the hypothesis of independence between  $C^{(j)}$  and  $\mu_*^{(i)}, \sigma_*^{(i)}$  the variance terms in Eq. (9) and (10) are computed according these two types of uncertainties:

$$Var [\log \tilde{\mu}^{(i)}] = Var [\log C^{(j)}] + Var [\log \mu_*^{(i)}] \quad (12)$$

$$Var [\log \tilde{\sigma}^{(i)}] = Var [\log C^{(j)}] + Var [\log \sigma_*^{(i)}] \quad (13)$$

The independence assumption between  $C^{(j)}$  and  $\mu_*^{(i)}, \sigma_*^{(i)}$  is not too restrictive as the target site Index Flood  $C^{(j)}$  is estimated independently from  $\mu_*^{(i)}, \sigma_*^{(i)}$ .

Note that  $Var [\log \mu_*^{(i)}]$  are estimated thanks to Fisher Information and the Delta method. Estimation of  $Var [\log C^{(j)}]$  is a special case and depends on the method for estimating the at-site Index Flood. Nevertheless, it is always possible to carry out an estimation of this variance, at least through standard errors.

### 3.3 Specificities of the proposed prior model

The construction of the prior distribution with the regional information was already suggested by (Northrop, 2004). Nevertheless, the location parameter - or equivalently the threshold in the GP case - was supposed to be known. Yet, from a theoretical point of view, the location parameter can not be known prior to having any information from the target site sample. Northrop (2004) developed a similar approach based on the Index Flood but uncertainty associated with the scale factor prediction was not considered. The prior distribution was elicited directly from the distribution of the ‘‘pseudo target site’’ estimates  $(\tilde{\mu}^{(i)}, \tilde{\sigma}^{(i)}, \tilde{\xi}^{(i)})$ . In this perspective,

“pseudo target site” estimates are viewed as constant and not as random variables. When dealing with sites with a long record, uncertainties on parameter distributions are low. On the contrary, this has a much more impact for the Index Flood as uncertainties are as much larger as the target site Index Flood is estimated without use of at-site data - even with long record length sites. Note that if the prior distribution is overly accurate, estimation and credibility intervals are influenced. For these reasons and unlike the approach proposed by Northrop (2004), the target site Index Flood in the proposed methodology is considered to be a random variable and not a constant.

Thus, our prior distribution is not too falsely “tight fit”. But it reflects “real” beliefs about target site behaviour without any use of target site sample.

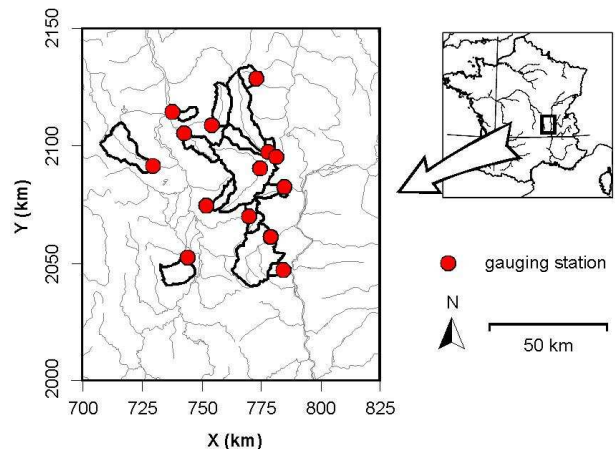
Madsen and Rosbjerg (1997) and Fill and Stedinger (1998) both presented a regional empirical Bayesian estimator. Both models used conjugate families for prior distributions. However, even if conjugate families are convenient devices, they should not only be used just because computations are easier. In their approaches, prior distributions are elicited with quantile regression on relevant physiographic characteristics.

Our approach differs from the two previous empirical Bayesian approaches (i.e. target site sample is not used to elicit priors) and respects in that way absolutely the Bayesian theory. Moreover, conjugate priors are not considered, but priors are suited to the data. For example, the lognormal distribution for both location and shape parameters is justified by a physical and theoretical lower bound as: a) discharge data are naturally non negative; so the location parameter should also be non negative; b) the scale parameter is strictly positive by definition of the GP distribution.

This prior model is quite different from the one proposed by Coles and Tawn (1996) who introduced a lognormal prior distribution only for the scale parameter. Note that it is possible to work with return levels (Coles and Tawn, 1996) or return periods (Crowder, 1992) instead of working with distribution parameters. However, regional information is suited to work directly with distribution parameters. For other studies, such prior models could be of interest if “suspicion” is based on return levels or return periods.

#### 4 Data description

Streamflow data were collected at 48 gauging stations in an area reaching from the 45<sup>th</sup> to 47<sup>th</sup> N and from the 3<sup>rd</sup> to the 8<sup>th</sup> E. The selection of the gauging sites was initially based on the 22 regions into which France is divided for the implementation of the Water Framework Directive (Wasson et al., 2004). Seven regions cover the area under study. These regions were manually delineated taking into account the spatial pattern of mean annual rainfall, elevation and underlying geology. All these variables might influence flood generation processes. Therefore this division is considered as a preliminary guide for pooling stations. According to Hosking and Wallis



**Fig. 1.** Location of the gauging stations within the studied area

(1997), pre-regions were slightly altered by identifying discordant sites while maximising the number of site within the region and meeting the heterogeneity test. Finally, a set of 14 stations was selected for this study. The heterogeneity statistic for this group is  $H_1 = 0.17 < 1$ . Consequently, the region can be considered as “acceptably homogeneous”.

The dataset includes seven tributaries to the Loire River and seven gauging stations located in the French part of the Rhône basin (Fig. 1, Tab. 1). The record length of the instantaneous discharge time series ranges from a minimum of 22 years to a maximum of 37 years, with a mean value of 32 years. The drainage areas vary from 32 to 792 km<sup>2</sup>. Moreover, most of the gauging stations monitored first-order stream catchments *i.e.* all but two pairs of catchments are un-nested. The large majorities of floods in the region occur during autumn and winter and are caused by heavy liquid precipitation.

Partial duration flood series were extracted from the time series for each station. Fig. 2 illustrates time series for stations U4505010, U4635010 and V3015010 and their associated thresholds. Threshold levels were selected to extract in average around two events per year while meeting the criteria of independence between floods (Lang et al., 1999).

Three stations U4505010, U4635010 and V3015010 were of particular interest because of their extended record length of 37 years. The time series of those three sites are displayed in Fig. 2.

In this case study, the scale factor was set to correspond to the 1-year return flood quantile - or equivalently the quantile associated with probability of non exceedance 0.5. Thus, our choice for the Index Flood is close to the sample median which was the reference in Robson and Reed (1999) but differs from Hosking and Wallis (1997) where the sample mean was used. This particular choice for the Index Flood is not unimportant as estimating the quantile with probability of non exceedance 0.5 is more robust than estimating the sample mean. Analysing the influence of Index Flood selection is beyond the scope of this work. The main point is to

Table 1. Characteristics of the stations of the homogeneous region

Code	Station	Area (km <sup>2</sup> )	X (km)	Y (km)	Record
K0624510	The Bonson river at St Marcellin	104	744.72	2053.90	1971-2003
K0663310	The Coise river at Larajasse	61	770.67	2072.11	1971-2003
K0704510	The Toranche river at St Cyr	62.3	752.63	2076.68	1977-2003
K0813020	The Aix river at St Germain Laval	193	729.48	2093.71	1973-2002
K0943010	The Rhins river at Amplepuis	114	754.52	2111.10	1973-2003
K0974010	The Gand river at Neaux	85	743.45	2107.75	1972-2003
K1004510	The Rhodon river at Perreux	32	738.40	2116.64	1973-2003
U4505010	The Ardieres river at Beaujeu	54.5	773.67	2130.75	1969-2003
U4624010	The Azergues river at Chatillon	336	779.07	2099.72	1970-2003
U4635010	The Brevenne river at Sain Bel	219	775.90	2092.57	1969-2003
U4644010	The Azergues river at Lozanne	792	782.56	2096.09	1981-2003
V3015010	The Yzeron river at Craponne	48	785.47	2084.50	1969-2003
V3114010	The Gier river at Rive de Gier	319	780.54	2062.67	1981-2003
V3315010	The Valencize river at Chavanay	36	786.54	2048.60	1978-2003

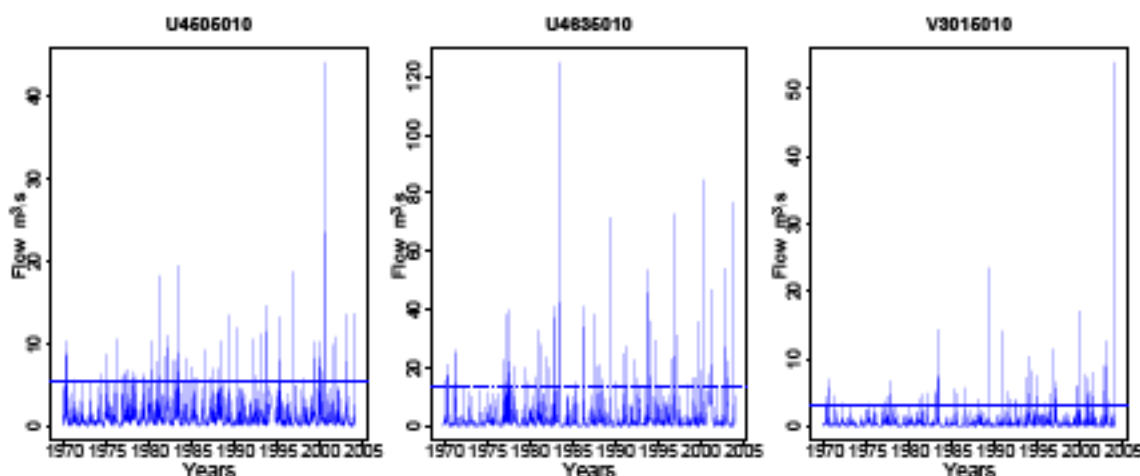


Fig. 2. Times series for sites U4505010, U4635010 and V3015010 and thresholds associated

keep the same Index Flood throughout the case study to compare approaches on the same basis.

### 5 Elicitation of the prior distribution

To estimate the target site Index Flood, the most popular way is to develop an empirical formula that relates the flow statistic to geomorphological, land-use and climatic descriptors. This relationship is usually established by multivariate regression procedures. In our case study, we consider a simple model for which only one explanatory variable is introduced in the regression analysis: the drainage area. The power form model is adopted:

$$C = aA^b \quad (14)$$

where  $A$  is the area of the catchment. Parameters  $a$  and  $b$  are computed through ordinary least square procedures on logarithmically transformed data.

However, more sophisticated models could be carried out. Nevertheless, for our case study, observations demonstrate that Eq. (14) is a good parametrisation for estimating the Index Flood. Fig. 3 and 4 illustrate the efficiency of the regressive and prior model for site

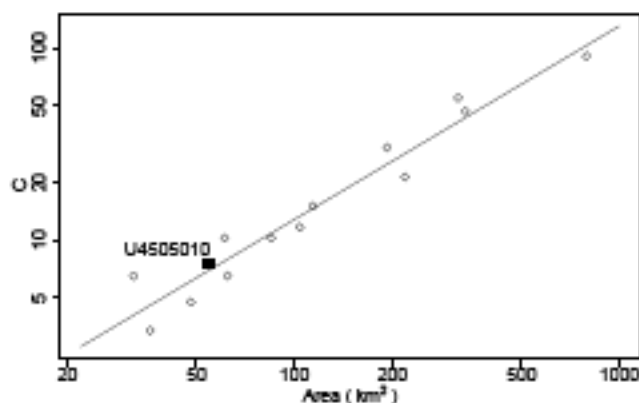


Fig. 4. Regression on the basin area for estimating the at-site index flow for station U4505010

U4505010 for which:

$$\hat{C} = 0.12A^{1.01} \quad (R^2 = 0.94) \quad (15)$$

Regional information was incorporated in the prior distribution through the Index Flood model. Moreover, uncertainties in the prior distribution were incorporated. Thus, the prior information is, on one hand, not too



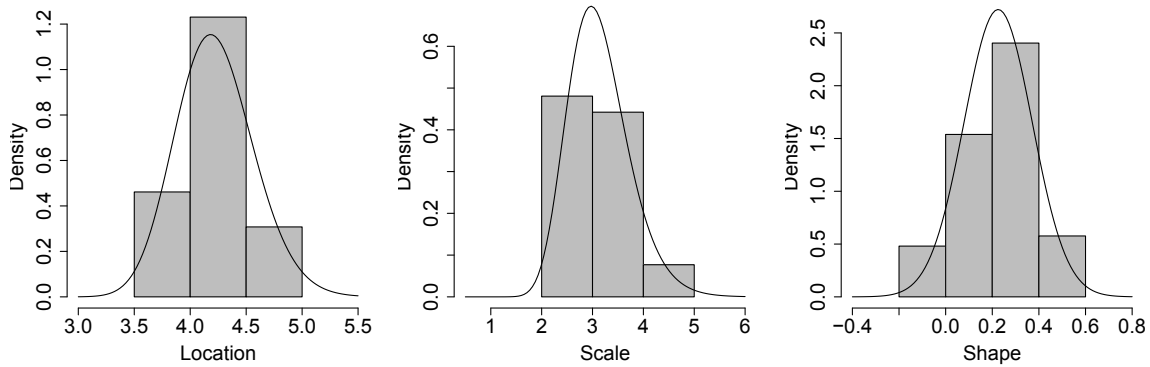


Fig. 3. Histograms of pseudo target site estimates of location, scale and shape parameters for site U4505010

falsely accurate and on the other hand, informative enough because of the supposed homogeneity of stations.

## 6 Performance of the Bayesian model on a homogeneous region

When making classical inference on small samples, the uncertainties may be too large. If an extremal event or too many “regular events” in this short record period are present, estimation can be impacted. It could lead to a dramatic overestimation or underestimation of quantiles corresponding to different return levels. A perfect model is expected : a) to perform well enough even with small samples; b) to be robust enough when an extreme event occurs in the sample; c) to be robust enough when too many “regular” events occur in the sample.

In this section, three different models will be applied. For this purpose, the three stations U4505010, U4635010 and V3015010 were selected to assess the robustness and efficiency of the local, regional and Bayesian regional models. These three different approaches correspond to : a) local: fit the GP distribution to the peaks over threshold data with the Maximum Likelihood Estimator (MLE), Unbiased Probability Weighted Moments (PWU) and the Biased Probability Weighted Moments (PWB); b) regional (REG): fit a regional GP distribution as described in section 2 and obtain the target site distribution; c) regional Bayesian (BAY): elicit the prior density from regional information, then compute the posterior density through MCMC techniques. As an illustration of MCMC output, Fig. 5 displays the prior and posterior marginal densities for the GP parameters of the proposed model. Marginal posterior distribution obtained from an uninformative prior model are also displayed. That is with the same prior model but with a large variance - i.e.  $d_i = 1000, i = 1 \dots 3$ .

Fig. 5 shows the relevance of regional information as the proposed prior model is clearly more accurate than an analysis directly from data. Moreover, for the proposed model and even with only 5 years record length, marginal posterior densities are more accurate than marginal prior densities - except for the shape parameter. Thus, combination of regional and target site information at two different stages is worthwhile, even when only few

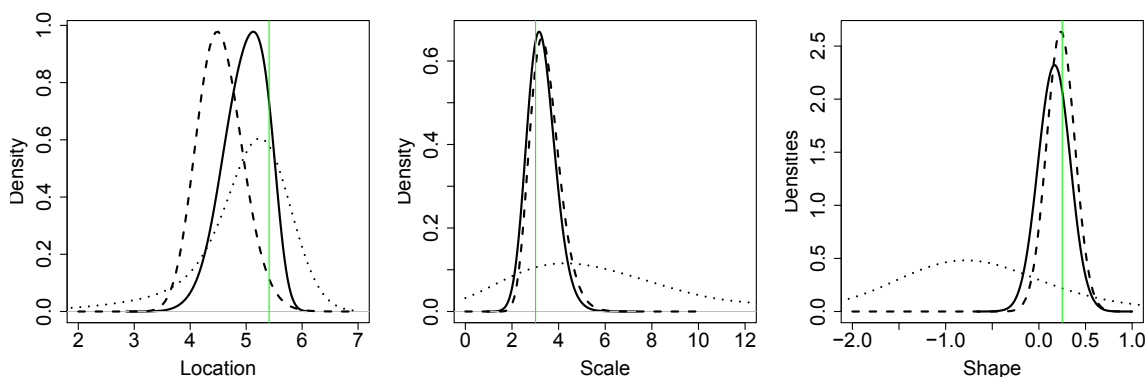
data are available. Location parameter is a special case as the modes of both marginal prior and posterior densities seem to be significantly dissimilar.

As the main goal of this work is to compare models on small samples, efficiency will be evaluated on sub-samples from the original data. Local Maximum Likelihood Estimation on the whole sample will be used as benchmark to assess the performance of each model. This particular case will be denoted THEO in the following sections. The choice of MLE estimate as a benchmark value is reasonable because of its theoretical motivation and asymptotic efficiency. Moreover, the MLE approach allows the calculation of profile confidence intervals. This is a key point as these profile confidence intervals are often more accurate than those based on the Delta Method and Fisher Information (Coles, 2001).

Furthermore, as interpretation on quantile estimates is more natural than for distribution parameter estimates, the analysis will focus on quantiles corresponding to return period 2, 5, 10 and 20 years. Benchmark values for these quantiles - and their associated 90% profile likelihood confidence intervals are detailed in Tab. 2. Benchmark values with return periods greater than 20 years will be considered unreliable - as uncertainties on these quantiles are too large with only 37 years of record.

Moreover, for such return periods, benchmark values are quite equivalent to those obtained with PWM estimates - with a mean bias of 0.89%. So, performance of each model is not too much impacted by the choice of the MLE estimator for benchmark values.

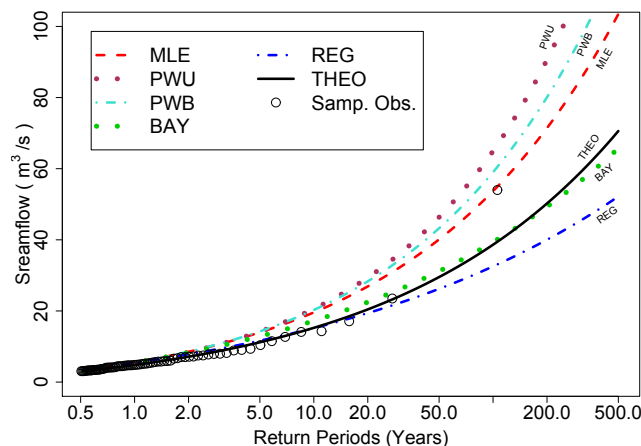
Different frequency curves for site V3015010 with only the last 15 years recording are displayed in Fig. 6. Let us focus on the largest observation. Return period related to this event is very high for the REG approach. All other models lead to significantly lower return periods. This flood event is extreme at regional scale but not anymore in a local context. This underestimation is due to the misuse of the target site sample to establish the regional distribution. On the other side, the regional Bayesian model performs well for all return periods. Indeed, Fig. 6 indicates that the return level curve is very similar to benchmark one. This is quite logical as it adds up the advantage of using efficiently the target site sample and a good “suspicion” on the global behaviour of the



**Fig. 5.** Proposed prior (dashed line), proposed posterior (solid line) and posterior from an uninformative prior (dotted line) marginal densities for GP parameters. Site U4505010 with 5 years record length. Vertical lines denotes benchmark values.

**Table 2.** Benchmark values for 2, 5, 10 and 20 years quantiles and the associated 90% profile likelihood confidence intervals in bracket

Station	$Q_2$	$Q_5$	$Q_{10}$	$Q_{20}$
U4505010	10.8(10.1, 11.7)	15.3 (13.9, 17.4)	19.5(17.2, 23.4)	24.4 (20.6, 31.5)
U4635010	33.0(30.0, 36.5)	52.2 (45.5, 62.5)	72.2(60.2, 95.4)	98.9(69.2, 200.5)
V3015010	7.5 ( 6.9, 8.3)	11.7 (10.4, 13.7)	15.9(13.6, 19.9)	21.3 (17.3, 28.8)



**Fig. 6.** Comparison of frequency curves for site V3015010 with 15 last years recording

flood peak distribution thanks to the so-called prior distribution. Local approaches suggest a very heavy tail as the extremal event of year 2004 (see Fig. 2, right panel) was in the last sequence of 15 years of records.

As one of the main goals of a RFFA procedure is to deal with small samples, the target site sample was truncated to obtain shorten periods of records of  $m$  years,  $m \in \{5, 10, 15, 20, 25, 30, 37\}$ . Robustness and efficiency of the methods to converge to the parameters of the target site distribution are measured. For this purpose, quantile estimates corresponding to return period 2, 5, 10, 20 years - corresponding to non-exceedance probabilities 0.75, 0.9, 0.95 and 0.975 respectively - are pointed. The evolution of quantile estimates as a function of the record length period is presented in Fig. 7. The figure is achieved considering only the first  $m$  years ; that is, for

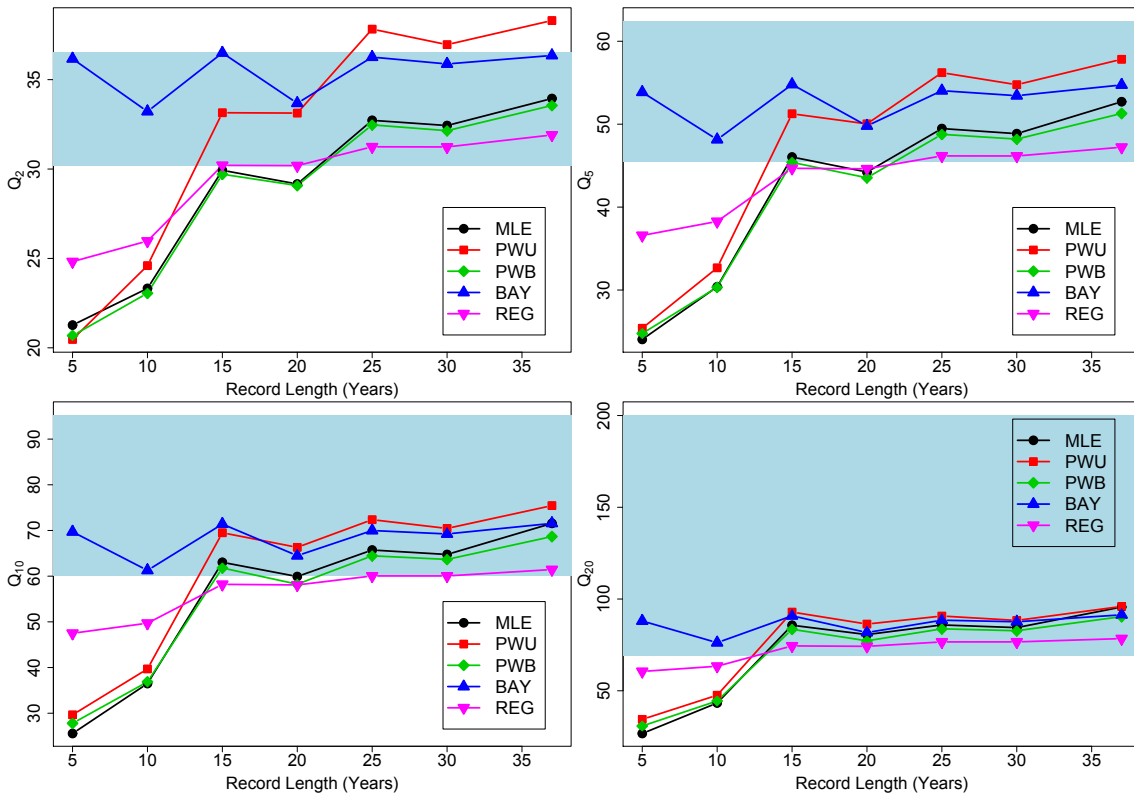
example, estimates related to the 5-year record length corresponds to the period 1969 – 1973.

Because of the extreme event observed in 1983 (see Fig. 2, middle panel), systematic underestimation of benchmark values for local and REG approaches can be noticed. This result shows that: on one hand, for small samples classical inference like MLE, PWB and PWU are too responsive if too many “regular” events occurred. On the other hand, for the Index Flood model, underestimation of quantiles is related to the underestimation of the scale factor  $C^{(S)}$  in Eq. (3) because of these “regular” events. Only the Bayesian model performs well enough even with record lengths lower than 15 years.

A monotonic increase of the design flood estimates with the sample length can be noticed in Fig. 7. This behaviour is easily explained by Fig. 2, middle panel. Indeed, only the last part of the time series shows really extreme events. As the record length increases, much more extreme events occur leading to higher estimates. The Bayesian approach is the only one which does not really present this monotonic behaviour.

Moreover, the Bayesian approach is by far the most robust and accurate model as, on the whole range of record length, and for all benchmark values, estimation lies in the 90% profile likelihood confidence interval. This is not true with any other model. The advantage of incorporating regional information within a Bayesian framework is certainly to define a “restricted space” where distribution parameters belong to. Thus, the impact of a very extremal event - or conversely too many low-level events - should be regarded as an extreme event related to this “restricted space”.

The gain of accuracy in the target site from using regional information is clearly established in section 6 (Fig. 6 and 7). The Bayesian approach seems to be robust even with small samples while being accurate with



**Fig. 7.** Evolution of  $Q_2, Q_5, Q_{10}, Q_{20}$  estimates as the size increases for the site U4635010 and 90% profile likelihood confidence interval for the benchmark values - light blue area

larger sample. The poor performance of the REG model is related to a bad selection of sites within the “homogeneous” region being considered and estimates may be more accurate if “better” regions were considered. Unfortunately, building up such region is difficult because of the purely deterministic relation (3). As the Bayesian approach relaxes the REG model, the search for more homogeneous regions could be ineffective. The goal of the next section is to measure the potential gain, for the Bayesian model, against homogeneity property.

### 7 Effect of heterogeneity degree on quantile estimation

As indicated in the previous section, we focus now on the impact of the level of homogeneousness of the region. For this purpose, we consider four different regions - denoted  $He^+, He, Ho$  and  $Ho^+$  - which correspond to increasingly homogeneous regions according to the test of Hosking and Wallis (1997). The  $Ho$  region corresponds to the region analysed in the previous section and described in Tab. 1. All regions have 14 site except for the most homogeneous one  $Ho^+$  which contains only 8 stations.  $He$  and  $He^+$  regions are derived from  $Ho$ . One to five sites are withdrawn and replaced by other stations to obtain larger heterogeneity measure. The  $Ho^+$  region is a sub-region of  $Ho$ . Heterogeneity statistics for these regions are summarised in Tab. 3.

**Table 3.** Heterogeneity statistics for the four region considered - statistics in bracket are obtained with the scale factor taken to be the 1-year quantile corresponding to non-exceedance probability 0.5

Region	$He^+$	$He$	$Ho$	$Ho^+$
$H_1$	7.11 (6.83)	1.35 (1.37)	0.17 (0.08)	-0.60 (-0.67)
$H_2$	3.46 (3.38)	1.00 (1.03)	0.41 (0.33)	-1.28 (-1.31)
$H_3$	1.40 (1.45)	0.30 (0.28)	-0.09 (-0.14)	-1.14 (-1.18)

To evaluate the influence of homogeneousness level of a region on quantile estimation, models are assessed using two performance criteria : the Normalised Bias (NBIAS) and the Normalised Root Mean Squared Error (NRMSE). These indices are defined as follows :

$$NBIAS = \frac{1}{k} \sum_{i=1}^k \frac{\hat{Q}_i - Q}{Q} \quad (16)$$

$$NRMSE = \sqrt{\frac{1}{k} \sum_{i=1}^k \left( \frac{\hat{Q}_i - Q}{Q} \right)^2} \quad (17)$$

where  $k$  is the number of estimates of  $Q$  and  $\hat{Q}_i$  is the  $i$ -th estimate of the benchmark value  $Q$ . To compute these two indices, we fit all models on all trimmed periods of size  $m$  years -  $m \in \{5, 10, 15, 20, 25, 30\}$ . Moreover, the overall performance of each model is evaluated using a rank score. This technique was already used to compare different models in Shu and Burn (2004).

To calculate the rank score, the  $p$  models are ordered thanks to their performance indices - 1 corresponding to the best model and  $p$  to the worst. For each model, the scores for the different criteria are summed to obtain the overall rank score  $R_o$  for the model. For convenience, the overall rank score  $R_o$  is standardised in such a way that it lies within the interval  $[0, 1]$ :

$$R_s = \frac{pq - R_o}{pq - q} \quad (18)$$

where  $p$  is the number of models being considered, and  $q$  the number of indices. A standardised rank score close to 1 - resp. 0 - is associated to a model with a good - resp. poor - performance.

Three quantiles are of particular interest  $Q_5, Q_{10}$  and  $Q_{20}$  - i.e. associated to probability of non-exceedance 0.9, 0.95 and 0.975 respectively. *NRMSE*, *NBIAS* and the standardised rank score for station U4635010 and a record length of 5 years are illustrated in Tab. 4. Notations for different models in this table consist of one lowercase letter referring to the Bayesian approach  $b$  or Regional Index Flood  $r$  and the denomination of the homogeneity degree of the region. Only the *MLE* model does not use these notations as it is completely independent of the homogeneity level.

Results from Tab. 4 demonstrate that the Bayesian model performs quite well independently of the region being considered. However, this model seems to perform even better when applied to a "acceptably homogeneous" or "probably heterogeneous" region. For the  $Ho^+$  region, the Bayesian approach performs poorly. This may be explained by the fact that the prior distribution is too informative and probably not consistent with the target site sample. This comment is yet not discrepant with the good overall performance of the REG model on this region. Indeed, as the prior distribution is elicited using equations (6)–(8), and the scale factor  $C^{(j)}$  is estimated without any use of the target site sample, this can lead to a misleading prior distribution while the REG model performs well. The bad estimation of the scale factor is less important with a more heterogeneous region as the prior information is less informative, thus the Bayesian model performance is not highly impacted.

On the other side, the overall rank score of the REG model increases with the homogeneity degree of the region. Yet, the overall rank score for the REG model never exceeds the value of 0.6 - reached for the  $Ho^+$  region. This value remains much lower than the best rank score for the Bayesian model - i.e. 0.85. These results corroborate the superiority of the Bayesian approach.

From Tab. 4, two conclusions can be established. On one hand, for small samples, the Bayesian approach is the most competitive model. On the other hand, results seem to indicate that there is no need to keep increasing the homogeneity of the region as it increases the risk of being too confident in the "homogeneous region" without increasing significantly the efficiency of the model.

These results are in line with similar results obtained for stations U4505010 and V3015010, except for the bad

behaviour of the Bayesian model on the  $Ho^+$  region. Indeed, for the other stations, the Bayesian model remains more efficient than the REG model within the  $Ho^+$  region. However, its overall rank score remains stable through out the different region -  $He, Ho$  and  $Ho^+$ . The "risk" to deal with too much homogeneous region - as  $Ho^+$  - is also corroborated as the overall rank score for the Index Flood model for station U4505010 decreases dramatically until 0.06. Thus, the Index Flood for the  $Ho^+$  region performs quite well for stations U4635010 and V3015010, while very surprisingly badly with U4505010.

In Fig. 8, the evolution of the overall rank score as a function of the record length is illustrated for station V3015010. The left panel corresponds to the REG part, while the right one stands for the Bayesian approach. In both panel, the *MLE* score is also presented. Fig. 8 indicates that the evolution of the overall rank score is more stable for regional models - that is REG and Bayesian models - than for the *MLE*. Furthermore, the benefit of increasing the homogeneity degree of the region is more relevant for the REG model than for the Bayesian model. Nevertheless, the worst Bayesian rank score is always quite close to the best REG rank score. This seems to indicate the superiority of the Bayesian approach. This last point is corroborated with the results corresponding to stations U4505010 and U4635010 except for the  $bHo^+$  model for station U4635010 because of the bad estimation of the scale factor  $C^{(j)}$  - as denoted earlier. The effect of bad estimation of the target site Index Flood on prior and thereby on posterior distributions is depicted in Fig. 9.

From Fig. 9, it is overwhelming that the prior model is not appropriate - particularly for the location parameter. Prior for the shape parameter is not too false as it does not depend on the target site Index Flood estimate.

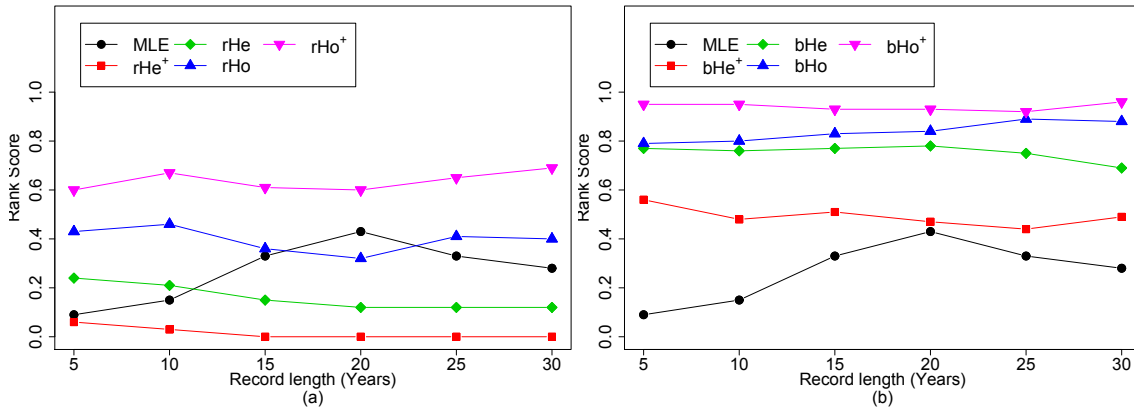
As the record length increases, the *MLE* model becomes more and more efficient. In particular, for record lengths greater than 15 years, it is more effective than  $rHe^+, rHe$  and  $rHo$  models. On one hand, for record lengths smaller than 15 years, *MLE* is always less efficient than Bayesian approaches and even significantly for  $bHe, bHo$  and  $bHo^+$  models. This is quite logical as Bayesian estimation can be looked at as a restrictive maximum likelihood estimator - restriction being defined by the prior distribution. So, under the hypothesis that the prior distribution is well-defined, the "restrictive estimator" is unbiased and has a smaller variance. On the other hand, for record lengths greater than 15 years, *MLE, bHe* and *bHo* seems to be equivalent.

## 8 Conclusion

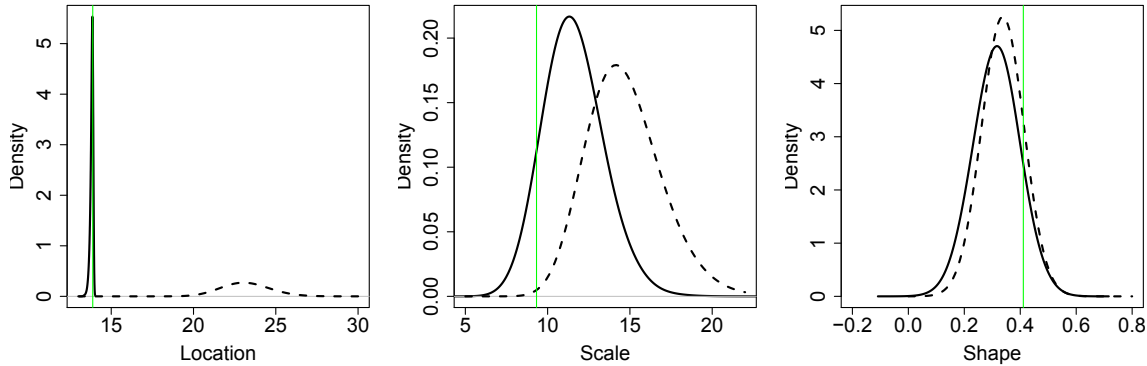
A framework to perform a regional Bayesian frequency analysis for partially gauged stations is presented. The proposed model has the advantage of being less restrictive than the most widely used regional model, that is the Index Flood. Several case studies from French sites were analysed to illustrate the superiority of the Bayesian approach in comparison to the traditional Index Flood and to local approaches. The influence of the

**Table 4.** Estimation of *NRMSE* and *NBIAS* for station U4635010 with a record length of 5 years

Model	<i>NRMSE</i>			<i>NBIAS</i>			Rank Score
	$Q_5$	$Q_{10}$	$Q_{20}$	$Q_5$	$Q_{10}$	$Q_{20}$	
<i>MLE</i>	0.33	0.34	0.39	0.01	-0.09	-0.18	0.26
<i>bHe</i> <sup>+</sup>	0.16	0.13	0.18	0.09	-0.02	-0.13	0.65
<i>rHe</i> <sup>+</sup>	0.27	0.30	0.37	-0.12	-0.22	-0.31	0.18
<i>bHe</i>	0.10	0.07	0.11	0.08	0.00	-0.09	0.85
<i>rHe</i>	0.27	0.26	0.28	-0.03	-0.10	-0.17	0.43
<i>bHo</i>	0.14	0.09	0.08	0.12	0.05	-0.02	0.76
<i>rHo</i>	0.27	0.26	0.27	0.01	-0.06	-0.12	0.58
<i>bHo</i> <sup>+</sup>	0.29	0.28	0.25	0.29	0.27	0.25	0.19
<i>rHo</i> <sup>+</sup>	0.28	0.27	0.26	0.02	-0.01	-0.04	0.60



**Fig. 8.** Score evolution as a function of record length for Station V3015010. (a) REG scores, (b) BAY scores



**Fig. 9.** Effect of bad estimation of target site Index Flood on marginal prior and posterior densities. Site U4635010 with 10 years record length.

homogeneous level of the pooling group on quantile estimates was also considered. Results demonstrate that working with quite large and homogeneous regions rather than small and strongly homogeneous regions is more efficient. Further work can focus on the regional estimation of other characteristics of the flood hydrograph. For instance, a regional Bayesian model can focus on Flood Duration Frequency.

All statistical analysis was carried out in the R Development Core Team (2005) framework. For this purpose, two packages were contributed to this software under the framework of the present research work. These two packages integrate the tools that were developed to carry out the modelling effort presented in this paper. The first

one **POT** performs statistical inference on Peaks Over Thresholds, while the second one, **RFA**, contains several tools to carry out a Regional Frequency Analysis. These two packages are available, free of charge, at the web site <http://www.R-project.org>, section CRAN, Packages.

*Acknowledgements.* The authors wish to thank the DIREN Rhne-Alpes for providing data. The authors are also very grateful to the two referees for their constructive remarks which improve the document.

## Appendix A. Properties of the Index Flood on GP parameters

We provide in this appendix the proof for the following theorem:

**Theorem 1.** *Let  $X$  be a random variable GP distributed. So  $X$  has the Cumulative Distribution Function defined by:*

$$F(x) = 1 - \left(1 + \frac{\xi(x - \mu)}{\sigma}\right)^{-1/\xi}$$

*Let  $Y = CX$  where  $C \in \mathbb{R}_*^+$ . Then,  $Y$  is also GP distributed with parameters  $(C\mu, C\sigma, \xi)$ .*

**Proof.** Let  $X$  be a *r.v.* GP distributed with parameters  $(\mu, \sigma, \xi)$  and  $Y = CX$  where  $C \in \mathbb{R}_*^+$ . Then:

$$\begin{aligned} \Pr[Y \leq y] &= \Pr\left[X \leq \frac{y}{C}\right] \\ &= 1 - \left(1 + \frac{\xi\left(\frac{y}{C} - \mu\right)}{\sigma}\right)^{-1/\xi} \\ &= 1 - \left(1 + \frac{\xi(y - \mu C)}{\sigma C}\right)^{-1/\xi} \end{aligned}$$

So,  $Y$  is also GP distributed with parameters  $(\mu C, \sigma C, \xi)$ . The proof for the GEV case can be established in the same way.

## References

- M Acreman and S Wiltshire. The regions are dead: Long live the regions. In *FRIENDS*, volume 187, pages 175–188, Washington, DC, 1989. International Association of Hydrological Sciences.
- A. Balkema and L. de Haan. Residual life time at great age. *Annals of Probability*, 2:792–804, 1974.
- D.H. Burn. Evaluation of regional flood frequency analysis with a region of influence approach. *Water Resources Research*, 26(10):2257–2265, 1990.
- S. Coles. *An Introduction to Statistical Modelling of Extreme Values*. Springer Series in Statistics. Springer Series in Statistics, London, 2001.
- S. Coles and L. Pericchi. Anticipating catastrophes through extreme value modelling. *J Royal Statistical Soc C*, 52(4):405–416, 2003.
- S. Coles and J. Tawn. A bayesian analysis of extreme rainfall data. *Journal of the Royal Statistical Society. Series C: Applied Statistics*, 45(4):463–478, 1996.
- M. Crowder. Bayesian priors based on a parameter transformation using the distribution function. *Annals of the Institute of Statistical Mathematics*, 44(3):405–416, 1992.
- T. Dalrymple. Flood frequency analysis. *U.S. Geol. Surv. Water Supply Pap.*, 1543 A:–, 1960.
- P. Embrechts, C. Klüppelberg, and T. Mikosch. *Modelling Extremal Events for Insurance and Finance*. Springer, New York, 1997.
- H. D. Fill and J. R. Stedinger. Using regional regression within index flood procedures and an empirical bayesian estimator. *Journal of Hydrology*, 210(1-4):128–145, 1998.
- R.A. Fisher and L.H. Tippett. Limiting forms of the frequency distribution of the largest or smallest member of a sample. In *Cambridge Phil. Soc.*, volume 24, 1928.
- S. Gabriele and N. Arnell. A hierarchical approach to regional flood frequency analysis. *Water Resources Research*, 27(6):1281–1289, 1991.
- GREHYS. Inter-comparison of regional flood frequency procedures for canadian rivers. *Journal of Hydrology*, 186(1-4):85–103, 1996.
- V. K. Gupta, O. J. Mesa, and D. R. Dawdy. Multiscaling theory of flood peaks: Regional quantile analysis. *Water Resources Research*, 30(12):3405–3421, 1994.
- M. J. Hall, A. W. Minns, and A. K. M. Ashrafuzzaman. The application of data mining techniques for the regionalisation of hydrological variables. *Hydrology and Earth System Sciences*, 6(4):685–694, 2002.
- J. R. M. Hosking and J. R. Wallis. Some statistics useful in regional frequency analysis. *Water Resources Research*, 29(2):271–281, 1993.
- J. R. M. Hosking and J. R. Wallis. *Regional Frequency Analysis*. Cambridge University Press, 1997.
- R.W. Katz, M.B. Parlange, and P. Naveau. Statistics of extremes in hydrology. *Advances in Water Resources*, 25(8-12):1287–1304, 2002.
- M. Lang, T. B. M. J. Ouarda, and B. Bobée. Towards operational guidelines for over-threshold modeling. *Journal of Hydrology*, 225(3-4):103–117, 1999.
- H. Madsen and D. Rosbjerg. Generalized least squares and empirical bayes estimation in regional partial duration series index-flood modeling. *Water Resources Research*, 33(4):771–781, 1997.
- P.J. Northrop. Likelihood-based approaches to flood frequency estimation. *Journal of Hydrology*, 292(1-4):96–113, 2004.
- E. Parent and J. Bernier. Encoding prior experts judgments to improve risk analysis of extreme hydrological events via pot modeling. *Journal of Hydrology*, 283(1-4):1–18, 2003.
- J. III Pickands. Statistical inference using extreme order statistics. *Annals of Statistics*, 3:119–131, 1975.
- R. Development Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2005. URL <http://www.R-project.org>. ISBN 3-900051-07-0.
- J. S. Robinson and M. Sivapalan. An investigation into the physical causes of scaling and heterogeneity of regional flood frequency. *Water Resources Research*, 33(5):1045–1059, 1997.
- A.J. Robson and D.W. Reed. *Flood Estimation Handbook*, volume 3. Institute of Hydrology, Wallingford, 1999.
- C. Shu and D.H. Burn. Artificial neural network ensembles and their application in pooled flood frequency analysis. *Water Resources Research*, 40(9), 2004.
- J.G. Wasson, A. Chandesris, H. Pella, and L. Blanc. Hydroecoregions : a functional approach of river typology for the european water framework directive. *Ingénieries - EAT - in french -*, 40(3-10), 2004.

### 5.4 Discussion

Ce modèle régional Bayésien a été validé sur une région supposée homogène de 14 stations françaises. Trois sites particuliers ont été analysés plus particulièrement. Une comparaison de la performance du modèle introduit face au modèle de l'indice de crue mais aussi aux estimateurs purement locaux traditionnellement utilisés en hydrologie statistique a été effectuée.

Il a été montré que le modèle proposé était de loin le plus robuste face à la variabilité d'échantillonnage au site cible mais aussi de la région toute entière. De plus, le modèle proposé a montré un taux de convergence nettement supérieur aux autres estimateurs. Finalement, cet estimateur semble être parfaitement adapté pour l'estimation des quantiles de crue en site partiellement jaugé - *i.e.* chronique de débits inférieure à 15 ans.

L'étude a également montré qu'il semble préférable pour l'estimation régionale des débits de référence de travailler avec des régions dites *homogènes* ou *probablement hétérogènes* selon la terminologie d'Hosking et Wallis [1997] plutôt que de rechercher l'homogénéité à tout prix. En effet, cette dernière option semble plus risquée et conduit parfois à des erreurs d'estimation très fortes. L'avantage de l'estimateur régional Bayésien est qu'il s'est révélé moins sensible face au degré d'homogénéité de la région considérée.

Bien qu'ayant montré des propriétés théoriques et empiriques attrayantes, le modèle que nous avons proposé admet des limites.

Premièrement, il n'a été validé que pour des quantiles ayant une période de retour inférieure à 20 ans. Il est donc pour le moment impossible de porter un quelconque jugement pour de grandes périodes de retour.

Deuxièmement, malgré une relaxation des hypothèses d'invariance au sein de la région, ce modèle impose des contraintes peut-être encore trop rigides. En particulier, dans sa forme présentée ici, les sites de la région homogène sont supposés mutuellement indépendants. Cette hypothèse est bien entendu forte et peut-être infondée pour certaines applications compte tenu de la structure des pluies et des dépendances amont/aval. Même lorsque les régions homogènes ne sont pas construites sur des critères géographiques, il peut subsister quelques sites présentant une forte dépendance. C'est donc un point sur lequel il faudra travailler car cela implique une définition de la loi *a priori* plus précise qu'elle ne devrait être en réalité, impactant de ce fait les intervalles de crédibilité. Pour notre cas d'étude toutefois, cette hypothèse n'était pas trop restrictive puisque les bassins considérés étaient deux à deux disjoints.

Sur un autre plan, les paramètres de la loi GPD sont supposés indépendants dans la définition de notre loi *a priori*. Si pour le paramètre de position, cela semble être une hypothèse raisonnable ; ce n'est plus toujours le cas pour les paramètres d'échelle  $\sigma$  et de forme  $\xi$  qui semblent présenter une dépendance négative. La Figure 5.2 illustre bien cette dépendance puisque la sous-estimation du paramètre d'échelle  $\sigma^{(S)}$  est compensée par une surestimation du paramètre de forme  $\xi^{(S)}$ . A l'avenir, il faudrait donc considérer un modèle plus général permettant la présence de dépendances entre les paramètres  $\sigma$  et  $\xi$ . En particulier, cela aurait pour avantage de préciser un peu plus encore la loi *a priori* et avec les estimations.

# 6

## Utilisation des Sauts Réversibles pour une Prédiction Plus Robuste des Extrêmes

DANS cette section, nous allons présenter un nouveau modèle Bayésien régional pour l'estimation des débits de référence. Ce nouveau modèle est en fait une prolongation du modèle précédent et est établi sur les mêmes bases théoriques. Toutefois, avant de rentrer plus en détail dans le coeur du modèle, nous allons présenter les raisons qui nous ont poussé à cette amélioration.

### 6.1 La Nécessité d'un Nouveau Modèle Régional

Dans la section précédente (§ 5.4), nous avons discuté le premier modèle régional Bayésien (**BAY**) proposé durant ce travail de recherche. En particulier, nous avons précisé que les performances obtenues pour ce modèle n'étaient établies que pour des quantiles de période de retour inférieure à 20 ans.

En effet, les chroniques dont nous disposions pour l'étude de performance ne dépassaient pas 37 ans. Dès lors, il paraissait hasardeux de réaliser une étude sur des quantiles dont les valeurs de référence étaient trop incertaines. Afin d'établir des conclusions fiables sur des quantiles plus extrêmes, nous avons étudié le comportement de l'estimateur précédent sur des données simulées et de ce fait entièrement contrôlées.

L'algorithme de simulation proposé devait donc être un algorithme de simulation de données régionales pouvant contrôler le degré d'homogénéité de la région simulée. Les détails algorithmiques et théoriques sont reportés dans le corps de l'article. Il est important de noter que cet algorithme est basé sur le concept de l'indice de crue. Ainsi, par construction, l'estimateur de l'indice de crue est légèrement avantage par rapport aux autres estimateurs lorsqu'il est appliqué aux régions ainsi générées.

Pour étudier la performance de l'estimateur BAY sur des données simulées, trois statistiques mesurant la performance de chaque estimateur ont été considérées : (a) le biais normalisé (**nbias**), (b) l'écart type (**sd**) et (c) l'erreur quadratique moyenne normalisée (**nrmse**). Ces statistiques sont



## 6. Utilisation des Sauts Réversibles pour une Prédiction Plus Robuste des Extrêmes

**Tab. 6.1:** Performance de chaque estimateur pour l'estimation des quantiles  $Q_2$ ,  $Q_{10}$  et  $Q_{20}$ . Nombre de régions simulées : 1000. Taille d'échantillon au site cible : 20.

Modèle	<i>nbias</i>			<i>sd</i>			<i>nrmse</i>		
	$Q_2$	$Q_{10}$	$Q_{20}$	$Q_2$	$Q_{10}$	$Q_{20}$	$Q_2$	$Q_{10}$	$Q_{20}$
<i>MLE</i>	0.049	-0.003	0.002	12.66	34.70	63.26	0.176	0.321	0.482
<i>PWU</i>	0.018	-0.013	-0.010	11.03	31.53	49.31	0.158	0.263	0.345
<i>PWB</i>	0.033	-0.020	-0.031	11.46	31.70	48.97	0.164	0.270	0.350
<i>MOM</i>	0.068	-0.033	-0.079	13.20	32.07	45.91	0.199	0.291	0.344
<i>BAY</i>	0.008	-0.012	-0.020	3.42	14.34	23.49	0.046	0.112	0.148
<i>IFL</i>	0.028	0.017	0.009	9.11	18.02	23.20	0.132	0.134	0.136

**Tab. 6.2:** Performance de chaque estimateur pour l'estimation des quantiles  $Q_{50}$ ,  $Q_{100}$  et  $Q_{1000}$ . Nombre de régions simulées : 1000. Taille d'échantillon au site cible : 20.

Modèle	<i>nbias</i>			<i>sd</i>			<i>nrmse</i>		
	$Q_{50}$	$Q_{100}$	$Q_{1000}$	$Q_{50}$	$Q_{100}$	$Q_{1000}$	$Q_{50}$	$Q_{100}$	$Q_{1000}$
<i>MLE</i>	0.05	0.14	1.13	158.1	333.5	4585.8	0.86	1.38	8.21
<i>PWU</i>	0.01	0.05	0.33	90.8	145.7	741.9	0.50	0.66	1.78
<i>PWB</i>	-0.03	-0.01	0.16	88.6	140.1	688.8	0.49	0.64	1.65
<i>MOM</i>	-0.13	-0.17	-0.27	73.6	104.1	303.9	0.42	0.48	0.70
<i>BAY</i>	-0.03	-0.03	-0.04	42.0	62.6	203.7	0.20	0.24	0.41
<i>IFL</i>	-0.00	-0.01	-0.05	32.1	41.2	96.2	0.14	0.15	0.19

définies par :

$$nbias = \frac{1}{k} \sum_{i=1}^k \frac{\hat{Q}_i - Q_i}{Q_i} \quad (6.1)$$

$$sd = \sqrt{\frac{1}{k-1} \sum_{i=1}^k \left( \frac{\hat{Q}_i - Q_i}{Q_i} - nbias \right)^2} \quad (6.2)$$

$$nrmse = \frac{1}{k} \sum_{i=1}^k \left( \frac{\hat{Q}_i - Q_i}{Q_i} \right)^2 \quad (6.3)$$

où  $\hat{Q}_i$  est la  $i$ -ème estimation du quantile de référence  $Q_i$  et  $k$  le nombre total de valeurs de référence.

Les Tableaux 6.1 et 6.2 illustrent les performances des estimateurs utilisés fréquemment en hydrologie statistique : maximum de vraisemblance (**MLE**), moments pondérés (**PWU**, **PWB**), moments (**MOM**) et de l'indice de crue (**IFL**). L'estimateur *BAY* est également reporté.

Le Tableau 6.1 confirme les conclusions établies par Ribatet *et al.* [2007b]. En effet, pour des périodes de retour inférieures à 20 ans, le modèle *BAY* est, en termes d'erreur quadratique moyenne, plus performant que tous les autres estimateurs.

En revanche, comme le montre le Tableau 6.2, ce n'est plus le cas dès lors que l'on s'attache à estimer des quantiles plus extrêmes. En effet, nous voyons que l'estimateur *BAY*, bien que toujours plus performant que les estimateurs locaux, est nettement moins performant que l'estimateur *IFL* pour l'estimation de  $Q_T$ ,  $T \geq 50$  ans.

Le Tableau 6.3 représente des statistiques de performance sur l'estimation des paramètres de position  $u$ , d'échelle  $\sigma$  et de forme  $\xi$  de la GPD. Ainsi, nous pouvons constater que la moins bonne performance de l'estimateur *BAY* pour l'estimation des plus forts quantiles est entièrement due à une mauvaise estimation du paramètre de forme  $\xi$ . En particulier, l'estimateur *BAY* admet une

**Tab. 6.3:** Performance de chaque estimateur pour l'estimation des paramètres de la GPD. Nombre de régions simulées : 1000. Taille d'échantillon au site cible : 20.

Modèle	<i>nbias</i>			<i>sd</i>			<i>nrmse</i>		
	<i>u</i>	$\sigma$	$\xi$	<i>u</i>	$\sigma$	$\xi$	<i>u</i>	$\sigma$	$\xi$
<i>MLE</i>	0.042	0.227	-0.626	1.115	11.290	0.369	0.060	0.572	1.552
<i>PWU</i>	0.042	0.076	-0.289	1.115	8.191	0.278	0.060	0.411	1.108
<i>PWB</i>	0.042	0.136	-0.455	1.115	8.723	0.276	0.060	0.431	1.153
<i>MOM</i>	0.042	0.258	-0.759	1.115	10.350	0.216	0.060	0.481	1.124
<i>BAY</i>	0.015	0.026	-0.131	0.757	1.473	0.085	0.038	0.058	0.354
<i>IFL</i>	0.007	0.064	-0.098	3.592	3.652	0.035	0.126	0.155	0.168

*nrmse* plus de deux fois supérieure à celle obtenue par l'*IFL*. De plus, cette trop forte erreur quadratique sur l'estimation de  $\xi$  est principalement en lien avec une trop grande variance d'estimation sur ce paramètre de forme.

Le fait qu'une mauvaise estimation du paramètre de forme  $\xi$  impacte fortement l'estimation des plus forts quantiles n'est pas une surprise. En effet, nous savons que le paramètre de forme  $\xi$  contrôle le comportement de la queue de distribution et donc des plus forts extrêmes. En particulier, s'il est positif (*resp.* négatif), la variable aléatoire est non bornée (*resp.* bornée) supérieurement.

A partir de ces résultats, et bien que, comme il a été dit, l'estimateur *IFL* soit ici artificiellement favorisé, il nous est donc impossible de recommander l'utilisation de l'estimateur *BAY* pour l'estimation des quantiles ayant une période de retour supérieure à 20 ans. Nous allons donc proposer un nouvel estimateur utilisant l'information régionale de manière plus optimale pour l'estimation de  $\xi$ .

## 6.2 Les Apports du Nouveau Modèle

Partant du constat que la variance d'estimation de  $\xi$  est trop pénalisante pour l'estimation des plus forts quantiles à partir du modèle précédent, nous avons eu l'idée de proposer un paramètre de forme régional  $\xi_{\text{Fix}}$ . Cette idée reprend le principe de l'indice de crue pour lequel il n'y a qu'une seule valeur pour  $\xi$  sur toute la région - *cf.* Équation (5.2). Toutefois, pour notre nouveau modèle, nous resterons cohérents avec la relaxation des hypothèses d'invariance sur la région puisque ce paramètre  $\xi_{\text{Fix}}$  ne sera pas **imposé** comme pour l'indice de crue mais **proposé**.

Intuitivement, nous voyons que la « proposition » d'un paramètre de forme  $\xi_{\text{Fix}}$  permet de réduire la variance d'estimation. Finalement, si le candidat  $\xi_{\text{Fix}}$  est pertinent, on peut espérer que le nouveau modèle sera plus performant pour l'estimation des plus forts quantiles, notamment lorsque  $T > 20$  ans.

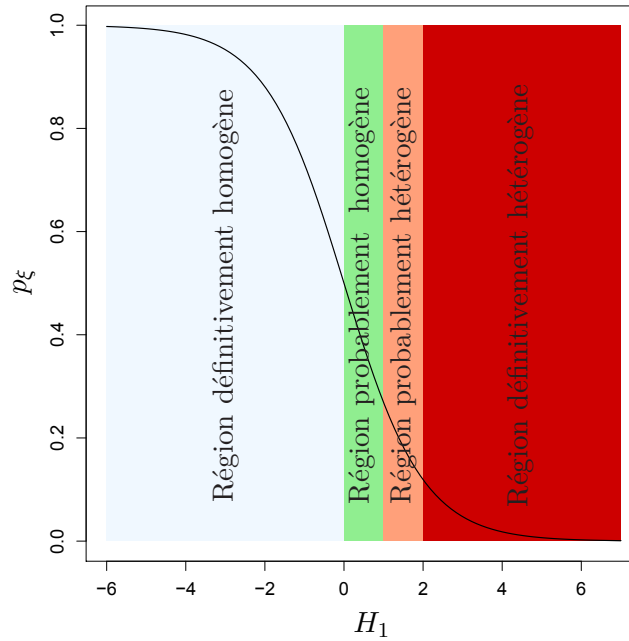
Puisque nous ne voulons aucunement imposer la valeur  $\xi_{\text{Fix}}$ , il est impossible de l'intégrer dans la fonction vraisemblance. Toujours dans un contexte Bayésien, cette intégration ne pourra se faire qu'au sein de la loi *a priori*.

Partant de la loi *a priori* initiale  $\pi_{\text{in}}(\theta)$  introduite pour l'estimateur *BAY*, nous définissons donc une nouvelle loi *a priori*  $\pi(\theta)$  comme suit :

$$\pi(\theta) = \begin{cases} (1 - p_\xi)\pi_{\text{in}}(\theta), & \text{pour } \theta \in \Theta \setminus \Theta_0 \\ p_\xi \pi_{\xi_{\text{Fix}}}(\theta), & \text{pour } \theta \in \Theta_0 \end{cases} \quad (6.4)$$

où  $p_\xi \in [0, 1]$ ,  $\Theta = \{(u, \sigma, \xi) : u, \sigma \in \mathbb{R}_*^+, \xi \in \mathbb{R}\}$ ,  $\Theta_0 = \{(u, \sigma, \xi_{\text{Fix}}) : u, \sigma \in \mathbb{R}_*^+\}$  et

$$\pi_{\xi_{\text{Fix}}}(\theta) = \frac{\pi_{\text{in}}(u, \sigma, \xi_{\text{Fix}})}{\int_{u, \sigma} \pi_{\text{in}}(u, \sigma, \xi_{\text{Fix}}) du d\sigma} \quad (6.5)$$



**Fig. 6.1:** Evolution de l'hyperparamètre  $p_\xi$  en fonction de la statistique  $H_1$ .

pour  $\theta \in \Theta_0$ .

Cette nouvelle loi *a priori*  $\pi(\theta)$  est en fait une mixture de deux lois *a priori* : (a) celle introduite lors du modèle précédent  $\pi_{\text{in}}(\theta)$  et (b) une nouvelle loi *a priori*  $\pi_{\xi_{\text{Fix}}}(\theta)$  où le paramètre de forme est fixé à  $\xi_{\text{Fix}}$ . L'hyperparamètre supplémentaire  $p_\xi$ , quant à lui, contrôle le *degré de mixture*. En particulier, pour  $p_\xi = 0$ ,  $\pi(\theta) = \pi_{\text{in}}(\theta)$  de sorte que nous retombons sur le modèle précédent. Au contraire, si  $p_\xi = 1$ ,  $\pi(\theta) = \pi_{\xi_{\text{Fix}}}(\theta)$  et le paramètre de forme  $\xi$  est alors imposé et égal à  $\xi_{\text{Fix}}$ .

L'hyperparamètre supplémentaire  $p_\xi$  contrôle donc la croyance *a priori* que nous pouvons avoir en  $\xi_{\text{Fix}}$  comme la vraie valeur du paramètre de forme du site cible. Il paraît donc logique de lier la valeur de  $p_\xi$  au degré d'homogénéité de la région, et en particulier à la statistique  $H_1$  d'[Hosking et Wallis \[1997\]](#). Pour notre modèle, nous avons introduit la paramétrisation suivante :

$$p_\xi = \frac{\exp(-H_1)}{1 + \exp(-H_1)} \quad (6.6)$$

La Figure 6.1 représente l'évolution de l'hyperparamètre  $p_\xi$  en fonction du degré d'homogénéité de la région - *cf.* Équation (6.6). Bien entendu, le choix de cette paramétrisation est purement intuitif. Néanmoins, elle respecte certaines propriétés essentielles :

- pour une région parfaitement homogène (*i.e.*  $H_1 \rightarrow -\infty$ ), le paramètre de forme  $\xi$  est supposé connu et égal à  $\xi_{\text{Fix}}$  (*i.e.*  $p_\xi = 0$ ) ;
- pour une région complètement hétérogène (*i.e.*  $H_1 \rightarrow +\infty$ ), le paramètre de forme est laissé libre et défini selon la loi *a priori* initiale  $\pi_{\text{in}}(\theta)$  (*i.e.*  $p_\xi = 1$ ).

Finalement, cette nouvelle loi *a priori* permet donc de réduire la variance d'estimation du paramètre de forme  $\xi$ . Cette réduction sera d'autant plus importante que la région d'étude aura un degré d'homogénéité plus important.

### 6.3 L'Article

## Usefulness of the Reversible Jump Markov Chain Monte Carlo Model in Regional Flood Frequency Analysis

M. Ribatet,<sup>1,2</sup> E. Sauquet,<sup>1</sup> J.-M. Grésillon,<sup>1</sup> and T.B.M.J. Ouarda<sup>2</sup>

**Abstract.** Regional flood frequency analysis is a convenient way to reduce estimation uncertainty when few data are available at the gauging site. In this work, a model that allows a non null probability to a regional fixed shape parameter is presented. This methodology is integrated within a Bayesian framework and uses reversible jump techniques. The performance on stochastic data of this new estimator is compared to two other models: a conventional Bayesian analysis and the index flood approach. Results show that the proposed estimator is absolutely suited to regional estimation when only a few data is available at the target site. Moreover, unlike the index flood estimator, target site index flood error estimation seems to have less impact on Bayesian estimators. Some suggestions about configurations of the pooling groups are also presented to increase the performance of each estimator.

**Keywords:** Regional Frequency Analysis, Extreme Value Theory, Generalized Pareto Distribution, Reversible Jumps, Markov Chain Monte Carlo.

### 1. Introduction

Extreme value theory is now widely applied when modeling block maxima or exceedences over a threshold is of interest. In particular, the Generalized Pareto Distribution (GPD) describes the limiting distribution of normalized excesses of a threshold as the threshold approaches the endpoint of the variable [Pickands, 1973]. The GPD has a distribution function defined by:

$$G(x; \mu, \sigma, \xi) = 1 - \left[ 1 + \frac{\xi(x - \mu)}{\sigma} \right]^{-1/\xi}, \quad x > \mu, 1 + \frac{\xi(x - \mu)}{\sigma} > 0 \quad (1)$$

where  $\sigma > 0$ ,  $\xi \in \mathbb{R}$ ,  $\mu$ ,  $\sigma$  and  $\xi$  are respectively the location, scale and shape parameters.

Thus, when extreme values must be estimated, this approximation is frequently used. Most applications based on this result are related to environmental sciences, as extreme wind speed [Payer and Kuechenhoff, 2004], extreme sea level [Bortot and Coles, 2000; Pandey et al., 2004] or extreme river discharge [Northrop, 2004].

However, one must often deal with small samples and large uncertainties on estimation. Several publications point out the problem of the shape parameter estimation. This parameter is of great interest as it determines the tail behaviour of the distribution. Therefore, many authors analyzed the performance of particular estimators given a specified range for the shape parameter: Rosbjerg et al. [1992] for the method of moments; Coles and Dixon [1999] for the maximum likelihood; Hosking and Wallis [1987] for the probability weighted moments; Judes and Schuany [2004] for the minimum density power divergence estimator; Martins and Stedinger [2000] for a proposed generalized maximum likelihood. However, these results provide the most accurate estimator given the shape parameter; which is never

the case in practice. Therefore, Park [2005] introduced a systematic way of selecting hyper-parameters for his proposed generalized maximum likelihood estimator.

All these approaches only deal with information from the target site sample. However, it is frequent in hydrology to perform a Regional Frequency Analysis (RFA). Traditional RFA consists of two steps: (a) delineation of homogeneous regions i.e. a pooling group of stations with similar behaviour; (b) regional estimation i.e. estimate target site distribution from the regional information.

More recently, Bayesian approaches have been applied with success to incorporate regional information in frequency analysis [Coles and Tawn, 1996; Northrop, 2004; Seidou et al., 2006; Ribatet et al., 2007]. Empirical Bayesian estimators have also been proposed [Kucsera, 1982; Madsen and Rosbjerg, 1997]. One of the advantages of these approaches is to distinguish the at site information from the other sites data in the estimation procedure. This is an important point as, no matter how high the homogeneity level may be, the only data which represents perfectly the target site is obviously the target site one. Thus, the whole information available is used more efficiently. In addition, according to Ribatet et al. [2007], the Bayesian approaches allow to relax the scale invariance property required by the most applied RFA model, that is, the index flood [Dairymple, 1960].

However, a preliminary study on simulated data showed that the approach developed by Ribatet et al. [2007] may lead to unreliable estimates for larger return periods ( $T > 20$  years) when small samples are involved. This poor performance is mainly due to the large variance on the shape parameter estimation. Consequently, for such cases, attention must be paid to the regional estimation procedure for the shape parameter.

The basis of our new development was formerly proposed by Stephenson and Tawn [2004]. They use reversible jump Markov chain Monte Carlo techniques [Green, 1995] to attribute a non null probability to the Gumbel case. Therefore, realizations are not supposed to be Gumbel distributed, but have a non null probability to be Gumbel distributed. An application to extreme rainfall and sea-level is given. In this work, this approach is extended to take into account a regional shape parameter, not only the Gumbel/Exponential

<sup>1</sup>CEMAGREF Lyon, Unité de Recherche Hydrologie-Hydraulique, 3 bis quai Chauvessou, CP220, 69638 Lyon cedex 09, FRANCE

<sup>2</sup>INRS-ETE, University of Québec, 490, de la Couronne Québec, Qc, G1K 9A9, CANADA.

case, within a RFA framework. The reversible jump technique allows to focus on a "likely" shape parameter value given by the hydrological relevance of the homogeneous region. Thus, this approach may reduce the shape parameter variance estimation while relaxing the scale invariance property.

The main objectives of this article is first to present new developments in the methodology proposed by Stephenson and Taun [2004] required for a RFA context; second to assess the quality of two Bayesian models based on the index flood hypothesis: the regional Bayesian model proposed by Ribatet et al. [2007] (BAY) and the new proposed Bayesian approach applying reversible jumps Markov chains (REV). They are compared to the classical index flood approach of Dalrymple [1960] (IFL). The assessment is developed through a stochastic generation of regional data performed in order to obtain realistic features of homogeneous regions. Detailing the index flood concept is out of the scope of this article. Estimation procedure can be found in Hosking and Wallis [1997].

The paper is organized as follows. The next two sections concentrate on methodological aspects. Section 2 describes the Bayesian framework including the specific Markov Chain Monte Carlo (MCMC) algorithm, required to extend the work by Stephenson and Taun [2004]. Section 3 presents the simple and efficient algorithm to generate stochastically hydrological homogeneous regions. A sensitivity analysis is performed in section 4 to assess how quantile estimates and related uncertainties are influenced by the values of two parameters of the reversible jump Markov chains. Section 5 compares the performance of each estimator on six representative case studies. The impact of the bias in the target site index flood estimation is analyzed in section 6, while suggestions for building efficient pooling groups are presented in section 7. Finally, some conclusions are drawn in section 8.

## 2. Methodology

In the Bayesian framework, the posterior distribution of parameters must be known to derive quantile estimates. The posterior distribution  $\pi(\theta|x)$  is given by the Bayes Theorem [Bayes, 1763]:

$$\pi(\theta|x) = \frac{\pi(\theta) \pi(x;\theta)}{\int_{\Theta} \pi(\theta) \pi(x;\theta) d\theta} \propto \pi(\theta) \pi(x;\theta) \quad (2)$$

where  $\theta$  is the vector of parameters of the distribution to be fitted,  $\Theta$  is the parameter space.  $\pi(x;\theta)$  is the likelihood function,  $x$  is the vector of observations and  $\pi(\theta)$  is the prior distribution.

In this study, as excesses over a high threshold are of interest, the likelihood function  $\pi(x;\theta)$  is related to the GPD - see equation (1).

### 2.1. Prior Distribution

In this section, the methodology to elicit the prior distribution is presented. In this study, regional information is used to define the prior distribution. Furthermore, the prior is specific as it must account for a fixed shape parameter  $\xi_{\text{Fix}}$  with a non null probability  $p_{\xi}$ . Let  $\Theta_0$  be a sub-space of the parameter space  $\Theta$  of  $\theta$ . More precisely,  $\Theta_0 = \{\theta \in \Theta : \xi = \xi_{\text{Fix}}\}$ .  $p_{\xi}$  is a hyper-parameter of the prior distribution. The approach is to construct a suitable prior distribution on  $\Theta$ ; then, for  $p_{\xi}$  fixed, to modify this prior to account for the probability of  $\Theta_0$ .

For clarity purposes, the prior distribution is defined in two steps. First, an initial prior distribution  $\pi_{\text{in}}(\theta)$  defined on  $\Theta$  is introduced. Second, a revised prior distribution  $\pi(\theta)$  is derived from  $\pi_{\text{in}}(\theta)$  to attribute a non null probability to the  $\Theta_0$  sub-sample.

#### 2.1.1. Initial prior distribution

As the proposed model is fully parametric, the initial prior distribution  $\pi_{\text{in}}(\theta)$  is a multivariate distribution entirely defined by its hyper-parameters. In our case study, the initial prior distribution corresponds to the one introduced by Ribatet et al. [2007]. Consequently, the marginal prior distributions were supposed to be independent lognormal for both location and scale parameters and normal for the shape parameter. Thus,

$$\pi_{\text{in}}(\theta) \propto J \exp\left[(\theta' - \gamma)^T \Sigma^{-1} (\theta' - \gamma)\right] \quad (3)$$

where  $\gamma, \Sigma$  are hyper-parameters,  $\theta' = (\log \mu, \log \sigma, \xi)$  and  $J$  is the Jacobian of the transformation from  $\theta'$  to  $\theta$ , namely  $J = 1/\mu\sigma$ .  $\gamma = (\gamma_1, \gamma_2, \gamma_3)$  is the mean vector,  $\Sigma$  is the covariance matrix. As marginal priors are supposed to be independent,  $\Sigma$  is a  $3 \times 3$  diagonal matrix with diagonal elements  $d_1, d_2, d_3$ .

Hyper-parameters are defined through the index flood concept, that is, all distributions are identical up to an at-site dependent constant. Consider all sites of a region except the target site - say the  $j$ -th site. A set of pseudo target site parameters can be computed:

$$\bar{\mu}_i = C^{(j)} \mu_i^{(i)} \quad (4)$$

$$\bar{\sigma}_i = C^{(j)} \sigma_i^{(i)} \quad (5)$$

$$\bar{\xi}_i = \xi_i^{(i)} \quad (6)$$

for  $i \neq j$ , where  $C^{(j)}$  is the target site index flood and  $\mu_i^{(i)}, \sigma_i^{(i)}, \xi_i^{(i)}$  are respectively the location, scale and shape at-site parameter estimates from the rescaled sample - e.g. normalized by its respective index flood estimate. Under the hypothesis of the index flood concept, pseudo-parameters are expected to be distributed as parameters of the target site.

Information from the target site sample can not be used to elicit the prior distribution. Thus,  $C^{(j)}$  in equations (4) and (5) must be estimated without use of the  $j$ -th sample site.

In this case study,  $C^{(j)}$  is estimated through a Generalized Linear Model (GLM) defined by:

$$\begin{cases} \mathbf{E}[\log C^{(j)}] & = \nu, & \nu = X\beta \\ \text{Var}[\log C^{(j)}] & = \phi V(\nu) \end{cases} \quad (7)$$

where  $X$  are basin characteristics (possibly log transformed),  $\phi$  is the dispersion parameter,  $V$  the variance function and  $\nu$  is the linear predictor. McCullagh and Nelder [1989] give a comprehensive introduction to GLM. Other alternatives for modeling the target site index flood can be considered such as Generalized Additive Models [Wood and Augustin, 2002], Neural Networks [Shu and Burn, 2004] or Kriging [Merz and Blöschl, 2005]. However, the variance of  $C^{(j)}$  should be estimated. Indeed, as  $C^{(j)}$  is estimated without use of the target site data, uncertainties due to this estimation must be incorporated in the prior distribution.

From these pseudo parameters, hyper-parameters can be computed:

$$\gamma_1 = \frac{1}{N-1} \sum_{i \neq j} \log \bar{\mu}_i^{(i)}, \quad d_1 = \frac{1}{N-1} \sum_{i \neq j} \text{Var}[\log \bar{\mu}_i^{(i)}] \quad (8)$$

$$\gamma_2 = \frac{1}{N-1} \sum_{i \neq j} \log \bar{\sigma}_i^{(i)}, \quad d_2 = \frac{1}{N-1} \sum_{i \neq j} \text{Var}[\log \bar{\sigma}_i^{(i)}] \quad (9)$$

$$\gamma_3 = \frac{1}{N-1} \sum_{i \neq j} \bar{\xi}_i^{(i)}, \quad d_3 = \frac{1}{N-2} \sum_{i \neq j} (\bar{\xi}_i^{(i)} - \gamma_3)^2 \quad (10)$$

Under the independence assumption between  $C^{(j)}$  and  $\mu_s^{(j)}, \sigma_s^{(j)}$ , the following relations hold:

$$\text{Var} \left[ \log \bar{\mu}^{(j)} \right] = \text{Var} \left[ \log C^{(j)} \right] + \text{Var} \left[ \log \mu_s^{(j)} \right] \quad (11)$$

$$\text{Var} \left[ \log \bar{\sigma}^{(j)} \right] = \text{Var} \left[ \log C^{(j)} \right] + \text{Var} \left[ \log \sigma_s^{(j)} \right] \quad (12)$$

The independence assumption is not too restrictive as the target site index flood is estimated independently from  $\mu_s^{(j)}, \sigma_s^{(j)}$ .

Note that  $\text{Var} \left[ \log \bar{\mu}^{(j)} \right]$  are estimated thanks to Fisher information and the delta method. Estimation of  $\text{Var} \left[ \log C^{(j)} \right]$  is a special case and depends on the method for estimating the at-site index flood. Nevertheless, it is always possible to carry out an estimation of this variance, at least through standard errors.

### 2.1.2. Revised prior distribution

The initial prior distribution  $\pi_{1s}(\theta)$  gives a null probability to the sub-sample  $\Theta_0$ . Thus, from this initial prior  $\pi_{1s}(\theta)$ , a revised prior  $\pi(\theta)$  is constructed to attribute a non null probability to the  $\Theta_0$  sub-sample. According to Stephenson and Tawn [2004],  $\pi(\theta)$  is defined as:

$$\pi(\theta) = \begin{cases} (1 - p_\xi) \pi_{1s}(\theta) & \text{for } \theta \in \Theta \setminus \Theta_0 \\ p_\xi \pi_{\xi_{\text{fix}}}(\theta) & \text{for } \theta \in \Theta_0 \end{cases} \quad (13)$$

where  $p_\xi \in [0, 1]$  and with

$$\pi_{\xi_{\text{fix}}}(\theta) = \frac{\pi_{1s}(\mu, \sigma, \xi_{\text{fix}})}{\int_{\mu, \sigma} \pi_{1s}(\mu, \sigma, \xi_{\text{fix}}) d\mu d\sigma} \quad (14)$$

for  $\theta \in \Theta_0$ . The integral in equation (14) can be easily evaluated by standard numerical integration methods.

By construction, the new prior distribution  $\pi(\theta)$  gives the required probability to the sub-space  $\Theta_0$ . Stephenson and Tawn [2004] have already applied formulations (13) and (14) with success for sea-level maxima and rainfall threshold exceedences.

### 2.2. Posterior Estimation

As it is often the case in Bayesian analysis, the integral in equation (2) is insolvable analytically. MCMC techniques are used to overcome this problem. Yet, due to the duality of  $\pi(\theta)$  distribution, standard Metropolis-Hastings [Hastings, 1970] within Gibbs [Geman and Geman, 1984] methods are not sufficient. Reversible jump techniques [Green, 1995] are used to allow moves from the two dimensional space  $\Theta_0$  to the three dimensional space  $\Theta \setminus \Theta_0$  and vice-versa.

The classical Bayesian analysis, on  $\Theta \setminus \Theta_0$ , is performed with Gibbs cycle over each component of  $\theta$  using Metropolis-Hastings updates, with random walk proposals [Coles and Tawn, 1996].

Stephenson and Tawn [2004] extended this algorithm to incorporate the mass on the Gumbel/Exponential case. However, as our approach does not only focus on the  $\xi_{\text{fix}} = 0$  case, a new algorithm must be implemented. To help understand the algorithmic developments, some details about the classical Metropolis-Hastings algorithm and the reversible jump case are reported in Appendix A.

The proposed algorithm must deal with two dimensional changes: a change to  $\Theta_0$  from  $\Theta \setminus \Theta_0$  space and vice-versa. These two types of special moves must be defined cautiously. As inspired by Stephenson and Tawn [2004], quantiles associated to a non exceedence probability  $p$  are set to be equal at current state  $\theta_t$  and proposal  $\theta_{\text{prop}}$ ,  $p$  being fixed.

For a proposal move to  $\Theta \setminus \Theta_0$  from  $\Theta_0$ , i.e.,  $\xi_t = \xi_{\text{fix}}$  and a proposal shape  $\xi_{\text{prop}} \neq \xi_{\text{fix}}$ , the candidate move is to

change  $\theta_t = (\mu_t, \sigma_t, \xi_t)$  to  $\theta_{\text{prop}} = (\mu_{\text{prop}}, \sigma_{\text{prop}}, \xi_{\text{prop}})$  where

$$\mu_{\text{prop}} = \mu_t \quad (15a)$$

$$\sigma_{\text{prop}} = \sigma_t \frac{\xi_{\text{prop}}(y^{-\xi_t} - 1)}{\xi_t(y^{-\xi_{\text{prop}}} - 1)} \quad (15b)$$

$$\xi_{\text{prop}} \sim \mathcal{N}(\bar{\xi}, s_\xi^2) \quad (15c)$$

where  $y = 1 - p$ ,  $p$  being fixed,  $\bar{\xi}$  is taken to be the mode of the marginal distribution for  $\xi$  when there is no mass on  $\Theta_0$  [Stephenson and Tawn, 2004], and  $s_\xi$  is the standard deviation selected to give good mixing properties to the chain. As it is usually the case with Metropolis-Hastings updates, this move is accepted with probability  $\min(1, \Delta)$  with

$$\Delta = \frac{\pi(\mu_{\text{prop}}, \sigma_{\text{prop}}, \xi_{\text{prop}} | x)}{\pi(\mu_t, \sigma_t, \xi_{\text{fix}} | x)} \frac{p_\xi}{1 - p_\xi} \left[ \phi(\xi_{\text{prop}}; \bar{\xi}, s_\xi^2) J_{\xi_{\text{fix}}}(\xi_{\text{prop}}) \right]^{-1} \quad (16)$$

where  $\phi(\cdot; m, s^2)$  denotes the density function of the Normal distribution with mean  $m$  and variance  $s^2$ , and  $J_{\xi_{\text{fix}}}$  is the Jacobian of the parameter transformation for quantile matching, that is:

$$J_{\xi_{\text{fix}}}(\xi) = \frac{\xi_{\text{fix}}}{\xi} \frac{y^{-\xi} - 1}{y^{-\xi_{\text{fix}}} - 1} \quad (17)$$

If the move is accepted, then  $\theta_{t+1} = (\mu_{\text{prop}}, \sigma_{\text{prop}}, \xi_{\text{prop}})$ , else  $\theta_{t+1} = \theta_t$ .

For a proposal move to  $\Theta_0$  from  $\Theta \setminus \Theta_0$ , i.e.,  $\xi_t \neq \xi_{\text{fix}}$  and a proposal shape  $\xi_{\text{prop}} = \xi_{\text{fix}}$ , the proposal is to change  $\theta_t = (\mu_t, \sigma_t, \xi_t)$  to  $\theta_{\text{prop}} = (\mu_{\text{prop}}, \sigma_{\text{prop}}, \xi_{\text{prop}})$  where

$$\mu_{\text{prop}} = \mu_t \quad (18a)$$

$$\sigma_{\text{prop}} = \sigma_t \frac{\xi_{\text{prop}}(y^{-\xi_t} - 1)}{\xi_t(y^{-\xi_{\text{prop}}} - 1)} \quad (18b)$$

$$\xi_{\text{prop}} = \xi_{\text{fix}} \quad (18c)$$

This move is accepted with probability  $\min(1, \Delta)$  where

$$\Delta = \frac{\pi(\mu_{\text{prop}}, \sigma_{\text{prop}}, \xi_{\text{fix}} | x)}{\pi(\mu_t, \sigma_t, \xi_t | x)} \frac{1 - p_\xi}{p_\xi} \phi(\xi_t; \bar{\xi}, s_\xi^2) J_{\xi_{\text{fix}}}(\xi_t) \quad (19)$$

If the move is accepted, then  $\theta_{t+1} = (\mu_{\text{prop}}, \sigma_{\text{prop}}, \xi_{\text{prop}})$  else  $\theta_{t+1} = \theta_t$ .

Obviously, special moves introduced in this study are not the only conceivable ones. Other reversible jumps can be explored - see for example Stephenson and Tawn [2004]. However, for this application, the proposed moves seem to be particularly well suited. Indeed, a preliminary study shows that the location parameter was well estimated by a regional Bayesian approach. Thus, a special move which only affects the shape and scale parameters should be consistent.

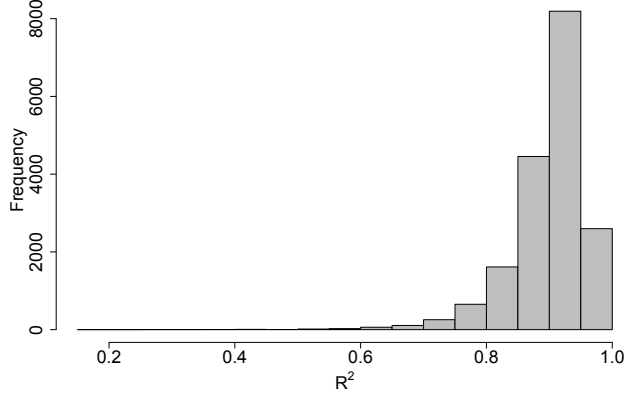
### 3. Generation Procedure

In this section, the procedure implemented to generate stochastic homogeneous regions is described. The idea consists in generating sample points in a neighborhood of the L-moment space (Mean, L-CV, L-Skewness). The generation procedure can be summarized as follows:

1. Set the center of the neighborhood i.e.  $(l_{1,R}, \tau_R, \tau_{2,R})$  or equivalently parameters of the regional distribution  $(\mu_R, \sigma_R, \xi_R)$ ;
2. Generate  $N$  points  $(l_{1,i}, \tau_i, \tau_{2,i})$  uniformly in the sphere  $B((l_{1,R}, \tau_R, \tau_{2,R}); \epsilon)$ ;
3. Generate  $N$  index floods  $C$  using the scaling model parametrization:

$$C = \alpha \text{Area}^\beta \quad (20)$$

Catchment areas are defined as realizations of a lognormal random variable.



**Figure 1.** Histogram of the coefficient of determination for the regressive model (7). Application of section 5.

4. For each  $(l_{1,i}, \tau_i, \tau_{3,i})$ , compute adimensional parameters by:

$$\xi_i^* = \frac{3\tau_{3,i} - 1}{1 + \tau_{3,i}} \quad (21a)$$

$$\sigma_i^* = (\xi_i^* - 1)(\xi_i^* - 2)l_{1,i}\tau_i \quad (21b)$$

$$\mu_i^* = l_{1,i} - \frac{\sigma_i^*}{1 - \xi_i^*} \quad (21c)$$

5. Then, compute at-site parameters from:

$$\xi_i = \xi_i^* \quad (22a)$$

$$\sigma_i = C_i \sigma_i^* \quad (22b)$$

$$\mu_i = C_i \mu_i^* \quad (22c)$$

6. Simulate samples from a GPD with parameters  $(\mu_i, \sigma_i, \xi_i)$ .

As a GLM is used to elicit the prior distribution, the scaling model (20) must be altered to avoid giving an advantage to the Bayesian approaches over the index flood model. For this purpose, a noise in relation (20) at step 3 is introduced. Thus, areas are altered by adding uniform random variables varying in  $(-0.5 \times Area, 0.5 \times Area)$ .

This distortion is necessary to ensure that the regressive model is not too competitive and is consistent with observations. Indeed large deviations to the area-index flood relationship are often encountered in practice. In the following applications,  $\alpha = 0.12$ ,  $\beta = 1.01$  and  $Area \sim \mathcal{LN}(4.8, 1)$ . These values arise from a previous study on a French data set [Ribatet et al., 2007] and ensure realistic magnitudes. For the application of section 5, the coefficients of determination for the regressive model (7) varies from 0.20 to 0.99, with a mean value of 0.89. The histogram of these coefficients of determination is presented in Figure 1. The radius  $\varepsilon$  in the generation algorithm is set to 0.04. This value is chosen to reflect variability met in practice while preserving a low dispersion around the regional distribution. The  $\varepsilon$  value primarily impacts the proportions of regions satisfying  $H_1 < 1$ . For specific applications, regions with a heterogeneity statistic  $H_1$  such as  $H_1 > 1$  may be discarded.

#### 4. Sensitivity Analysis

In this section, a sensitivity analysis for the algorithm introduced in section 2.2 is carried out. The primary goal is to check if results are not too impacted by the choice of the two user-selectable parameters  $p_\xi$  and  $\xi_{\text{Fix}}$ . For this purpose, the effect of both  $p_\xi$  and  $\xi_{\text{Fix}}$  values on estimates and

credibility intervals is examined. For this sensitivity analysis, the parameters of the regional distribution is set to be (0.64, 0.48, 0.26). The regions have 20 sites with a sample size of 70. For the whole sensitivity analysis, 10 000 regions were generated. The target site has a sample size of 10. We concentrate on estimates at sites with very few data, to exhibit the main differences in the most restricting configuration. Other configurations were found to demonstrate features similar to Fig. 3 and Fig. 2.

##### 4.1. Effect of $p_\xi$

The evolution of the normalized biases (expressed in percent) for return levels with non exceedence probabilities 0.75, 0.95 and 0.995 associated to several  $p_\xi$  values are depicted in Fig. 3. Each boxplot is obtained from at-site estimates computed on more than 365 stochastic homogeneous regions. The case  $p_\xi = 0$  corresponds to a classical Bayesian approach free from any point mass. In addition, to analyze only the effect of the parameter  $p_\xi$ ,  $\xi_{\text{Fix}}$  is temporarily fixed to be equal to the theoretical regional shape parameter.

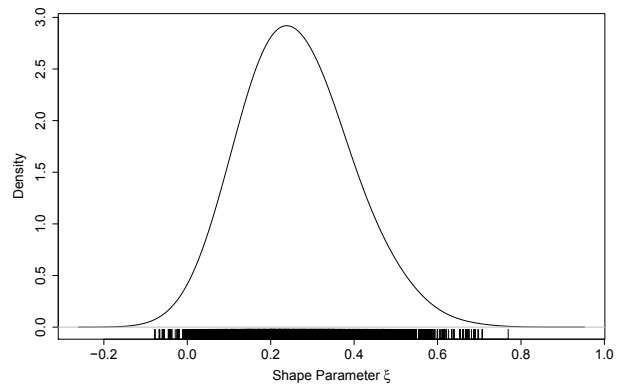
From Fig. 3, the quantile estimates distribution seems to be stationary, provided that  $p_\xi > 0$ . Introducing a point mass does not impact  $Q_{0.75}$  estimates, whereas significant reduction in median biases and scatter of estimates is noticeable for more extremal quantiles.

Fig. 4 shows the posterior distributions of return levels and 90% posterior credibility intervals for several  $p_\xi$  values.

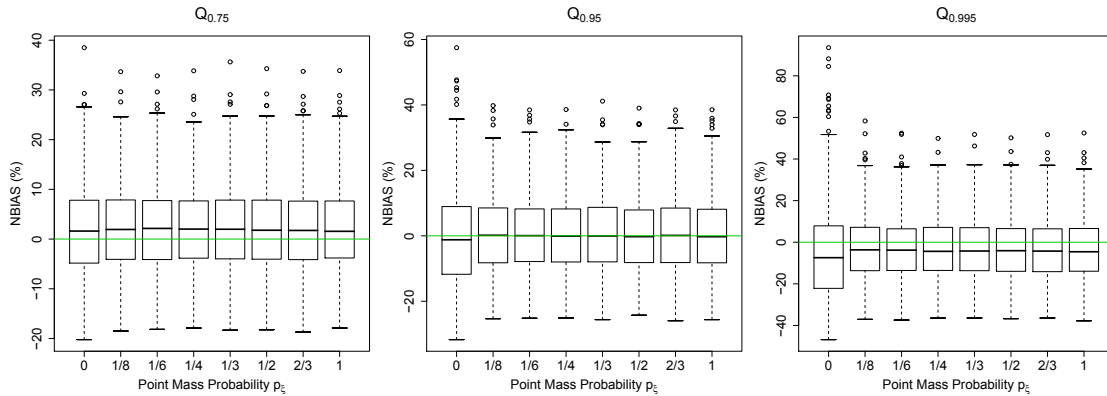
It is clear that credibility intervals are sensitive to the  $p_\xi$  value. This result is consistent as more and more proposals in the MCMC simulation belong to  $\Theta_0$  as  $p_\xi$  increases. Thus, by construction, the Markov chain is less variable. As denoted by Stephenson and Tawn [2004], the special case

**Table 1.** Posterior proportions (in percent) of events  $\{\theta \in \Theta_0\}$  for different values of  $p_\xi$  and  $\xi_{\text{Fix}}$ . Target Sample Size 60.

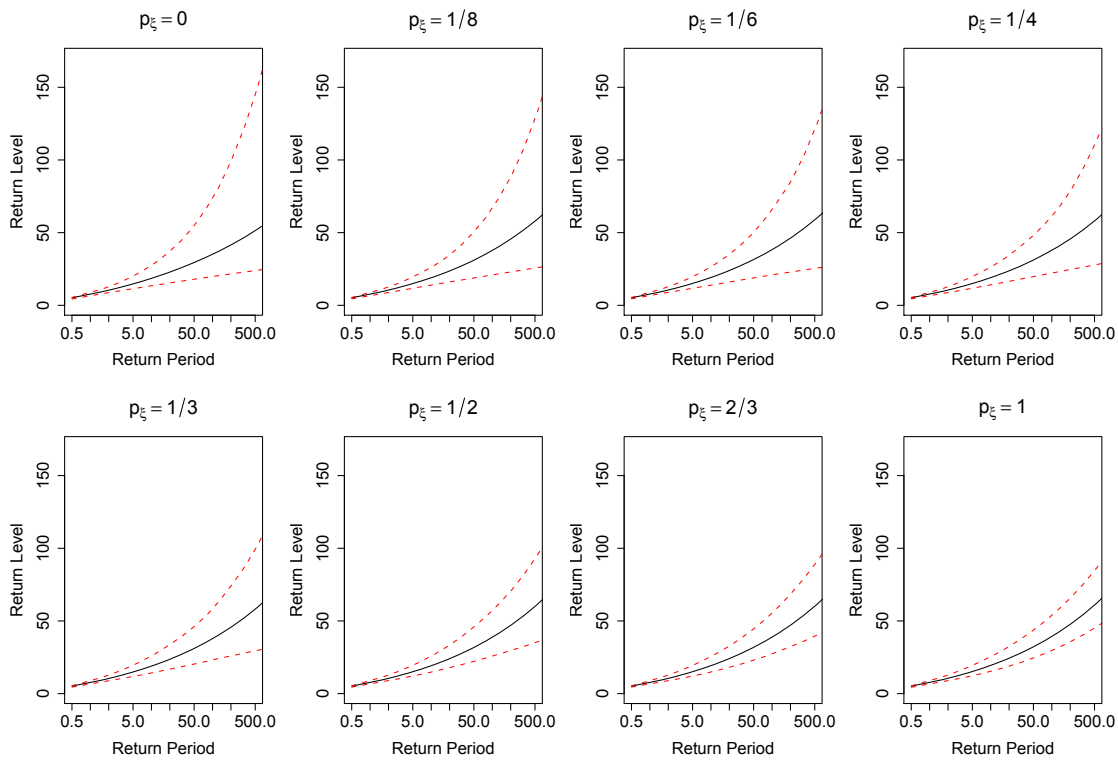
$\xi_{\text{Fix}}$ features		$p_\xi$ values					
$R_{\text{Shape}}$	$D_{\text{Shape}}$	1/8	1/6	1/4	1/3	1/2	2/3
-0.50	$2e^{-5}$	0.00	0.03	0.00	0.00	0.05	0.00
0.00	0.06	10.07	14.55	17.27	21.99	41.53	61.84
0.50	0.70	38.88	46.94	59.96	67.42	81.88	92.17
0.83	1.00	46.21	57.33	67.53	76.08	85.33	92.20
1.00	0.87	48.24	55.14	68.90	76.16	86.05	91.85
1.50	0.41	32.72	45.61	54.62	66.18	82.11	89.90
2.00	0.10	22.95	22.83	35.06	49.82	57.86	81.92
2.50	0.01	13.93	7.04	9.86	36.21	38.89	42.28



**Figure 2.** Posterior marginal density for the shape parameter.



**Figure 3.** Effect of  $p_\xi$  value on quantile estimation with non exceedence probabilities 0.75, 0.95 and 0.995. Sample size 10.  $\xi_{\text{Fix}} = 0.26$ .



**Figure 4.** Effect of  $p_\xi$  value on 90% posterior credibility interval. Sample Size 10.

$p_\xi = 1$  is particular as uncertainty in the shape parameter is not considered. In that case, credibility intervals could be falsely narrow.

#### 4.2. Effect of $\xi_{\text{Fix}}$

It is important to analyze the influence of the choice of  $\xi_{\text{Fix}}$  on the simulated Markov chains; and thus, its impact on estimations. Indeed, when specifying an unreasonable  $\xi_{\text{Fix}}$  value, the estimations must not differ significantly from the conventional Bayesian ones. For this purpose, Tab. 1 displays the posterior proportions of events  $\{\theta \in \Theta_0\}$  for several  $\xi_{\text{Fix}}$  and  $p_\xi$  values. This table is obtained with a target site sample size of 60. For each specified  $\xi_{\text{Fix}}$  value, two features are computed to measure the relevance of the  $\xi_{\text{Fix}}$  value: (a)  $R_{\text{Shape}}$  the ratio of  $\xi_{\text{Fix}}$  to the true shape parameter ; and

(b)  $D_{\text{Shape}}$  the ratio of the marginal posterior density from a conventional Bayesian analysis evaluated in  $\xi_{\text{Fix}}$  and  $\xi$ .

$R_{\text{Shape}}$  characterizes how much the point Mass differs from the true value.  $D_{\text{Shape}}$  quantifies the distance of  $\xi_{\text{Fix}}$  from the estimator of the shape parameter proposed by Ribatet *et al.* [2007]. Thus, from these two statistics, consistency of the posterior proportions with deviations from theoretical and empirical values can be analyzed.

The results in Tab. 1 show that values of  $\xi_{\text{Fix}}$  that are not consistent with the data imply low proportions of state in  $\Theta_0$ . Thus, for such values, the proposed model is quite similar to a conventional Bayesian analysis. However, for two different values of  $\xi_{\text{Fix}}$  ( $R_{\text{Shape}}$  equal to 0.83 and 1), the posterior proportions are quite equivalent. This emphasizes the large uncertainty on the shape parameter estimation for small sample sizes. Uncertainty on the shape parameter



estimation is also corroborated by the posterior marginal distribution of a conventional Bayesian analysis - see Fig. 2.

As noticed above, these results are obtained with a target site sample size of 60. This particular sample size was selected as it is the most illustrative case. However, the posterior proportions are quite similar when dealing with other target site sample sizes - even if for very small sample sizes, this is less noticeable.

## 5. Simulation Study

In this section, performance of three different estimators are analyzed: a conventional Bayesian estimator (BAY) introduced by Ribatet et al. [2007], the proposed estimator based on reversible jumps (REV) and the index flood estimator (IFL). In particular, the BAY estimator is related to the initial prior distribution defined in Section 2.1.1. Thus, the BAY estimator is identical to the REV approach with  $p_\xi = 0$ .

For the proposed estimator, the point Mass probability  $p_\xi$  was set to be a function of the  $H_1$  statistic of Hosking and Wallis [1997]; that is:

$$p_\xi = \frac{\exp(-H_1)}{1 + \exp(-H_1)} \quad (23)$$

For this parametrization, necessary requirements are satisfied; i.e.  $p_\xi \rightarrow 0$  when  $H_1 \rightarrow +\infty$  and  $p_\xi \rightarrow 1$  when  $H_1 \rightarrow -\infty$ . Moreover, for  $H_1 = 0$ ,  $p_\xi = 0.5$  which corresponds to the estimator introduced by Stephenson and Tawn [2004]. Note that  $p_\xi$  in Eq. (23) is defined with the negative inverse of the so called logit function.

Thus, for this choice, as underlined by the sensitivity analysis, credibility intervals are related to the degree of confidence of the point Mass  $\xi_{\text{true}}$  to be the true shape parameter and implicitly to the level of homogeneity of the regions.

In addition, the non exceedence probability  $p$  used for quantiles matching in our algorithm (see Section 2.2) is equal to  $1 - 1/2n$ , where  $n$  is the target site sample size. This last point guarantees that quantiles associated with non exceedence probability  $1 - 1/2n$  for both proposal and current state of the Markov chain are identical. Other choices for  $p$  are arguable. Here, we introduce a quantile matching equation for a value closely related to the scale parameter and for which uncertainties are not too large.

The analysis was performed on six different case studies summarized in Tab. 2. The configurations differ by the way information is distributed in space; that is, (a) "small regions" with well instrumented but few sites (Conf1 and Conf4); (b) "large regions" with less instrumented and numerous sites (Conf2 and Conf5) and (c) "medium regions" with well instrumented sites and an intermediate number of gauging stations. Conf1 (resp. Conf2, Conf3) correspond to Conf4 (resp. Conf5, Conf6) apart from the  $(\mu_R, \sigma_R, \xi_R)$  values. The target site sample size takes the values in 10, 25 and 40. 1000 regions were generated for each configuration.

Table 2. Characteristics of the sixth case studies. The target site is omitted in the couple  $(n_{\text{Site}}, n_{\text{Site}})$  and has a sample size of: 10, 25 and 40.

	$(\mu_R, \sigma_R, \xi_R)$	$N_{\text{Site}}$	$(n_{\text{Site}}, n_{\text{Site}})$	$N_{\text{Events}}$
Conf1	(0.64, 0.48, 0.26)	10	(9, 50)	450
Conf2	(0.64, 0.48, 0.26)	20	(9, 30) × (10, 18)	450
Conf3	(0.64, 0.48, 0.26)	15	(14, 50)	700
Conf4	(0.66, 0.48, 0.08)	10	(9, 50)	450
Conf5	(0.66, 0.48, 0.08)	20	(9, 30) × (10, 18)	450
Conf6	(0.66, 0.48, 0.08)	15	(14, 50)	700

Markov chains of length 15 000 were generated. To ensure good mixing properties for all simulated Markov chains, an automated trial and error process was used to define proposal standard deviations of the MCMC algorithm. Furthermore, the first 2000 iterations were discarded to ensure that the equilibrium was reached.

The performance of each estimator is assessed through the three following statistics:

$$NBIAS = \frac{1}{k} \sum_{i=1}^k \frac{\hat{Q}_i - Q_i}{Q_i} \quad (24)$$

$$SD = \sqrt{\frac{1}{k-1} \sum_{i=1}^k \left( \frac{\hat{Q}_i - Q_i}{Q_i} - NBIAS \right)^2} \quad (25)$$

$$NMSE = \frac{1}{k} \sum_{i=1}^k \left( \frac{\hat{Q}_i - Q_i}{Q_i} \right)^2 \quad (26)$$

where  $\hat{Q}_i$  is the estimate of the theoretical value  $Q_i$  and  $k$  is the total number of theoretical values.

### 5.1. BAY vs. IFL Approach

Table 3 shows that, for a small target site sample size and quantiles  $Q_{0.75}$  and  $Q_{0.95}$ , the BAY approach is more competitive than the IFL one. Indeed, the three BAY statistics (NBIAS, SD, NMSE) are smaller than the ones related to IFL. However, for Conf2 and Conf5, IFL  $Q_{0.95}$  estimates are more competitive. These two case studies correspond to the same configuration - i.e. numerous sites with short records. IFL estimates for  $Q_{0.95}$  are always more accurate than BAY for all configurations.

These results indicate that the relative performance of BAY compared to IFL depends on the pooling group. Thus, for the BAY approach and quantiles  $Q_{0.75}$  and  $Q_{0.95}$ , it seems preferable to work with less gauging stations but which have larger data series, independently of the target site sample size. The sensitivity to the configuration of the sites and the availability of long time series is a drawback for the application of this Bayesian approach.

These conclusions obtained on stochastic regions are in line with a previous analysis on a French data set [Ribatet et al., 2007]. The BAY approach is suited to work with "small" or "medium" regions and well instrumented gauging stations. In addition, this approach is accurate for "reasonable" quantile estimation - see the bad performance of BAY for  $Q_{0.95}$  in table 3.

However, the white noise introduced in the generation procedure is independent of the target site sample size. It only regards both Bayesian approaches. Thus, the performance of the BAY estimator for large sample sizes may be too impacted. Indeed, while the IFL estimation procedure is not altered, both Bayesian approaches must deal with artificially generated biases.

The main idea for the REV approach is to combine the good performance of the BAY estimator for "reasonable" quantiles and the efficiency of the IFL approach for larger quantiles.

### 5.2. BAY vs. REV Approach

The comparison of the two Bayesian estimators is summarized in Tab. 4. REV leads to more accurate estimated quantiles, in particular for  $Q_{0.95}$  and  $Q_{0.995}$ . This last point confirms the benefits of using a regional shape parameter through a reversible jump approach.

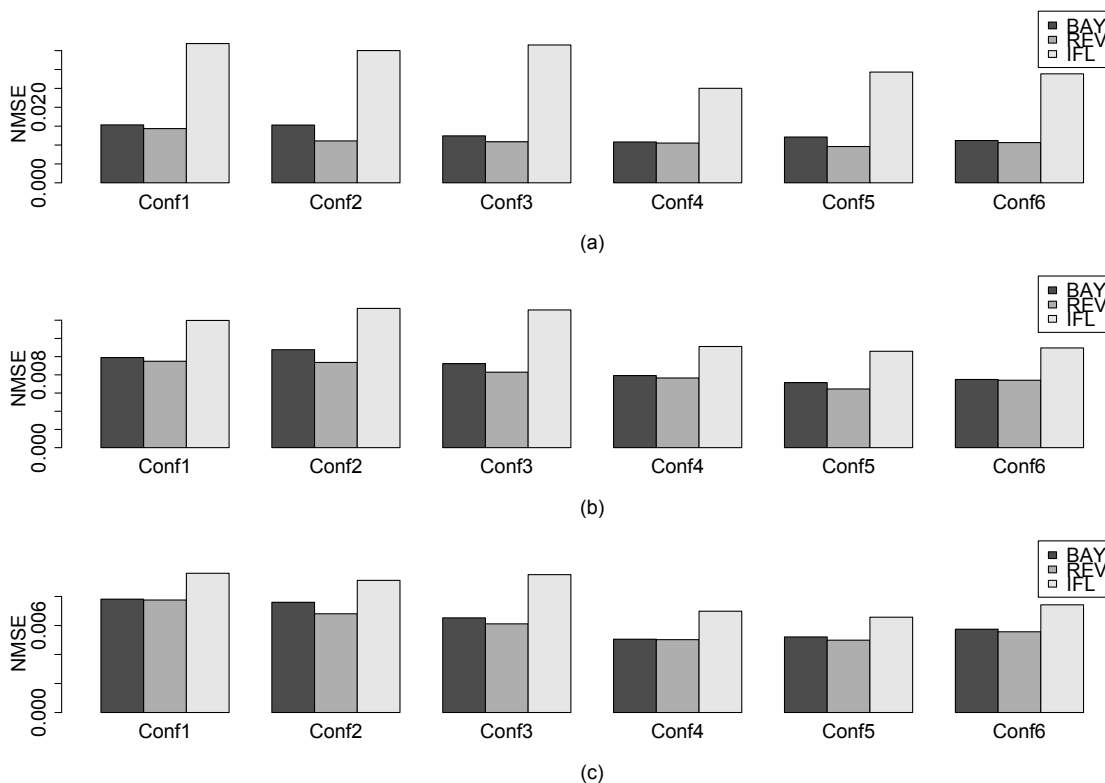
By construction of the algorithm described in section 2.2, Markov chains generated from the REV approach are less variable than the ones generated from the BAY model.

**Table 3.** Performance of *BAY* and *IFL* estimators for quantile  $Q_{0.75}, Q_{0.95}$  and  $Q_{0.995}$ . Target site sample size: 10.

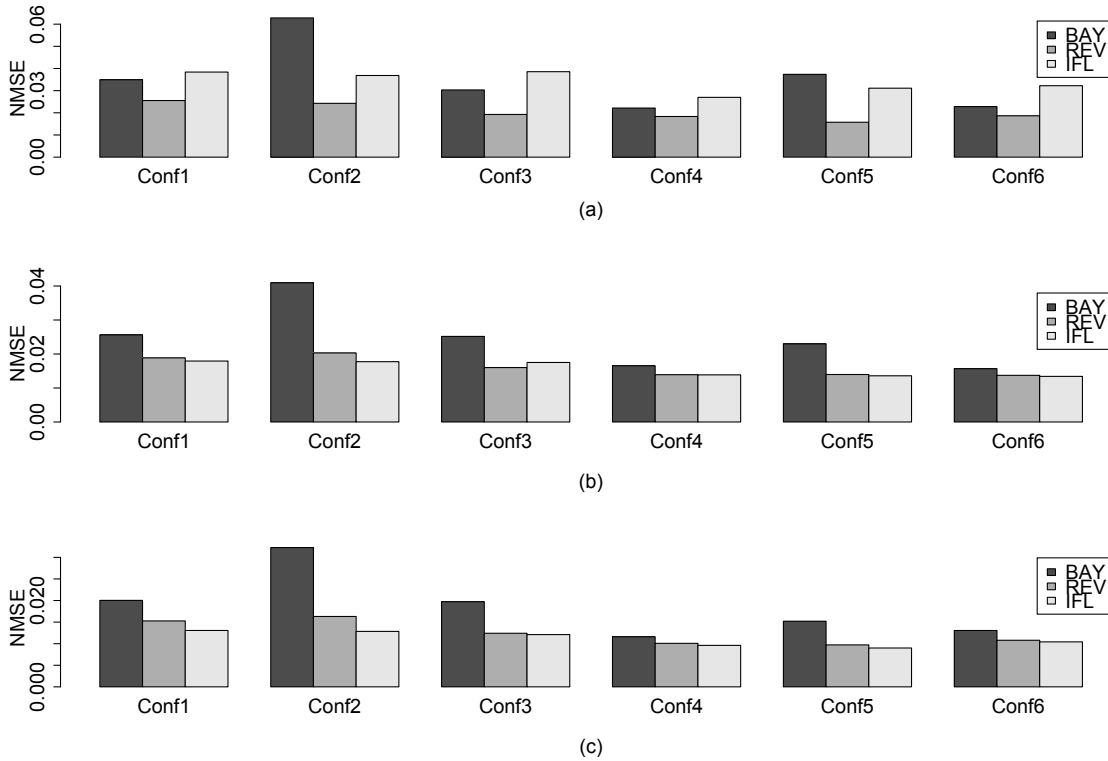
Model	$Q_{0.75}$			$Q_{0.95}$			$Q_{0.995}$		
	<i>NBIAS</i>	<i>SD</i>	<i>NMSE</i>	<i>NBIAS</i>	<i>SD</i>	<i>NMSE</i>	<i>NBIAS</i>	<i>SD</i>	<i>NMSE</i>
Conf1									
<i>BAY</i>	0.015	0.123	0.015	0.001	0.187	0.035	-0.006	0.318	0.101
<i>IFL</i>	0.037	0.189	0.037	0.025	0.195	0.038	-0.004	0.230	0.053
Conf2									
<i>BAY</i>	0.019	0.122	0.015	0.030	0.249	0.063	0.110	0.561	0.326
<i>IFL</i>	0.041	0.183	0.035	0.025	0.191	0.037	-0.022	0.221	0.049
Conf3									
<i>BAY</i>	0.019	0.110	0.012	0.006	0.174	0.030	-0.003	0.292	0.085
<i>IFL</i>	0.035	0.188	0.037	0.025	0.195	0.039	-0.002	0.222	0.049
Conf4									
<i>BAY</i>	0.009	0.104	0.011	-0.007	0.149	0.022	-0.021	0.233	0.054
<i>IFL</i>	0.023	0.157	0.025	0.022	0.163	0.027	0.022	0.192	0.037
Conf5									
<i>BAY</i>	0.018	0.109	0.012	0.012	0.193	0.037	0.033	0.378	0.144
<i>IFL</i>	0.036	0.168	0.029	0.033	0.173	0.031	0.024	0.197	0.039
Conf6									
<i>BAY</i>	0.024	0.103	0.011	0.001	0.151	0.023	-0.038	0.222	0.050
<i>IFL</i>	0.028	0.168	0.029	0.028	0.177	0.032	0.028	0.202	0.042

**Table 4.** Performance of *BAY* and *REV* estimators for quantile  $Q_{0.75}, Q_{0.95}$  and  $Q_{0.995}$ . Target site sample size: 10.

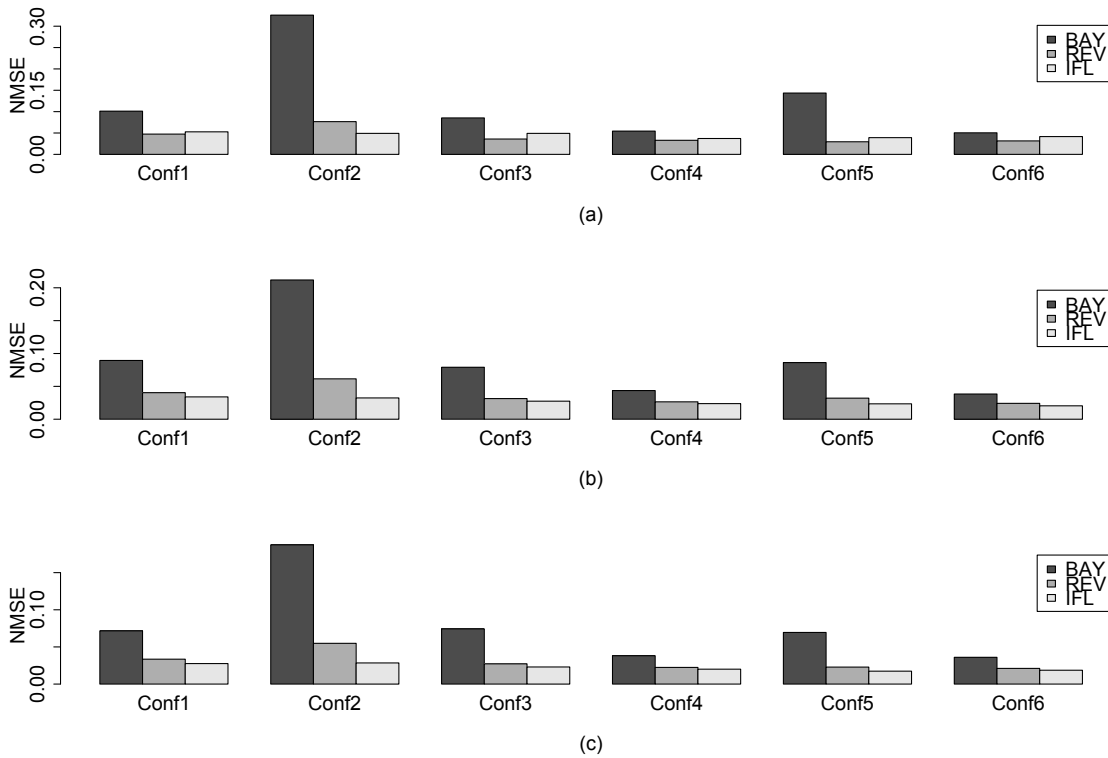
Model	$Q_{0.75}$			$Q_{0.95}$			$Q_{0.995}$		
	<i>NBIAS</i>	<i>SD</i>	<i>NMSE</i>	<i>NBIAS</i>	<i>SD</i>	<i>NMSE</i>	<i>NBIAS</i>	<i>SD</i>	<i>NMSE</i>
Conf1									
<i>BAY</i>	0.015	0.123	0.015	0.001	0.187	0.035	-0.006	0.318	0.101
<i>REV</i>	0.011	0.119	0.014	-0.012	0.159	0.026	-0.046	0.213	0.047
Conf2									
<i>BAY</i>	0.019	0.122	0.015	0.030	0.249	0.063	0.110	0.561	0.326
<i>REV</i>	0.005	0.105	0.011	-0.026	0.154	0.024	-0.066	0.269	0.077
Conf3									
<i>BAY</i>	0.019	0.110	0.012	0.006	0.174	0.030	-0.003	0.292	0.085
<i>REV</i>	0.014	0.103	0.011	-0.008	0.139	0.019	-0.042	0.185	0.036
Conf4									
<i>BAY</i>	0.009	0.104	0.011	-0.007	0.149	0.022	-0.021	0.233	0.054
<i>REV</i>	0.010	0.102	0.011	0.002	0.136	0.018	-0.001	0.182	0.033
Conf5									
<i>BAY</i>	0.018	0.109	0.012	0.012	0.193	0.037	0.033	0.378	0.144
<i>REV</i>	0.013	0.097	0.010	0.000	0.126	0.016	-0.014	0.171	0.030
Conf6									
<i>BAY</i>	0.024	0.103	0.011	0.001	0.151	0.023	-0.038	0.222	0.050
<i>REV</i>	0.031	0.099	0.011	0.033	0.133	0.019	0.034	0.174	0.032



**Figure 5.** Evolution of the *NMSE* for quantile  $Q_{0.75}$  in function of the region configuration. Target site sample size: (a) 10, (b) 25 and (c) 40.



**Figure 6.** Evolution of the  $NMSE$  for quantile  $Q_{0.95}$  in function of the region configuration. Target site sample size: (a) 10, (b) 25 and (c) 40.

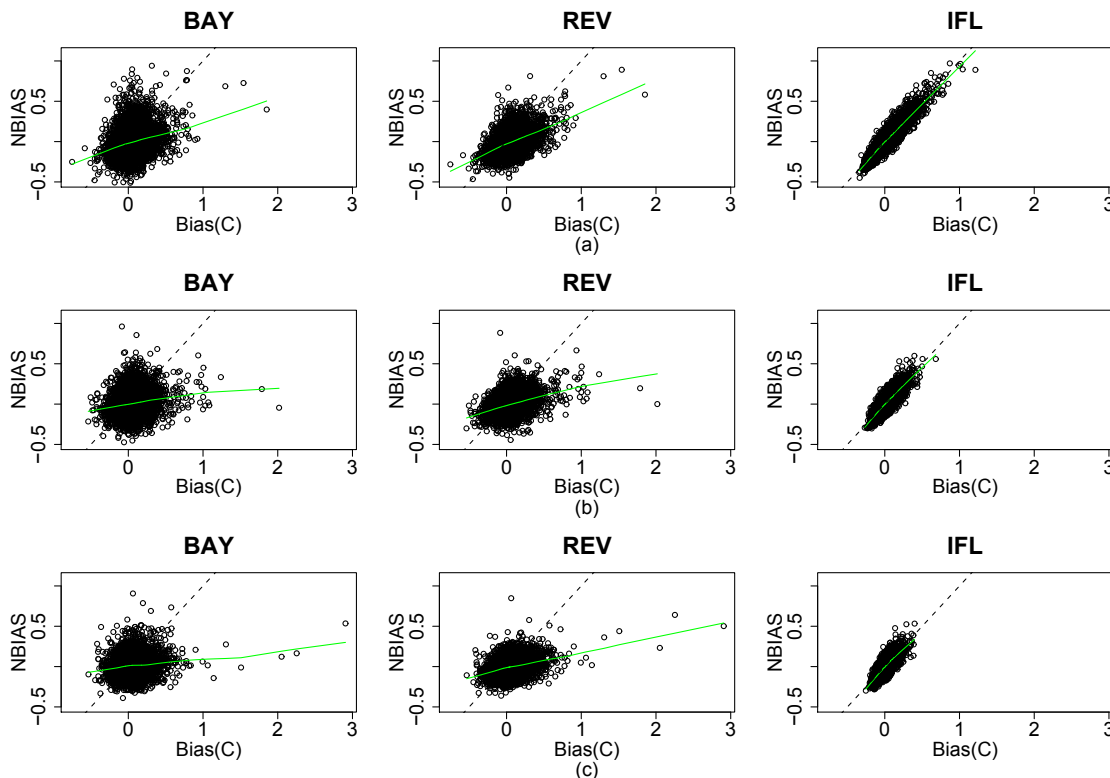


**Figure 7.** Evolution of the  $NMSE$  for quantile  $Q_{0.995}$  in function of the region configuration. Target site sample size: (a) 10, (b) 25 and (c) 40.

should have the same biases than *BAY*. Thereby, the *REV* estimator always leads to a smaller  $NMSE$ .

Figures 5 to 7 illustrate the results for different target site sample sizes and regions. We concentrate on the  $NMSE$  cri-

5.3. Global Comparison



**Figure 8.** Evolution of  $NBIAS$  for  $Q_{0.95}$  in function of the normalized bias on target site index flood estimation ( $Bias(C)$ ). Target site sample size: (a) 10, (b) 25 and (c) 40. Solid green lines: local smoothers, black dashed lines:  $y = x$ .

teria since it measures variation of the estimator around the true parameter value.

From Figure 5, it is clear that Bayesian estimations, i.e.  $BAY$  and  $REV$ , of  $Q_{0.75}$  are more accurate; specially for a target site sample size of 10. For larger target site sample sizes, Bayesian approaches are always more competitive than the  $IFL$  estimator, even if this is less clear-cut on the graphs. Furthermore,  $BAY$  and  $REV$  estimators often have the same performance. This result is logical as the  $Q_{0.75}$  value is mostly impacted by the location parameter  $\mu$ . Thus, reversible jumps do not have a significant result on  $REV$   $Q_{0.75}$  estimation.

The plots in Figure 6 and those displayed in Figure 5 are quite different. For a target site sample size of 10, both Bayesian approaches are the most accurate - except for  $BAY$  applied on  $Conf2$  and  $Conf5$  - and the  $REV$  estimator leads always to the smallest  $NMSE$ . Thus,  $REV$  is the most competitive model. For larger target site sample sizes,  $REV$  is at least as accurate as  $IFL$ , except for  $Conf2$ .

For  $Q_{0.995}$  and a target site sample size of 10,  $REV$  is the most accurate model, except for  $Conf2$ . As the target site sample size increases, the  $IFL$  approach becomes more efficient. However, for these cases,  $NMSE$  for the  $REV$  estimator are often close to the  $IFL$  ones. Although the  $BAY$  approach performs poorly for  $Q_{0.995}$ , its  $NMSE$  for  $Conf6$  is close to the  $REV$  and  $IFL$  ones.

In conclusion, these results illustrate the good overall performance of the  $REV$  model. Indeed, this approach benefits from the efficiency of the  $BAY$  estimator for quantiles with small non exceedence probabilities while being as competitive as the  $IFL$  approach for larger non exceedence probabilities.

However, the Bayesian approaches outperform the index flood model but differences in accuracy seem to be less and less significant as the sample site increases. This may be

related to the white noise introduced in the generation procedure. Indeed, this white noise is independent of the target site sample size and may strongly penalize the performances of the both Bayesian approaches. The next section tries to outline the effect of the target site index flood estimation error to the quantile estimates.

## 6. Effect of Bias on the Target Site Index Flood Estimation

According to the model being considered, two types of biases are encountered for the target site index flood estimation. Indeed, on one hand, the index flood for the  $IFL$  model is derived from the target site sample. On the other hand, for  $BAY$  and  $REV$  approaches, the index flood is estimated from a scaling model. Thus, biases on index flood estimation are due to the relevance of this scaling model but also to the index flood error estimation for the other sites within the region.

To illustrate these two types of biases, the normalized bias on target site index flood estimation is computed as follows:

$$Bias(C) = \frac{\hat{C} - C}{C} \quad (27)$$

where  $C$  is the target site index flood, and  $\hat{C}$  is an estimate of  $C$ . Figure 8 depicts changes in  $NBIAS$  for quantile  $Q_{0.95}$  in function of  $Bias(C)$ . As normalized biases are considered, statistics for the six configurations are plotted in the same graphic. Solid lines correspond to local polynomial regression fits to help underline trends.

Scatter-plots in Figure 8 show clearly these two types of biases. Indeed, on one hand, the range of  $Bias(C)$  is not the same for  $IFL$  than for  $BAY$  and  $REV$ , particularly for

**Table 5.** Changes in  $NMSE$  for  $Q_{0.75}$ ,  $Q_{0.95}$  and  $Q_{0.995}$  in function of the region configuration and the target site sample size for the  $IFL$  estimator. Related standard errors are displayed in brackets.

Model	Heavy Tail			Light Tail		
	<i>Conf1</i>	<i>Conf2</i>	<i>Conf3</i>	<i>Conf4</i>	<i>Conf5</i>	<i>Conf6</i>
Target site sample size 10						
$Q_{0.75}$	0.037 (3e-3)	0.035 (3e-3)	0.037 (3e-3)	0.025 (2e-3)	0.029 (2e-3)	0.029(3e-3)
$Q_{0.95}$	0.038 (3e-3)	0.037 (3e-3)	0.039 (3e-3)	0.027 (4e-3)	0.031 (2e-3)	0.032 (3e-3)
$Q_{0.995}$	0.053 (4e-3)	0.049 (3e-3)	0.049 (4e-3)	0.037 (2e-3)	0.039 (3e-3)	0.042 (4e-3)
Target site sample size 25						
$Q_{0.75}$	0.014 (8e-4)	0.015 (1e-3)	0.015 (1e-3)	0.011 (7e-4)	0.011 (7e-4)	0.011(7e-4)
$Q_{0.95}$	0.018 (1e-3)	0.018 (1e-3)	0.018 (1e-3)	0.014 (9e-4)	0.014 (9e-4)	0.013 (9e-4)
$Q_{0.995}$	0.034 (2e-3)	0.032 (2e-3)	0.027 (2e-3)	0.024 (2e-3)	0.023 (2e-3)	0.020 (1e-3)
Target site sample size 40						
$Q_{0.75}$	0.010 (6e-4)	0.009 (6e-4)	0.010 (6e-4)	0.007 (4e-4)	0.007 (4e-4)	0.007 (5e-4)
$Q_{0.95}$	0.013 (8e-4)	0.013 (8e-4)	0.012 (8e-4)	0.010 (6e-4)	0.009 (5e-4)	0.010 (6e-4)
$Q_{0.995}$	0.028 (2e-3)	0.028 (2e-3)	0.023 (2e-3)	0.020 (1e-3)	0.017 (1e-3)	0.019 (1e-3)

a target site sample size of 25 and 40. On the other hand, for the  $BAY$  and  $REV$  approaches, biases on index flood estimation are independent of the target site sample size; while this is not the case for  $IFL$ . This last point is also illustrated as the bias ranges for the Bayesian approaches remain the same for all target site sample size. Thus, for large sample size, efficiency of the Bayesian estimators may be too much impacted as the artificial bias introduced in the generation procedure is too penalizing.

The Bayesian approaches do not have the same behaviour than the  $IFL$  model. In particular,  $BAY$  and  $REV$  seem to be less sensitive to a large bias in target site index flood estimation.  $NBIAS$  for the  $IFL$  model are clearly linear with a response  $y = x$ . This last point is an expected result. Indeed, apart from sampling variability, if a unique regional distribution exists, quantile  $IFL$  estimate biases are only induced by biases on target site index flood estimates. Thus, the relevance of the generation procedure is corroborated.

The main difference between the  $BAY$  and  $REV$  estimators is the dispersion around local smoothers. Indeed,  $REV$  has a smaller range while preserving the same robustness to the bias on target site index flood estimation.

These results and conclusions are independent of the target site index flood estimation procedure. However, the performance of the two Bayesian estimators is related to the bias and variance of the target site index flood estimate. Thus, for similar variance, these results should be identical if GAMs or Kriging were used.

## 7. Suggestions for Region Configuration

This section attempts to present some suggestions for building suitable pooling groups according to the considered estimator. *Hosking and Wallis* [1997] already advice not to build regions greater than 20 sites because of the small gain affected with additional stations. However, they only focus on the  $IFL$  methodology. We attempt to do the same for the two Bayesian estimators considered in this study. For this purpose, tables 5, 6 and 7 include the  $NMSE$  and the related standard errors for each configuration and target site sample size.

From Table 5, the  $IFL$  estimator seems to have the same performance level independently of the configuration. This result points out that the information is not used optimally as regions with the most information (i.e. *Conf3* and *Conf6*) do not always lead to better estimations. This last point corroborates a previous comments of *Ribatet et al.* [2007].

Table 6 shows that the  $BAY$  estimator is more accurate with “medium” regions, i.e. *Conf3* and *Conf6*. However, results for “small” regions, i.e. *Conf1* and *Conf4*, are often close to the best ones - especially for a light tail. Thus, it is preferable to work with well-instrumented sites, i.e. *Conf1*, *Conf3*, *Conf4* and *Conf6*.

Table 7 shows that the  $REV$  estimator more efficient with “medium” regions, i.e. *Conf3* and *Conf6*. In addition, it

seems to be more accurate with few but well-instrumented gauging stations rather more but less-instrumented ones. Nevertheless for a light tail, all configurations seems to lead to similar performance levels.

Tables 5, 6 and 7 show that the estimation of  $Q_{0.75}$  is independent of the region configuration for all estimators. Thus, it seems that the regional information is not relevant for quantiles with small non exceedence probabilities.

## 8. Conclusions

This article introduced a new Bayesian estimator which uses regional information in an innovative way. The proposed model accounts for a fixed regional shape parameter with a non null probability. Thus, as in *Ribatet et al.* [2007], the regional information is still used to elicit the prior distribution. However, the prior distribution is now a mixture of a GEV/GPD and a GEV/GPD with only two parameters - the remaining one corresponds to the fixed regional shape parameter.

The estimation procedure is achieved using reversible jump Markov chains [*Green*, 1995]; and theoretical details for simulated suited Markov chains were presented. A sensitivity analysis for the proposed algorithm was performed. The results showed that the estimates are consistent provided that the probability attributed to the fixed regional shape parameter is positive. In addition, as noticed by *Stephenson and Tawn* [2004], the credibility intervals are sensitive to this probability value. Thus, the proposed estimator relates this probability value to the homogeneity degree of the region - using the heterogeneity statistic of *Hosking and Wallis* [1997]. Therefore, the credibility intervals take into account the belief about the fixed regional shape parameter to be the true value.

A performance analysis was carried out on stochastic data for three different estimators. For this purpose, another algorithm which generates stochastic homogeneous regions was implemented. The good overall performance of the proposed estimator has been demonstrated. Indeed, on one hand, this approach combines the accuracy of the regional Bayesian approach of *Ribatet et al.* [2007] for quantiles associated to small exceedence probabilities. On the other hand, the duality of the prior distribution (and the fixed regional shape parameter) allows the proposed estimator to be at least as efficient as the index flood model. Thus, this new estimator seems very suited for regional estimation when the target site is not well instrumented.

Furthermore, the two Bayesian approaches considered here appear to be less sensitive to biases on target site index flood estimation than the index flood estimator. Thus, the Bayesian approaches are more readily adaptable which is a major advantage as errors on the index flood estimation are often uncontrollable.

**Table 6.** Changes in  $NMSE$  for  $Q_{0.75}$ ,  $Q_{0.95}$  and  $Q_{0.995}$  in function of the region configuration and the target site sample size for the *BAY* estimator. Related standard errors are displayed in brackets.

Model	Heavy Tail			Light Tail		
	<i>Conf1</i>	<i>Conf2</i>	<i>Conf3</i>	<i>Conf4</i>	<i>Conf5</i>	<i>Conf6</i>
Target site sample size 10						
$Q_{0.75}$	0.015 (1e-3)	0.015 (1e-3)	0.012 (8e-4)	0.011 (6e-4)	0.012 (9e-4)	0.011(9e-4)
$Q_{0.95}$	0.035 (2e-3)	0.063 (4e-3)	0.030 (2e-4)	0.022 (1e-3)	0.037 (3e-3)	0.023 (2e-3)
$Q_{0.995}$	0.101 (1e-2)	0.326 (3e-2)	0.085 (6e-3)	0.054 (5e-3)	0.144 (1e-2)	0.050 (3e-3)
Target site sample size 25						
$Q_{0.75}$	0.010 (6e-4)	0.011 (7e-4)	0.009 (5e-4)	0.008 (5e-4)	0.007 (4e-4)	0.007(5e-4)
$Q_{0.95}$	0.026 (2e-3)	0.041 (3e-3)	0.025 (1e-3)	0.017 (1e-3)	0.023 (1e-3)	0.016 (9e-4)
$Q_{0.995}$	0.089 (8e-3)	0.212 (2e-2)	0.079 (4e-3)	0.044 (3e-3)	0.086 (6e-3)	0.038 (2e-3)
Target site sample size 40						
$Q_{0.75}$	0.008 (5e-4)	0.008 (5e-4)	0.007 (4e-4)	0.005 (3e-4)	0.005 (3e-4)	0.006 (4e-4)
$Q_{0.95}$	0.020 (1e-3)	0.032 (2e-3)	0.020 (1e-3)	0.012 (8e-4)	0.015 (9e-4)	0.013 (8e-4)
$Q_{0.995}$	0.072 (5e-3)	0.187 (2e-2)	0.074 (5e-3)	0.038 (3e-3)	0.070 (6e-3)	0.036 (2e-3)

**Table 7.** Changes in  $NMSE$  for  $Q_{0.75}$ ,  $Q_{0.95}$  and  $Q_{0.995}$  in function of the region configuration and the target site sample size for the *REV* estimator. Related standard errors are displayed in brackets.

Model	Heavy Tail			Light Tail		
	<i>Conf1</i>	<i>Conf2</i>	<i>Conf3</i>	<i>Conf4</i>	<i>Conf5</i>	<i>Conf6</i>
Target site sample size 10						
$Q_{0.75}$	0.014 (1e-3)	0.011 (7e-4)	0.011 (7e-4)	0.011 (6e-4)	0.010 (7e-4)	0.011(9e-4)
$Q_{0.95}$	0.026 (2e-3)	0.024 (2e-3)	0.019 (1e-3)	0.018 (1e-3)	0.016 (1e-3)	0.019 (2e-3)
$Q_{0.995}$	0.047 (3e-3)	0.077 (2e-2)	0.036 (2e-3)	0.033 (2e-3)	0.030 (2e-3)	0.032 (3e-3)
Target site sample size 25						
$Q_{0.75}$	0.009 (6e-4)	0.009 (6e-4)	0.008 (5e-4)	0.008 (5e-4)	0.006 (4e-4)	0.007(5e-4)
$Q_{0.95}$	0.019 (1e-3)	0.020 (2e-3)	0.016 (9e-4)	0.014 (1e-3)	0.014 (9e-4)	0.014 (9e-4)
$Q_{0.995}$	0.040 (3e-3)	0.061 (1e-2)	0.031 (2e-3)	0.026 (2e-3)	0.032 (3e-3)	0.024 (2e-3)
Target site sample size 40						
$Q_{0.75}$	0.008 (5e-4)	0.007 (5e-4)	0.006 (4e-4)	0.005 (3e-4)	0.005 (3e-4)	0.006 (3e-4)
$Q_{0.95}$	0.015 (1e-3)	0.016 (1e-3)	0.012 (9e-4)	0.010 (7e-4)	0.010 (5e-4)	0.011 (6e-4)
$Q_{0.995}$	0.034 (2e-3)	0.055 (1e-2)	0.027 (2e-3)	0.022 (2e-3)	0.023 (2e-3)	0.021 (1e-3)

As noticed by *Ribatet et al.* [2007], the index flood model does not use information optimally. This point is corroborated in this study as the model initiated by *Dalrymple* [1960] is not inevitably more accurate as the information within the pooling group increases. This is not the case for the Bayesian approaches. In addition, they seem to be more accurate when dealing with regions with well instrumented sites, particularly for large quantiles.

All statistical analysis were carried out by use of *R Development Core Team* [2006]. For this purpose, the algorithm presented in section 2.2 was incorporated in the *evdbayes* packages [*Stephenson and Ribatet*, 2006]. The algorithm for the generation procedure is available on request from the author.

**Acknowledgments.** The authors wish to thank Alec Stephenson for providing the original codes of his article. The financial support provided by the National Science and Engineering Research Council of Canada (NSERC) is acknowledged. We are also grateful to the editor, the associate editor and two anonymous referees for useful criticism of the original version of the paper.

## The Metropolis-Hastings Algorithm

In this section, the Metropolis-Hastings algorithm is presented. According to the results derived by *Green* [1995], some details will be given to consider the reversible jump case. The basic idea of the Metropolis-Hastings algorithm is to obtain a Markov chain that converges to a known stationary distribution. The strength of the Metropolis-Hastings approach is that the convergence is reached whatever the initial state of the Markov chain is and that the distributions could be known up to a constant.

Let  $f$  denote the target distribution of interest. Most often, in Bayesian inference,  $\pi$  will be the posterior distribution for the parameters. Let  $q(\cdot, x)$  be the proposal distribution i.e. the proposal states will be sampled from

this proposal distribution given the current state  $x_t$ . The Metropolis-Hastings algorithm can be summarized as follows:

1. Generate  $u$  from a uniform distribution on  $[0, 1]$ ;
2. Generate  $x_{\text{prop}}$  from  $q(\cdot, x_t)$
3.  $\Delta_{\text{class}} \leftarrow \frac{f(x_{\text{prop}}) q(x_t | x_{\text{prop}})}{f(x_t) q(x_{\text{prop}} | x_t)}$
4. **if**  $u < \min(1, \Delta_{\text{class}})$  **then**
5.  $x_{t+1} \leftarrow x_{\text{prop}}$
6. **else**
7.  $x_{t+1} \leftarrow x_t$
8. **endif**
9. Go to 1.

The initial Metropolis-Hastings algorithm can not account for dimensional switch. For this purpose, the ‘‘jumps’’ between sub-spaces must be defined (see equations (15a)–(15c) and (18a)–(18c)) and the quantity  $\Delta_{\text{class}}$  must be redefined each time a jump is considered. Here, only a simple case of the reversible jumps approach is considered (see Section 3.3 of *Green* [1995]). If only two moves  $m_1(x_t)$  and  $m_2(x_t)$  can occur with probabilities  $p_1$  and  $p_2$  respectively, then the quantity  $\Delta_{\text{class}}$  must be replaced by  $\Delta_{\text{rev}}$ . Consequently, for a proposal move of type  $m_1$ :

$$\Delta_{\text{rev}} = \Delta_{\text{class}} \frac{p_1}{p_2} J_1 \quad (\text{A1})$$

where  $J_1$  is the jacobian of the transformation  $x_t \mapsto m_1(x_t)$ . If the proposal move is of type  $m_2$ , then

$$\Delta_{\text{rev}} = \Delta_{\text{class}} \frac{p_2}{p_1} J_2 \quad (\text{A2})$$

where  $J_2$  is the jacobian of the transformation  $x_t \mapsto m_2(x_t)$ .

## References

- Bayes, T. (1763), An essay towards solving a problem in the doctrine of chance, *Philosophical Transaction of the Royal Society*, 53.
- Bortot, P., and S. Coles (2000), The multivariate gaussian tail model: An application to oceanographic data, *Journal of the Royal Statistical Society. Series C: Applied Statistics*, 49(1), 31–49.
- Coles, S., and M. Dixon (1999), Likelihood-based inference for extreme value models, *Extremes*, 2(1), 5–23.
- Coles, S., and J. Tawn (1996), A bayesian analysis of extreme rainfall data, *Journal of the Royal Statistical Society. Series C: Applied Statistics*, 45(4), 463–478.
- Dalrymple, T. (1960), Flood frequency analysis, *U.S. Geol. Surv. Water Supply Pap.*, 1543 A.
- Geman, S., and D. Geman (1984), Stochastic relaxation, gibbs distributions, and the bayesian restoration of images., *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6(6), 721–741.
- Green, P. (1995), Reversible jump markov chain monte carlo computation and bayesian model determination, *Biometrika*, 82, 711–732.
- Hastings, W. K. (1970), Monte carlo sampling methods using markov chains and their applications, *Biometrika*, 57, 97–109.
- Hosking, J., and J. Wallis (1987), Parameter and quantile estimation for the generalized pareto distribution, *Technometrics*, 29(3), 339–349.
- Hosking, J. R. M., and J. R. Wallis (1997), *Regional Frequency Analysis*, Cambridge University Press.
- Juárez, S., and W. Schucany (2004), Robust and efficient estimation for the generalized pareto distribution, *Extremes*, 7(3), 237–251.
- Kuczera, G. (1982), Combining at-site and regional information: An empirical bayes approach, *Water Resources Research*, 18(2), 306–314.
- Madsen, H., and D. Rosbjerg (1997), Generalized least squares and empirical Bayes estimation in regional partial duration series index-flood modeling, *Water Resources Research*, 33(4), 771–781.
- Martins, E., and J. Stedinger (2000), Generalized maximum-likelihood generalized extreme-value quantile estimators for hydrologic data, *Water Resources Research*, 36(3), 737–744.
- McCullagh, P., and J. A. Nelder (1989), *Generalized Linear Models*, Chapman and Hall.
- Merz, R., and G. Blöschl (2005), Flood frequency regionalisation – Spatial proximity vs. catchment attributes, *J. Hydrol.*, 302(1-4), 283–306.
- Northrop, P. (2004), Likelihood-based approaches to flood frequency estimation, *Journal of Hydrology*, 292(1-4), 96–113.
- Pandey, M., P. Van Gelder, and J. Vrijling (2004), Dutch case studies of the estimation of extreme quantiles and associated uncertainty by bootstrap simulations, *Environmetrics*, 15(7), 687–699.
- Park, J.-S. (2005), A simulation-based hyperparameter selection for quantile estimation of the generalized extreme value distribution, *Mathematics and Computers in Simulation*, 70(4), 227–234.
- Payer, T., and H. Kuchenhoff (2004), Modelling extreme wind speeds at a german weather station as basic input for a subsequent risk analysis for high-speed trains, *Journal of Wind Engineering and Industrial Aerodynamics*, 92(3-4), 241–261.
- Pickands, J. I. (1975), Statistical inference using extreme order statistics, *Annals of Statistics*, 3, 119–131.
- R Development Core Team (2006), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.
- Ribatet, M., E. Sauquet, J.-M. Grésillon, and T. B. M. J. Ouarda (2007), A regional bayesian pot model for flood frequency analysis, *Stochastic Environmental Research and Risk Assessment (SERRA)*, 21(4), 327–339.
- Rosbjerg, D., H. Madsen, and P. Rasmussen (1992), Prediction in partial duration series with generalised pareto- distributed exceedances, *Water Resources Research*, 28(11), 3001–3010.
- Seidou, O., T. Ouarda, M. Barbet, P. Bruneau, and B. Bobée (2006), A parametric bayesian combination of local and regional information in flood frequency analysis, *Water Resour. Res.*, 42(11), W11408.
- Shu, C., and D. H. Burn (2004), Artificial neural network ensembles and their application in pooled flood frequency analysis, *Water Resources Research*, 40(9), W09301.
- Stephenson, A., and M. Ribatet (2006), *A User's Guide to the evdbayes Package (Version 1.1)*.
- Stephenson, A., and J. Tawn (2004), Bayesian inference for extremes: Accounting for the three extremal types, *Extremes*, 7(4), 291–307.
- Wood, S., and N. Augustin (2002), GAMs with integrated model selection using penalized regression splines and applications to environmental modelling, *Ecol. Model.*, 157(2-3), 157–177.

M. Ribatet, Unité de Recherche HH, Cemagref Groupement de Lyon, 3bis quai Chauveau CP220, 69336 Lyon Cedex 09, FRANCE. (ribatet@lyon.cemagref.fr)

## 6.4 Discussion

Ce modèle [Ribatet *et al.*, 2007c] a permis d'étendre le domaine d'estimation du modèle précédent [Ribatet *et al.*, 2007b]. En effet, nos travaux montrent que ce nouveau modèle conserve les mêmes performances que le modèle précédent pour les quantiles ayant une période de retour  $T \leq 20$  ans. En revanche, il est nettement plus efficace lorsque  $T > 20$  ans. Pour les sites très peu jaugés (moins de 5 années d'enregistrement), il est bien plus performant que le modèle de l'indice de crue. Pour un plus grand nombre de données au site cible, le modèle à sauts réversibles est au moins aussi performant que celui de l'indice de crue.

Ces résultats sont donc très encourageants puisque nous obtenons, finalement, un estimateur performant sur toute la gamme des périodes de retour et permettant l'estimation des débits de pointe mais aussi les incertitudes associées. De plus, ce modèle a été validé sur des données simulées et en particulier favorables à l'estimateur proposé par Dalrymple [1960]. Il est donc intéressant que ce nouvel estimateur soit plus performant que le modèle de Dalrymple sur des données le favorisant.

Toutefois, ce modèle n'est pas rigide et de nombreuses améliorations sont possibles. Notamment, toutes les critiques faites sur le modèle *BAY* lors de la Section 5.4 restent valides ; à savoir la prise en compte de la dépendance sur les paramètres de la GPD et sur les sites de la région homogène. Ainsi, toutes améliorations sur ces points sensibles bénéficieraient aux modèles *BAY* et *REV* simultanément.

En ce qui concerne la particularité de ce modèle, un point sensible est le choix de la paramétrisation de l'hyperparamètre supplémentaire  $p_\xi$ . Pour notre application, nous avons fait un choix raisonnable et peu risqué - *cf.* Figure 6.1. Toutefois, les performances de ce modèle sont bien évidemment très sensibles à cette paramétrisation et une étude de sensibilité, plus complète que celle réalisée pour l'article, devrait être menée sur :

- la paramétrisation pour  $p_\xi$  ;
- la définition des sauts réversibles.

Une telle étude permettrait de mieux cerner le comportement global de notre estimateur, mais aussi éventuellement de proposer des paramétrisations plus pertinentes.

Enfin, un point faible de ce modèle concerne la gestion de la dépendance entre le paramètre de forme  $\xi$  et d'échelle  $\sigma$ . Ce n'est pas nouveau et nous avons déjà fait la remarque pour le modèle initial - *cf.* Section 5.4. Toutefois, vu que nous proposons un candidat particulier  $\xi_{\text{Fix}}$  pour le paramètre de forme, il devrait donc avoir une « zone candidate » toute aussi particulière pour le paramètre d'échelle. En particulier, la définition des sauts réversibles permettant la transition entre les espaces  $\Theta \setminus \Theta_0$  et  $\Theta_0$  joue également un rôle important. Une étude de sensibilité aux choix des sauts pourrait donc aussi être considérée.





# 7

## Modélisation de Tous les Excès Selon une Structure de Dépendance Extrême

DANS cette section, nous sortons du cadre de l'estimation régionale pour revenir dans un contexte plus classique : l'estimation **purement locale**. Comme mentionné lors du Chapitre 2, nous entendons par purement locale l'estimation à partir de la seule chronique de débit disponible au site cible. Pourquoi revenir dans un cadre d'estimation purement locale alors que les deux modèles régionaux précédents se sont avérés pertinents ? Pour y répondre, nous reprenons la notion d'utilisation de l'information disponible.

L'élément justifiant le développement de deux nouveaux modèles régionaux vient du constat que le modèle de référence, l'indice de crue, n'utilisait pas l'information (plus particulièrement celle du site cible) de manière rationnelle et efficace. Nos deux modèles précédents ont permis de différencier l'information au site cible, la plus précieuse, de l'information régionale pure. Néanmoins, a-t-on réellement utilisé toute l'information dont nous disposons au site d'étude ?

### 7.1 Une Utilisation Totale de l'Information Purement Locale

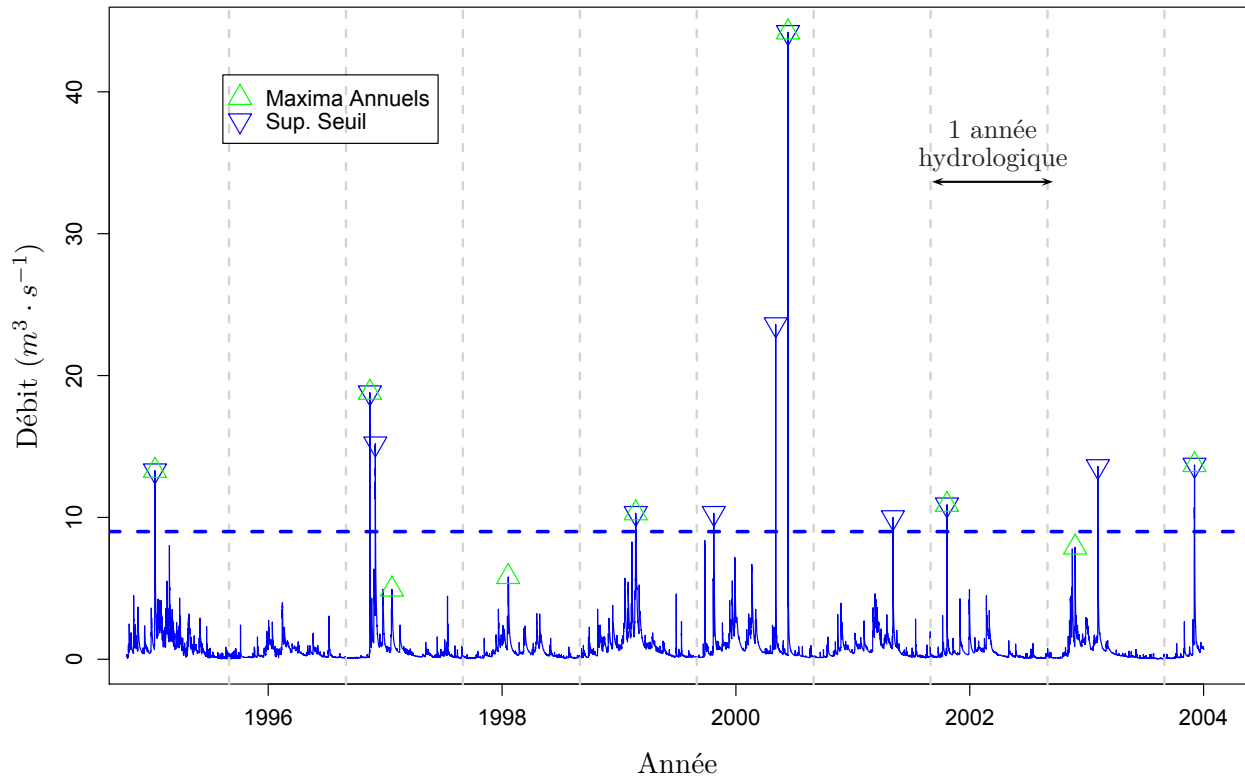
Lors d'une analyse des valeurs extrêmes, le praticien s'intéresse uniquement aux maxima annuels, voire dans les meilleurs cas, à des dépassements (jugés indépendants) au dessus d'un seuil suffisamment élevé. La Figure 7.1 illustre la différence d'utilisation des données disponibles pour ces deux approches.

Toutefois, il est nécessaire de rappeler que ces données extrêmes sont loin d'être les données de base, mais sont en fait un sous-échantillon de la chronique de débit initiale. En particulier, le passage d'une série quasiment continue en un échantillon réduit de valeurs de débit entraîne, de fait, une perte d'information non négligeable.

Ainsi, il nous est venu à l'esprit de proposer une approche limitant au maximum cette perte d'information sur les données du site cible. Cette approche se justifie de la manière suivante : quelle que soit la qualité de l'information supplémentaire (*i.e.* historique, GRADEX, régionale, ...), il paraît raisonnable de penser qu'elle ne sera jamais aussi fiable que celle dont nous disposons au site d'étude. Dès lors, avant de rechercher de l'information extérieure, il paraît pertinent d'utiliser l'information purement locale de manière efficace.

Le modèle que nous introduisons dans cette section est une généralisation de l'approche POT. Lors de la réduction de la série temporelle en un échantillon restreint, l'approche POT s'assure d'obtenir des observations indépendantes en imposant des contraintes d'espacement temporel et de redescente par rapport au débit de pointe [Davison et Smith, 1990; Bayliss et Jones, 1993; Lang *et al.*, 1999] : c'est l'étape de *declusterisation*.

Nous proposons de relâcher la nécessité d'indépendance entre les valeurs supérieures à seuil en introduisant un modèle qui prendrait en compte la dépendance potentielle entre ces dépassements. La Figure 7.2 explicite la différence d'utilisation de la chronique de débit entre l'approche POT



**Fig. 7.1:** Perte d'information : passage d'une chronique de débit à un échantillon de taille réduite. Approches classiques des maxima annuels et dépassements au dessus d'un seuil. L'Ardières à Beaujeu.

et celle que nous proposons. Seul le maxima de chaque *cluster* est retenu lors d'une analyse POT tandis que sur cet exemple, 19 points sont exploités avec l'approche que nous proposons.

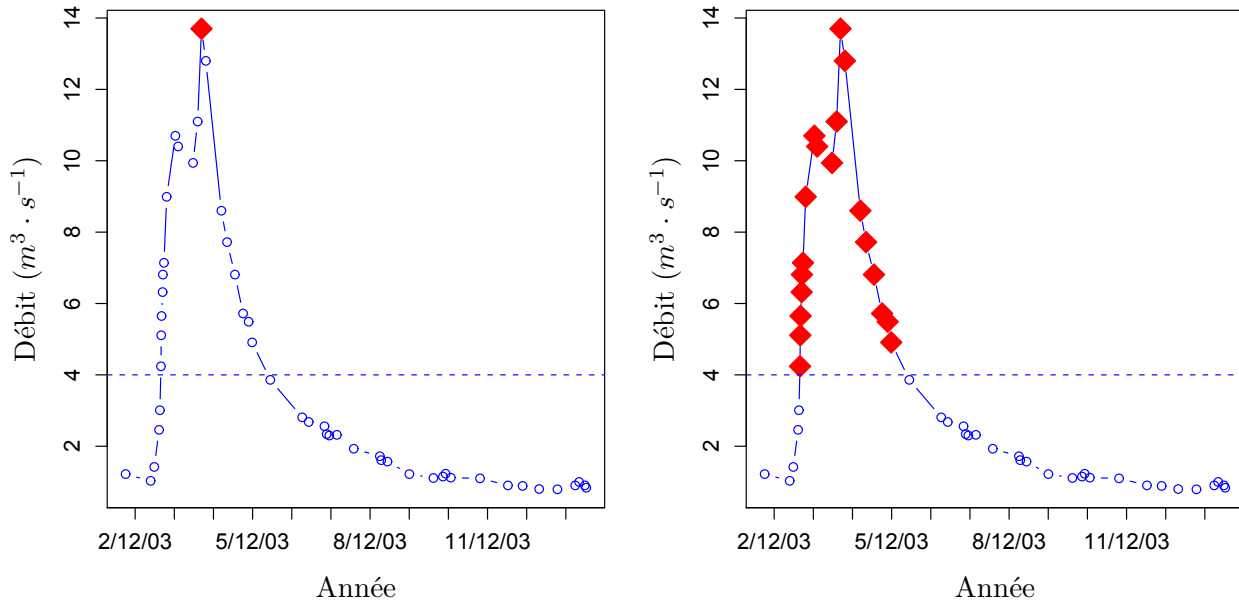
La modélisation de tous les excès apporte un avantage direct par rapport à l'approche POT. Elle évite notamment la procédure de declusterisation. Cette declusterisation est une étape sensible et souvent subjective. Notons néanmoins que [Ferro et Segers \[2003\]](#) ont dernièrement proposé une technique de declusterisation *quasi automatique* basée sur la distribution théorique du laps de temps écoulé entre deux dépassements - dépendants ou non. Toutefois, à notre connaissance, cette méthodologie n'a pas encore été utilisée pour les chroniques de débit. Notre retour d'expérience sur cette approche montre que cela conduit à des échantillons parfaitement cohérents avec ceux déduits des méthodes décrites par [Davison et Smith \[1990\]](#), [Bayliss et Jones \[1993\]](#) et [Lang et al. \[1999\]](#).

Un autre avantage de la modélisation de tous les excès résulte dans la taille nettement plus importante de l'échantillon servant à l'ajustement. Toutefois, n'oublions pas que cela apporte un degré de difficulté supplémentaire puisqu'il faut parvenir à modéliser la dépendance entre les observations. Il reste donc à vérifier que l'utilisation d'un échantillon étendu comportant des valeurs dépendantes apporte un réel gain pour l'estimation des quantiles.

La modélisation de la dépendance temporelle entre les débits successifs permet également d'élargir les variables d'études. En effet, la connaissance de cette structure de dépendance renseigne sur la dynamique des crues ; tout comme la dimension temporelle ainsi modélisée peut être utile à la construction d'hydrogrammes de projet.

## 7.2 La Modélisation de la Dépendance

L'utilisation de la théorie des valeurs extrêmes pour la modélisation des extrêmes est maintenant bien établie et reconnue. Néanmoins, les résultats dans le cadre multivarié sont très méconnus dans



**Fig. 7.2:** Différence d'utilisation de l'information entre l'approche classique POT (à gauche) et la modélisation de tous les excès (à droite). L'Ardières à Beaujeu. Losange rouge : observation exploitée lors de la procédure d'estimation.

les sciences environnementales et restent le domaine de garde des journaux plus théoriques. Notamment, en hydrologie, la théorie multivariée des valeurs extrêmes (**MEVT**) est souvent délaissée au profit des approches par copules [Favre *et al.*, 2004]. Bien que ces deux approches admettent des points de convergence, les copules permettent l'extension aux domaines multivariés indépendamment du caractère extrême des données. En revanche, la MEVT est spécifique à la modélisation en dimensions supérieures des extrêmes ; autant l'utiliser.

Pour notre modèle, nous nous servons de cette théorie afin de modéliser la dépendance entre les observations successives. En particulier, cette dépendance est intégrée au sein d'une chaîne de Markov d'ordre un :

$$\ell(y_1, \dots, y_n) = f(y_1) \prod_{i=2}^n f(y_i | y_{i-1}) = \frac{\prod_{i=2}^n f(y_{i-1}, y_i)}{\prod_{i=2}^{n-1} f(y_i)} \quad (7.1)$$

où  $f(y_i)$  est la densité marginale,  $f(y_i | y_{i-1})$  la densité conditionnelle,  $f(y_{i-1}, y_i)$  la densité jointe et  $\ell(y_1, \dots, y_n)$  la vraisemblance évaluée pour les observations  $y_1, \dots, y_n$ .

En particulier, la densité marginale est définie en ayant recours à la théorie des valeurs extrêmes univariée. La densité jointe est, quant à elle, définie à partir des résultats multivariés.

Du fait de la densité jointe évaluée pour le couple  $(y_{i-1}, y_i)$ , nous voyons apparaître une difficulté. En effet, le couple  $(y_{i-1}, y_i)$  peut appartenir à l'une des quatre régions définies par la Figure 7.3. D'après la théorie multivariée des valeurs extrêmes, nous ne pouvons utiliser une loi bivariée des valeurs extrêmes que pour la région *A*. Toutefois, à l'aide de l'approche par vraisemblance censurée [Ledford et Tawn, 1996], nous pouvons évaluer la vraisemblance (7.1) pour les trois régions restantes *B*, *C* et *D*.

La *censure en dessous d'un seuil  $u$*  consiste à dire que l'on connaît si une observation  $y_i$  est inférieure à  $u$  mais que l'on ne connaît pas sa valeur exacte. Pour notre application, c'est un peu différent puisque nous connaissons bel et bien la valeur  $y_i$ . Par contre, nous ne connaissons pas la distribution de probabilité régissant  $Y_i$ ,  $Y_i \leq u$  puisque nous ne sommes plus dans la région où l'approximation par une GPD est supposée valide. Ainsi, la densité évaluée au point  $y_i$ ,  $y_i \leq u$  est

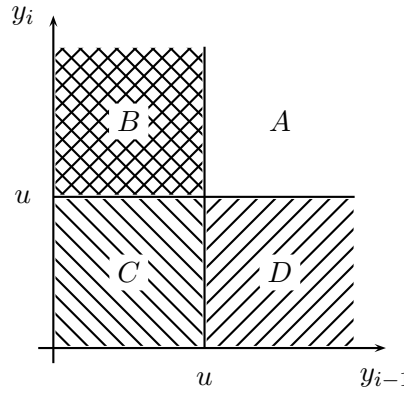


Fig. 7.3: Définition des 4 différentes régions du couple  $(y_{i-1}, y_i)$ .

donnée par :

$$f(y_i) = \Pr[Y \leq u] \quad (7.2)$$

De même, l'extension de la censure au couple  $(y_i, y_{i-1})$  est donnée par :

$$f(y_{i-1}, y_i) = \begin{cases} \frac{\partial^2 F}{\partial y_{i-1} \partial y_i} |_{(y_{i-1}, y_i)}, & (y_{i-1}, y_i) \in A \\ \frac{\partial F}{\partial y_i} |_{(u, y_i)}, & (y_{i-1}, y_i) \in B \\ F(u, u), & (y_{i-1}, y_i) \in C \\ \frac{\partial F}{\partial y_{i-1}} |_{(y_{i-1}, u)}, & (y_{i-1}, y_i) \in D \end{cases} \quad (7.3)$$

La vraisemblance censurée permet donc de modéliser toutes les observations de la chronique de débit. De sorte que, finalement, toute l'information disponible localement est utilisée dans la procédure d'estimation. C'est donc une différence considérable par rapport aux approches de type maxima annuels et POT.

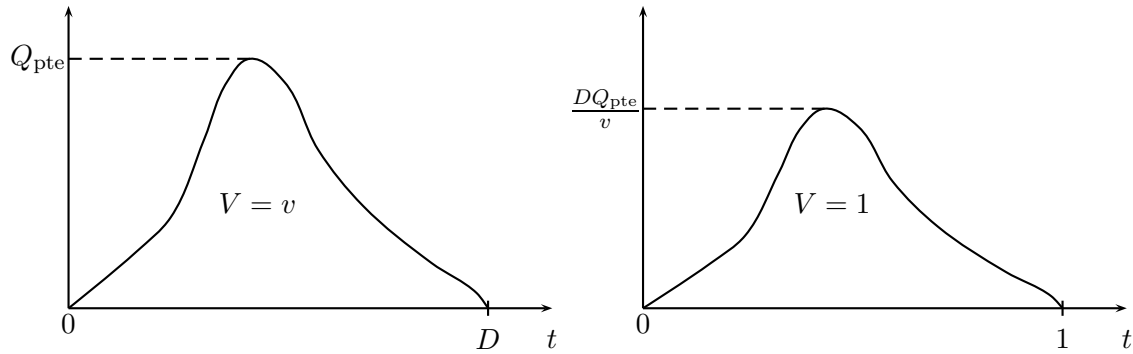
De plus, du fait de l'utilisation de la théorie des valeurs extrêmes multivariée, le modèle impose une structure de dépendance adaptée aux extrêmes pour tous les points appartenant aux régions A, B et D. Pour la région C, la structure de dépendance des extrêmes n'étant plus justifiée, nous utilisons la seule information possible *i.e.* les deux points  $y_{i-1}$  et  $y_i$  sont inférieurs au seuil  $u$ .

## 7.3 La Modélisation des Hydrogrammes Rendue Possible

Le modèle introduit dans cette section se distingue des autres modèles utilisés classiquement en hydrologie (*cf.* Chapitre 2) pour l'estimation des quantiles de crue puisqu'il utilise toutes les observations disponibles au sein de la chronique de débit - *cf.* Équation (7.1). En conséquence, il est désormais possible de simuler des chroniques de débits à partir de ce modèle et d'avoir, par exemple, une estimation de la dynamique des crues. Nous présentons auparavant deux approches actuellement disponibles pour estimer la dynamique des crues.

### 7.3.1 Deux Approches pour l'Estimation de la Dynamique des Crues

Comme nous l'avons présenté en Introduction, la seule connaissance des débits de pointe n'est pas toujours suffisante. L'estimation du volume, de la durée mais aussi de la forme des hydrogrammes peut également s'avérer utile en entrée des modélisations hydrauliques [Yue *et al.*, 2002]. Avant de présenter la manière dont le modèle (7.1) permet d'estimer la dynamique des crues, nous présentons, dans cette section, deux approches permettant de construire des hydrogrammes de projet.



**Fig. 7.4:** Illustration de la description des hydrogrammes de crue par une densité de probabilité. Panneau de gauche : hydrogramme brut. Panneau de droite : hydrogramme modifié et d'aire unité. Figure inspirée de [Yue et al. \[2002\]](#).

### Modélisation des Hydrogrammes de Crue

Plusieurs auteurs se sont intéressés à la description de la forme des hydrogrammes. Ces approches imposent une forme géométrique préétablie pour représenter les crues. La diversité des formes géométriques observées est alors déterminée par un modèle mathématique imposé dont le nombre de paramètres à estimer est également contraint.

Par exemple, [Sackl et Bergmann \[1987\]](#) et [Bertoli et Moisello \[1994\]](#) proposent respectivement les paramétrisations suivantes :

$$Q(t) = Q_{\text{pte}} \left\{ \left( \frac{t}{t_m} \right)^\beta \exp \left[ 1 - \left( \frac{t}{t_m} \right)^\beta \right] \right\}^n \quad (7.4)$$

$$Q(t) = \begin{cases} Q_b + (Q_{\text{pte}} - Q_b) \left( \frac{t}{t_m} \right)^n, & t \leq t_m \\ Q_b + (Q_{\text{pte}} - Q_b) \exp \left[ - \left( \frac{t-t_m}{k} \right) \right], & t > t_m \end{cases} \quad (7.5)$$

où  $Q_b$  correspond au débit de base,  $t_m$  au temps de montée et  $\beta$ ,  $n$  et  $k$  sont les paramètres des modèles.

Dernièrement, l'utilisation de lois de probabilité pour estimer la dynamique de crue a emporté un large succès [[Haktanir et Sezen, 1990](#); [Yue et al., 2002](#); [Bhunya et al., 2004](#)]. Contrairement aux approches précédentes, l'utilisation de lois de probabilité facilite grandement la procédure d'estimation en rendant celle-ci plus robuste numériquement et assise sur des bases théoriques établies [[Yang et Han, 2006](#)]. La Figure 7.4 schématise la procédure pouvant contraindre un hydrogramme de crue observé en un hydrogramme d'aire unité ; autorisant ainsi sa modélisation par une densité de probabilité.

Parmi ces approches, [Haktanir et Sezen \[1990\]](#) et [Yue et al. \[2002\]](#) reproduisent la forme des hydrogrammes en ajustant une loi Beta et Gamma à deux paramètres. [Bhunya et al. \[2004\]](#) suggèrent une modélisation par la loi Beta à trois paramètres. Dernièrement, [Bhunya et al. \[2007\]](#) comparent la performance des lois Gamma, du  $\chi^2$ , de Weibull et Beta pour la reconstitution des hydrogrammes de crue. Ils montrent que les distributions Beta et Weibull paraissent les plus aptes à reproduire la dynamique des crues.

Bien souvent, lors de la construction d'un hydrogramme de projet, le praticien devra multiplier l'hydrogramme unitaire par une constante de sorte que le débit de pointe (*resp.* le volume) soit égal au débit de pointe (*resp.* volume) de la période de retour d'intérêt [[Yue et al., 2002](#)]. Ainsi, indépendamment du modèle mathématique ou statistique décrivant la forme des crues, ce type d'approche s'appuie bien souvent sur la théorie de l'hydrogramme unitaire [[Sherman, 1932](#)] - approche supposant que tous les hydrogrammes sont identiques à un paramètre multiplicatif près.

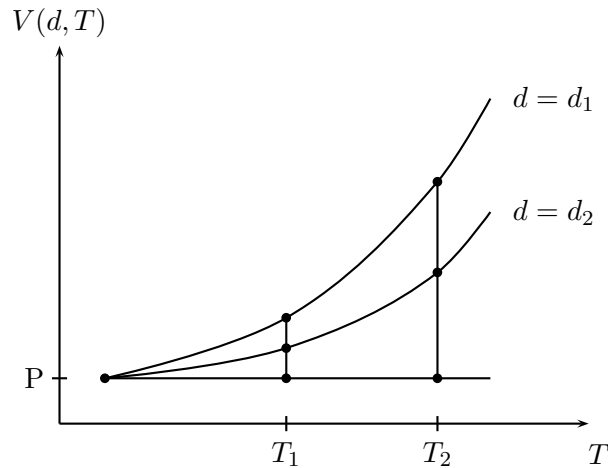


Fig. 7.5: Convergence des distributions des débits moyens associés aux durées  $d_1$  et  $d_2$ .

### Estimation des Débits Moyens et d'une Durée Caractéristique de Crue

Le Cemagref est à l'origine du modèle QdF autrement appelé modèle Débit Durée Fréquence. Les derniers développements de ce modèle [Javelle, 2001] permettent d'obtenir l'estimation des débits de pointe, des débits moyens  $V(d, T)$  associés à la durée  $d$  (cf. Figure 1.3), mais aussi l'estimation d'une durée caractéristique de crue  $\Delta$ . Il existe également une version du modèle QdF pour les débits continuellement dépassés  $Q(d, T)$ . Ce modèle repose sur deux hypothèses validées empiriquement :

1. les distributions des débits relatives à chaque durée convergent en un unique point (cf. Figure 7.5)
2. les distributions des débits relatives à chaque durée admettent une forme hyperbolique dépendant de la durée  $d$  et de la distribution des débits de pointe

$$V(d, T) = \frac{Q(0, T) - P}{1 + \frac{d}{\Delta}} + P \quad (7.6a)$$

$$Q(d, T) = \frac{Q(0, T) - P}{\left(1 + \frac{d}{\Delta}\right)^2} + P \quad (7.6b)$$

où  $Q(0, T)$  est la distribution des débits de pointe,  $V(d, T)$  (resp.  $Q(d, T)$ ) la distribution des débits moyens (resp. continuellement dépassés) associés à la durée  $d$ ,  $T$  la période de retour,  $P$  l'ordonnée d'affinité des distributions et  $\Delta$  une durée caractéristique de crue. En pratique,  $P$  est souvent considéré comme nul.

La force du modèle QdF est de fournir, au travers de l'estimation des débits de pointe, moyens et d'une durée caractéristique de crue, une appréciation de la dynamique des crues pour le bassin considéré.

Cependant, plusieurs critiques peuvent être émises. Premièrement, le modèle QdF mêle les notions de probabilité au sein d'un modèle empirique mais ajusté sans tenir compte de la vraisemblance. Notamment, les paramètres du modèle sont ajustés en deux temps [Javelle, 2001] : (a) estimation de la loi de  $Q(0, T)$  à partir des observations adimensionnalisées et (b) estimation de  $\Delta$  à partir d'un problème d'optimisation sachant les paramètres de la loi de  $Q(0, T)$ .

Le deuxième point sensible du modèle QdF concerne la durée caractéristique  $\Delta$ . Si la connaissance d'une durée caractéristique de crue est tout aussi essentielle que celle des débits de pointe, le modèle QdF en l'état actuel ne modélise pas cette durée comme une variable aléatoire. Si ce modèle propose une réelle avancée, il reste limité quant à la dynamique des crues en la réduisant

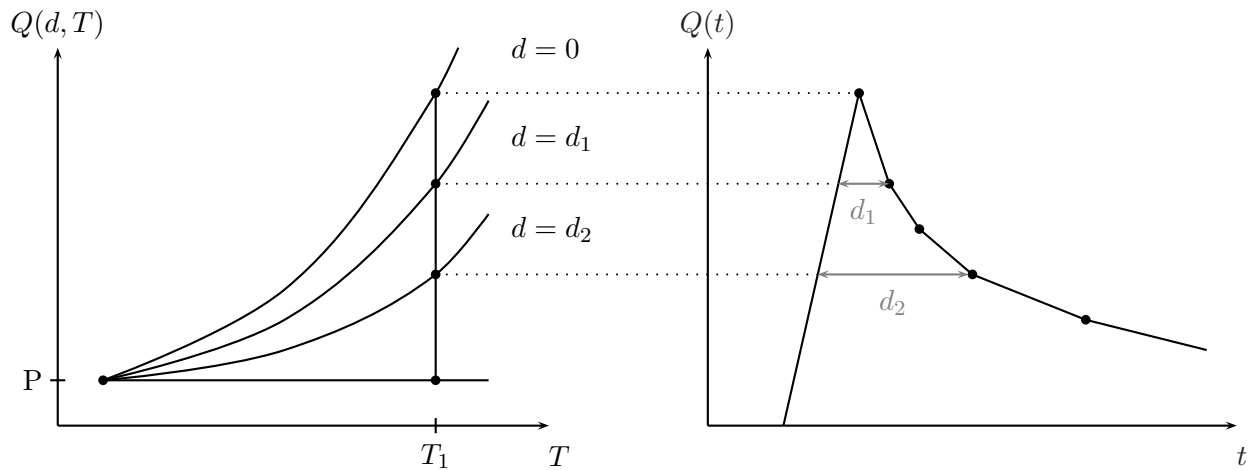


Fig. 7.6: Construction de l'HSMF à partir des courbes QdF des débits continuellement dépassés.

à une unique valeur. En effet, il paraîtrait aberrant de réduire les débits de pointe à une unique valeur. Il en est de même pour la durée caractéristique de crue  $\Delta$  qui est généralement fonction des débits de pointe. Même si la variable  $\Delta$  est intrinsèquement moins variable que les débits, sa variabilité doit être prise en compte puisqu'elle peut influencer de manière non négligeable la forme des hydrogrammes.

Le modèle QdF autorise la construction de crue de projet dit Hydrogrammes Synthétiques Mono Fréquence (**HSMF**) [Sauquet et al. \[2003\]](#). L'HSMF présente l'avantage, par construction, d'être homogène par rapport à une période de retour  $T$  fixée. La montée de l'HSMF est supposé linéaire; alors que la redescente est telle que le débit maximal  $Q_{\text{dep}}$  dépassé en continu sur la durée  $d$  est égal au quantile du débit seuil associé à cette même durée  $d$  et de période de retour  $T$  - *i.e.*  $Q(d, T)$ . La Figure 7.6 illustre la construction d'un HSMF à partir des courbes QdF des débits continuellement dépassés. Implicitement, lorsque  $P = 0$ , les familles d'HSMF répondent aux hypothèses de l'hydrogramme unitaire.

### 7.3.2 Les Possibilités Offertes par le Modèle Utilisant Tous les Excès

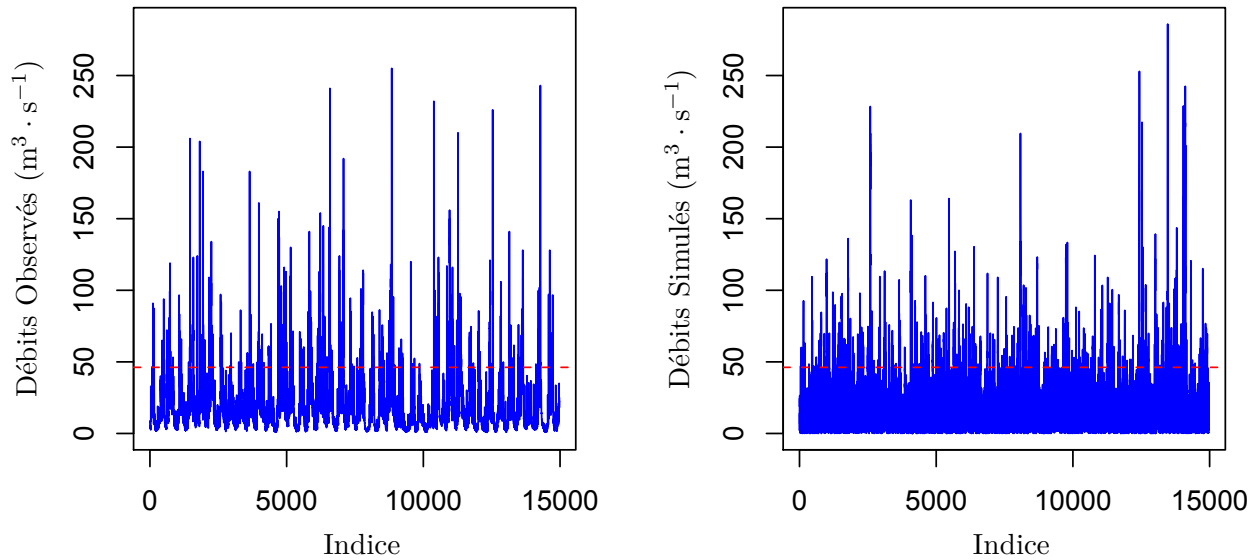
Le modèle présenté dans ce chapitre fut initialement proposé pour utiliser au maximum l'information disponible au site d'étude et, bien entendu, obtenir des estimations des débits caractéristiques de crue plus précises. Néanmoins, par construction du modèle, nous pouvons obtenir plus que les seules estimations des débits caractéristiques et ainsi proposer une alternative aux méthodes introduites plus haut.

À partir de l'Équation (7.1), il est possible de mieux connaître la dynamique des crues mais également de simuler des chroniques de débit. La Figure 7.7 représente une chronique de débit de la Sioule à St-Priest-des-Champs et celle simulée à partir de notre modèle. Deux constats peuvent être établis.

Premièrement, le modèle semble bien reproduire les plus forts débits; notamment l'ordre de grandeur semble préservé. En revanche, ce n'est visiblement pas le cas pour les débits de moyennes eaux - *i.e.* inférieur au seuil de  $46 \text{ m}^3 \cdot \text{s}^{-1}$ . Ces remarques ne sont pas surprenantes puisque la structure de dépendance imposée pour les observations consécutives est issue de la MEVT et donc spécifiques aux extrêmes. Il n'est donc pas étonnant que : (a) les débits extrêmes semblent bien reproduits (*cf.* Figure 7.8) et (b) que ce ne soit pas du tout le cas pour les débits fréquents. Néanmoins, lorsque l'analyse fréquentielle ne porte que sur les plus fortes valeurs, cette mauvaise performance disparaît.

A partir de ces chroniques de débits simulées, il est donc possible d'obtenir une estimation de





**Fig. 7.7:** Chronique de débits journaliers observés (à gauche) et simulés (à droite). La Sioule à St-Priest-des-Champs. Seuil à  $46 \text{ m}^3 \cdot \text{s}^{-1}$ .

la dynamique des crues, mais aussi des hydrogrammes probabilisés. Ce modèle nous a paru innovant puisqu'il mêle la vision fréquentielle du modèle QdF, mais aussi la vision probabiliste des hydrogrammes. De plus, contrairement aux approches présentées lors de la Section 7.3.1, les hydrogrammes se déduisent directement de la chronique simulée et non d'un modèle mathématique plus ou moins subjectif. Certes, un modèle mathématique est toujours imposé (chaîne de Markov d'ordre un, structure de dépendance extrême). Toutefois, il est imposé sur les dépassements successifs et non sur l'hydrogramme tout entier. De plus, la structure de dépendance extrême est justifiée théoriquement par la MEVT ce qui n'est pas le cas des paramétrisations proposées pour la modélisation des hydrogrammes.

## 7.4 L'Article

# Modeling All Exceedances Above a Threshold Using an Extremal Dependence Structure: Inferences on Several Flood Characteristics

Mathieu Ribatet<sup>\*,†</sup>    Taha B.M.J. Ouarda<sup>†</sup>    Eric Sauquet<sup>\*</sup>    Jean-Michel Grésillon<sup>\*</sup>

Submitted to: *Water Resources Research*

<sup>\*</sup> Cemagref Lyon, Unité de Recherche Hydrologie-Hydraulique, 3 bis quai Chauveau, CP220, 69336 Lyon cedex 09, France

<sup>†</sup> INRS-ETE, University of Québec, 490, de la Couronne Québec, Qc, G1K 9A9, CANADA.

Corresponding author: M. Ribatet; Email: ribatet@lyon.cemagref.fr

Phone: +33 4 72 20 87 64; Fax: +33 4 78 47 78 75

## Abstract

Flood quantile estimation is of great importance for many engineering studies and policy decisions. However, practitioners must often deal with small data available. Thus, the information must be used optimally. In the last decades, to reduce the waste of data, inferential methodology has evolved from annual maxima modeling to peaks over a threshold one. To mitigate the lack of data, peaks over a threshold are sometimes combined with additional information - mostly regional and historical information. However, whatever the extra information is, the most precious information for the practitioner is found at the target site. In this study, a model that allows inferences on the whole time series is introduced. In particular, the proposed model takes into account the dependence between successive extreme observations using an appropriate extremal dependence structure. Results show that this model leads to more accurate flood peak quantile estimates than conventional estimators. In addition, as the time dependence is taken into account, inferences on other flood characteristics can be performed. An illustration is given on flood duration. Our analysis shows that the accuracy of the proposed models to estimate the flood duration is related to specific catchment characteristics. Some suggestions to increase the flood duration predictions are introduced.

## 1 Introduction

Estimation of extreme flood events is an important stage for many engineering designs and risk management. This is a considerable task as the amount of data available is often small. Thus, to increase the precision and the quality of the estimates, several authors use extra information in addition to the target site one. For example, Ribatet et al. [2007a] Kjeldsen and Jones [2007], Kjeldsen and Jones [2006] and Cunderlik and Ouarda [2006] add information from other homogeneous gaging stations. Werritty et al. [2006] and Reis Jr. and Stedinger [2005] use historical information to improve inferences. Incorporation of extra information in the estimation procedure is attractive but it should not be more prominent than the original data [Ribatet et al., 2007b]. Before looking at other kinds of information, it seems reasonable to use efficiently the one available at the target site. Indeed, practitioners have initially the whole time series, not only the extreme observations. In particular, it is a considerable waste of information to reduce a time series to a sample of Annual Maxima (**AM**).

In this perspective, the Peaks Over Threshold (**POT**) approach is less wasteful than the AM one as more than one event per year could be inferred. However, the declustering method used to identify independent events is quite subjective. Furthermore, even though a “quasi automatic” procedure was recently introduced by Ferro and Segers [2003], there is still a waste of information as only cluster maxima are used.

Ledford and Tawn [1996] propose an innovative approach using Markov chain models that use all exceedances and accounts for temporal dependence between successive observations. Finally, the entire information available within the time series is taken into account. Fawcett and Walshaw [2006] give an illustrative application of the Markov chain model to extreme wind speed modeling.

In this study, extreme flood events are of interest. The performance of the Markov chain model is compared to the conventional POT approach. The data analyzed consist of a collection of 50 French gaging stations. The area under study ranges from 2°W to 7°E and from 45°N to 51°N. The drainage areas vary from 72 to 38300 km<sup>2</sup> with a median value of 792 km<sup>2</sup>. Daily observations were recorded from 39 to 105 years, with a mean value of 60

years. For the remainder of this article, the quantile benchmark values are derived from the maximum likelihood estimates on the whole times series using a conventional POT analysis.

The paper is organized as follows. Section 2 introduces the theoretical aspects for the Markov chain model, while Section 3 checks the relevance of the Markovian model hypothesis. Section 4 and 5 analyze the performance of the Markovian model to estimate the flood peaks and durations respectively. Finally, some conclusions and perspectives are drawn in Section 6.

## 2 A Markov Chain Model for Cluster Exceedances

In this section, the extremal Markov chain model is presented. In the remainder of this article, it is assumed that the flow  $Y_t$  at time  $t$  depends on the value  $Y_{t-1}$  at time  $t-1$ . The dependence between two consecutive observations is modeled by a first order Markov chain. Before introducing the theoretical aspects of the model, it is worth justifying and describing the main advantages of the proposed approach.

It is now well-known that the univariate Extreme Value Theory (**EVT**) is relevant when modeling either AM or POT. Nevertheless, its extension to the multivariate case is surprisingly rarely applied in practice. This work aims to motivate the use of the multivariate EVT. In our application, the multivariate results are used to model the dependence between a set of lagged values in a times series. Consequently, compared to the AM or the POT approaches, the amount of observations used in the inference procedure is clearly larger. Indeed, with the Markovian model, all exceedances are inferred while only cluster maxima are used in the POT approach. Thus, if the first order Markov chain and the dependence structure is suited, the Markovian models are expected to be more accurate.

### 2.1 Likelihood function

Let  $Y_1, \dots, Y_n$  be a stationary first-order Markov chain with a joint distribution function of two consecutive observations  $F(y_1, y_2)$ , and  $F(y)$  its marginal distribution. Thus, the likelihood function  $L$  evaluated at points  $(y_1, \dots, y_n)$  is:

$$L(y_1, \dots, y_n) = f(y_1) \prod_{i=2}^n f(y_i | y_{i-1}) = \frac{\prod_{i=2}^n f(y_i, y_{i-1})}{\prod_{i=2}^{n-1} f(y_i)} \quad (1)$$

where  $f(y_i)$  is the marginal density,  $f(y_i | y_{i-1})$  is the conditional density, and  $f(y_i, y_{i-1})$  is the joint density of two consecutive observations.

To model all exceedances above a sufficiently large threshold  $u$ , the joint and marginal densities must be known. Standard univariate EVT arguments [Coles, 2001] justify the use of a Generalized Pareto Distribution (**GPD**) for  $f(y_i)$ . Similarly, multivariate EVT arguments [Resnick, 1987] argue for a bivariate extreme value distribution for  $f(y_i, y_{i-1})$ .

As a consequence, the marginal distribution is defined by:

$$F(y) = 1 - \lambda \left( 1 + \xi \frac{y - u}{\sigma} \right)_+^{-1/\xi}, \quad y \geq u \quad (2)$$

where  $x_+ = \max(0, x)$ ,  $\lambda = \Pr[Y \geq u]$ ,  $\sigma$  and  $\xi$  are the scale and shape parameters respectively. The joint distribution is defined by:

$$F(y_1, y_2) = \exp[-V(z_1, z_2)], \quad y_1 \geq u, \quad y_2 \geq u \quad (3)$$

where  $V$  is a homogeneous function of order -1 satisfying  $V(z_1, \infty) = z_1^{-1}$  and  $V(\infty, z_2) = z_2^{-1}$ , and  $z_i = -1/\log F(y_i)$ ,  $i = 1, 2$ .

Contrary to the univariate case, there is no finite parametrization for the  $V$  functions. Thus, it is common to use specific parametric families for  $V$  such as the logistic [Gumbel, 1960], the asymmetric logistic [Tawn, 1988], the negative logistic [Galambos, 1975] or the asymmetric negative logistic [Joe, 1990] models. Some details for these parametrisations are reported in Annex A. These models, as all models of the form (3) are asymptotically dependent. Roughly speaking, it means that the probability that two consecutive extreme (enough) events could occur is strictly positive. Other parametric families exist to consider simultaneously asymptotically dependent and independent cases [Bortot and Tawn, 1998]. However, apart from a few particular cases (see Section 3), the data analyzed here seem to belong to the asymptotically dependent class. Consequently, in this work, only asymptotically dependent models are considered - i.e. of the form (1)–(3).

## 2.2 Inference

The Markov chain model is fitted using maximum censored likelihood estimation [Ledford and Tawn, 1996]. The contribution  $l_n(y_1, y_2)$  of a point  $(y_1, y_2)$  to the numerator of equation (1) is given by:

$$l_n(y_1, y_2) = \begin{cases} \exp[-V(z_1, z_2)] [V_1(z_1, z_2)V_2(z_1, z_2) - V_{12}(z_1, z_2)] K_1 K_2, & \text{if } y_1 > u, y_2 > u \\ \exp[-V(z_1, z_2)] V_1(z_1, z_2) K_1, & \text{if } y_1 > u, y_2 \leq u \\ \exp[-V(z_1, z_2)] V_2(z_1, z_2) K_2, & \text{if } y_1 \leq u, y_2 > u \\ \exp[-V(z_1, z_2)], & \text{if } y_1 \leq u, y_2 \leq u \end{cases} \quad (4)$$

where  $z_j = -1/\log(F(y_j))$ ,  $K_j = -\lambda_j \sigma^{-1} t_j^{1+\xi} z_j^2 \exp(1/z_j)$ ,  $t_j = [1 + \xi(y_j - u)/\sigma]_+^{-1/\xi}$  and  $V_j, V_{12}$  are the partial derivative with respect to the component  $j$  and the mixed partial derivative respectively. The contribution  $l_d(y_j)$  of a point  $y_j$  to the denominator of equation (1) is given by:

$$l_d(y_j) = \begin{cases} \sigma^{-1} \lambda [1 + \xi(y_j - u)/\sigma]_+^{-1/\xi - 1}, & \text{if } y_j > u, \\ 1 - \lambda, & \text{otherwise.} \end{cases} \quad (5)$$

Finally, the log-likelihood is given by:

$$\log L(y_1, \dots, y_n) = \sum_{i=2}^n \log l_n(y_{i-1}, y_i) - \sum_{i=2}^{n-1} l_d(y_i) \quad (6)$$

From equation (6), it can be seen that the whole time series observations are used in the estimation procedure. Particularly, if the Markov chain model and the extreme value dependence structure are suited, the extreme event predictions may be more accurate.

## 2.3 Return levels

Most often, the major issue of an extreme value analysis is the quantile estimation. As for the POT approach, return level estimates can be computed. However, as all exceedances are inferred, this is done in a different way as the dependence between successive observations must be taken into account. For a stationary sequence  $Y_1, Y_2, \dots, Y_n$  with a marginal distribution function  $F$ , Lindgren and Rootzen [1987] have shown that:

$$\Pr[\max\{Y_1, Y_2, \dots, Y_n\} \leq y] \approx F(y)^{n^\theta} \quad (7)$$

where  $\theta$  is the extremal index. In particular, for daily independent data,  $n = 365$  and  $\theta = 1$ ; so that

$$\Pr[\max\{Y_1, Y_2, \dots, Y_{365}\} \leq y] = \prod_{i=1}^{365} F(y) = F(y)^{365} \quad (8)$$

As a consequence, the quantile  $Q_T$  corresponding to the  $T$ -year return period is defined by

$$Q_T = u - \sigma \xi^{-1} \left( 1 - \left\{ \lambda^{-1} \left[ 1 - (1 - 1/T)^{1/(n^\theta)} \right] \right\}^{-\xi} \right) \quad (9)$$

It is worth emphasizing equation (7) as it has a large impact on both theoretical and practical aspects. Indeed, for the AM approach, equation (7) is replaced by

$$\Pr[\max\{Y_1, Y_2, \dots, Y_n\} \leq y] \approx G(y) \quad (10)$$

where  $G$  is the distribution function of the random variable  $M_n = \max\{Y_1, Y_2, \dots, Y_n\}$ , that is a generalized extreme value distribution. In particular, the equations (7) and (10) differ as the first one is fitted to the whole observations  $Y_i$ , while the latter is fitted to the annual maxima ones. By definition, the number  $n_Y$  of the  $Y_i$  observations is much larger than the size  $n_M$  of the AM data set. Especially, for daily data,  $n_Y = 365n_M$ .

From equation (9), the extremal index  $\theta$  must be known to obtain quantile estimates. The methodology applied in this study is similar to the one suggested by Fawcett and Walshaw [2006]. Once the Markovian model is fitted, 100 Markov chains of length 2000 were generated. For each chain, the extremal index is estimated using the estimator proposed by Ferro and Segers [2003] to avoid an individual declustering procedure. In particular, the extremal index  $\theta$  is estimated using the following equations:

$$\hat{\theta}(u) = \begin{cases} \max \left( 1, \frac{2 \left[ \sum_{i=1}^{N-1} (T_i - 1)^2 \right]^2}{(N-1) \sum_{i=1}^{N-1} T_i^2} \right), & \text{if } \max\{T_i : 1 \leq i \leq N-1\} \leq 2 \\ \max \left( 1, \frac{2 \left( \sum_{i=1}^{N-1} T_i \right)^2}{(N-1) \sum_{i=1}^{N-1} (T_i - 1)(T_i - 2)} \right), & \text{otherwise} \end{cases} \quad (11)$$

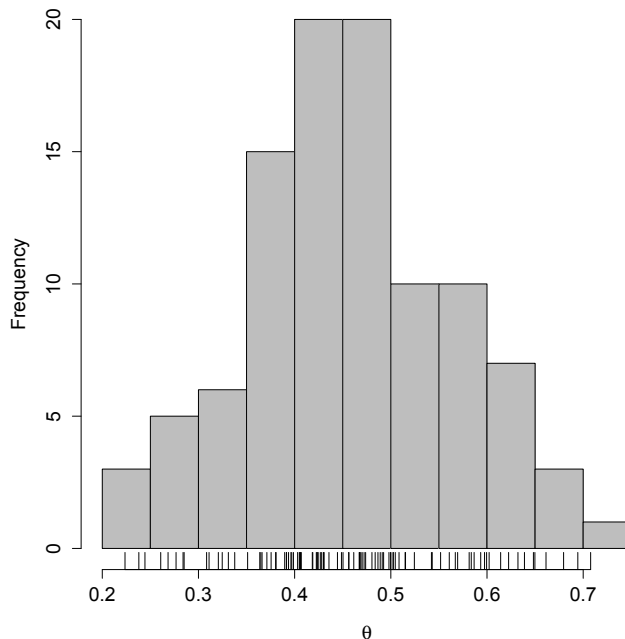


Figure 1: Histogram of the extremal index estimations from the 100 simulated Markov Chains of length 2000.

where  $N$  is the number of observations exceeding the threshold  $u$ ,  $T_i$  is the inter-exceedance time, e.g.  $T_i = S_{i+1} - S_i$  and the  $S_i$  is the  $i$ -th exceedance time.

Lastly, the extremal index related to a fitted Markov chain model is estimated using the sample mean of the 100 extremal index estimations. Figure 1 represents the histogram of these 100 extremal index estimations. In this study, as lots of time series are involved, the number and length of the simulated Markov chains may be too small to lead to the most accurate extremal index estimations; but avoid intractable CPU times. If less sites are considered, it is preferable to increase these two values.

A preliminary study (not shown here) demonstrates that this estimation procedure was more accurate than a direct estimation of  $\theta$  from the observed time series. Indeed, results showed that quantiles were poorly estimated with the latter approach. This confirms the conclusions drawn by Fawcett [2005] for the extreme wind speed data.

### 3 Extreme Value Dependence Structure Assessment

Prior to performing any estimations, it is necessary to test whether: (a) the first order Markov chain assumption and (b) the extreme value dependence structure (equation (3)) are suited to model successive observations above the threshold  $u$ . Though it is an important stage because of its consequences on quantile estimates [Ledford and Tawn, 1996; Bortot and Coles, 2000], verifying the (b) hypothesis is a considerable task. An overwhelming dependence between consecutive observations at finite levels is not sufficient as it does not give any information about the dependence relation at asymptotic levels.

Figures 2 and 3 plot the auto-correlation functions and the scatter plots between two consecutive observations for two different gaging stations. As the partial autocorrelation coefficients at larger lags become less relevant, Figure 2 and 3 (left panels) advocate the use of a first order Markov chain. However, despite an overwhelming dependence at lag 1 (Figure 2 and 3, right panels), it does certainly not justify the use of an asymptotic dependent model.

For this purpose, we have to check if two consecutive observations  $(Y_t, Y_{t+1})$  are asymptotically dependent; that is if

$$\chi = \lim_{\omega \rightarrow 1} \chi(\omega) = \lim_{\omega \rightarrow 1} \Pr [F(Y_{t+1}) > \omega | F(Y_t) > \omega] > 0 \quad (12)$$

$$\bar{\chi} = \lim_{\omega \rightarrow 1} \bar{\chi}(\omega) = \lim_{\omega \rightarrow 1} \frac{2 \log(1-\omega)}{\log \Pr[F(Y_t) > \omega, F(Y_{t+1}) > \omega]} - 1 = 1 \quad (13)$$

Figures 4 and 5 plot the evolution of the  $\chi(\omega)$  and  $\bar{\chi}(\omega)$  statistics as  $\omega$  increases for two different sites. For these figures, the confidence intervals are derived by bootstrapping contiguous blocks to take into account the successive observations dependence [Ledford and Tawn, 2003]. The  $\chi(\omega)$  and  $\bar{\chi}(\omega)$  statistics seem to depict two

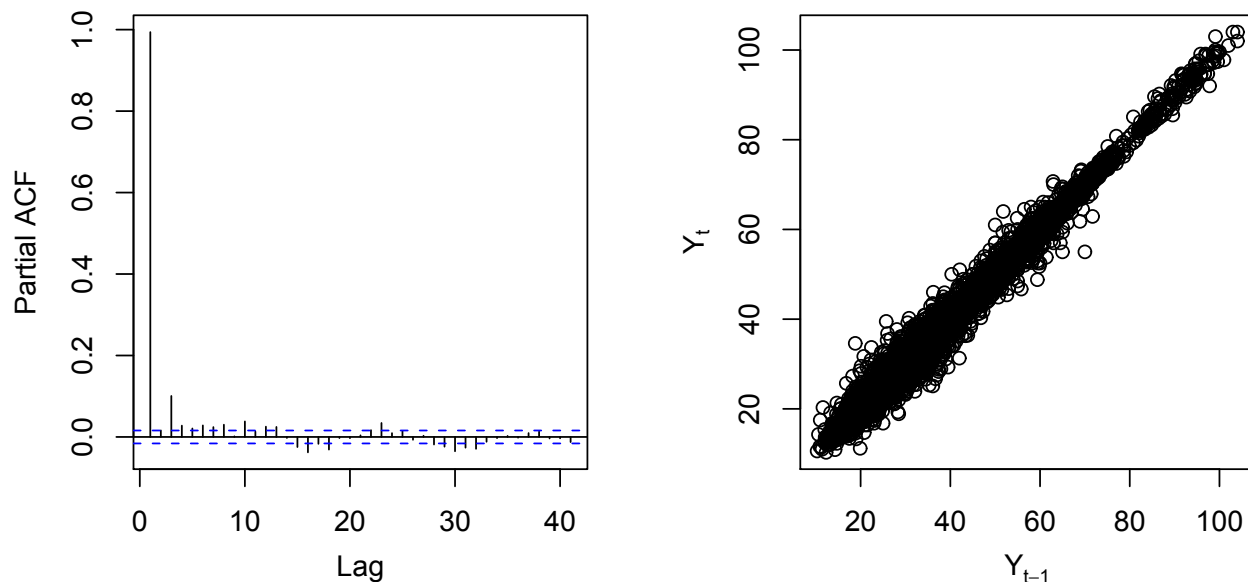


Figure 2: Autocorrelation plot (left panel) and scatterplot of the time series at lag 1 (right panel) for the Somme river at Abbeville (E6470910).

different asymptotic extremal dependence. From Figure 4, it seems that  $\lim \chi(\omega) \gg 0$  and  $\lim \bar{\chi}(\omega) = 1$  for  $\omega \rightarrow 1$ . On the contrary, Figure 5 advocates for  $\lim \chi(\omega) = 0$  and  $\lim \bar{\chi}(\omega) < 1$  for  $\omega \rightarrow 1$ . Consequently, Figure 4 seems to conclude for an asymptotic dependent case while Figure 5 for an asymptotic independent case.

Despite the fact that asymptotic (in)dependence should not be assessed using scatterplots, these two different features can however be deduced from Figures 2 and 3. For Figure 2, the scatterplot  $(Y_{t-1}, Y_t)$  is increasingly less spread as the observations becomes larger; while increasingly more spread for Figure 3. In other words, for the first case, the dependence seems to become stronger at larger levels while this is the contrary for the second case.

Two specific cases for different asymptotic dependence structures were illustrated. Table 1 shows the evolution of the  $\chi(\omega)$  statistics as  $\omega$  increases for all the sites under study. Most of the stations have significantly positive  $\chi(\omega)$  values. In addition, only 12 sites have a 95% confidence interval that contains the 0 value. For 9 of these stations, the 95% confidence intervals correspond to the theoretical lower and upper bounds; so that uncertainties are too large to determine the extremal dependence class. For the  $\bar{\chi}$  statistic, results are less clear-cut. Figure 6 represents the histograms for  $\bar{\chi}(\omega)$  for successive  $\omega$  values. Despite only a few observations being close to 1, most of the stations have a  $\bar{\chi}(\omega)$  value greater than 0.75. These values can be considered as significantly high as  $-1 < \bar{\chi}(\omega) \leq 1$ , for all  $\omega$ . Consequently, models of the form (1)–(3) may be suited to model the extremal dependence between successive observations.

Other methods exist to test the extremal dependence but were unconvincing for our application [Ledford and Tawn, 2003; Falk and Michel, 2006]. Indeed, the approach of Falk and Michel [2006] does not take into account the dependence between  $Y_{t-1}$  and  $Y_t$ ; while the test of Ledford and Tawn [2003] appears to be poorly discriminatory for our case study.

## 4 Performance of the Markovian Models on Quantile Estimation

### 4.1 Comparison between Markovian estimators

In this section, the performance of the six different extremal dependence structures is analyzed. These models are: *log* for the logistic, *nlog* for the negative logistic, *mix* for the mixed models and their relative asymmetric counterparts - e.g. *alog*, *anlog* and *amix*. For this purpose, quantiles with return periods 2, 10, 20, 50 and 100 years are aimed. In our study, the benchmark values are derived from a conventional POT analysis from the whole time series. As small samples are of interest, the performance of the Markovian models is analyzed on all sub time series of length 5, 10, 15 and 20 years. In addition, to assess the performance on all the gaging stations, the normalized bias (**nbias**), the standard deviation (**sd**) and the normalized mean squared error (**nmse**) are computed:

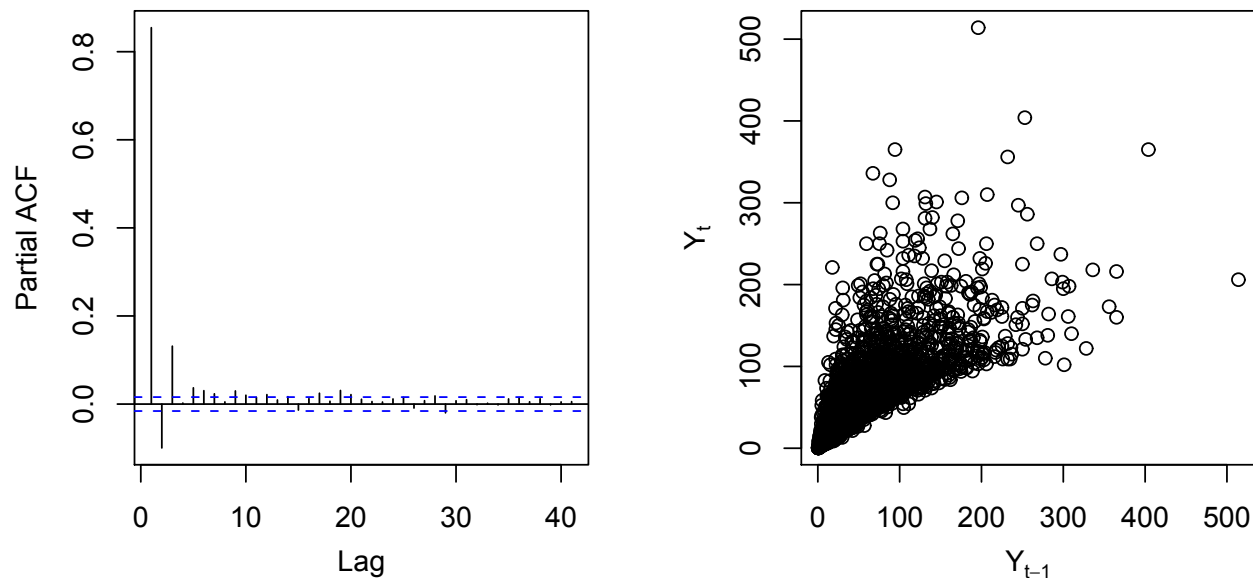


Figure 3: Autocorrelation plot (left panel) and scatterplot of the time series at lag 1 (right panel) for the Moselle river at Noirgueux (A4200630).

$$nbias = \frac{1}{N} \sum_{i=1}^N \frac{\hat{Q}_{i,T} - Q_T}{Q_T} \quad (14)$$

$$sd = \sqrt{\frac{1}{N-1} \sum_{i=1}^N \left( \frac{\hat{Q}_{i,T} - Q_T}{Q_T} - nbias \right)^2} \quad (15)$$

$$nmse = \frac{1}{N} \sum_{i=1}^N \left( \frac{\hat{Q}_{i,T} - Q_T}{Q_T} \right)^2 \quad (16)$$

where  $Q_T$  is the theoretical  $T$ -year return level and  $\hat{Q}_{i,T}$  is the  $i$ -th estimate of  $Q_T$ .

Figure 7 depicts the densities for the normalized biases for  $Q_{20}$  with a record length of 5 years. It is overwhelming that the extremal dependence structure has a great impact on the estimation of  $Q_{20}$ . Comparing the two panels, it can be noticed that the symmetric Markovian models give spreader densities; that is, more variable estimates. Independently of the symmetry, Figure 7 shows that the mixed dependence model is more accurate. Indeed, the densities for the *mix* and *amix* estimators are the narrowest and have a mode near zero.

Table 2 shows the *nbias*, *sd* and *nmse* statistics for all the Markovian estimators as the record length increases for quantile  $Q_{50}$ . This table confirms results derived from Figure 7. Indeed, the asymmetric dependence structures give less variable and biased estimates - as their *nbias* and *sd* statistics are smaller. In addition, whatever the record length is, the Markovian models perform with the same hierarchy; that is the *mix* and *amix* models are by far the most accurate estimators - i.e. with the smallest *nmse* values. The same results (not shown) have been found for other quantiles.

From an hydrological point of view, these two results are not surprising. The symmetric models suppose that the variables  $Y_t$  and  $Y_{t+1}$  are exchangeable. In our context, exchangeability means that the time series are reversible - e.g. the time vector direction has no importance. When dealing with AM or POT and stationary time series, it is a reasonable hypothesis. For example, the MLE remains the same with any permutations of the AM/POT sample. However, when modeling all exceedances, the time direction can not be considered as reversible as flood hydrographs are clearly non symmetric. As a consequence, the extremal dependence structure has to respect this feature and structures that do not allow asymmetry are more biased as they do not reflect natural properties.

Figure 8 plots normalized hydrographs simulated from the logistic and the asymmetric logistic models on an artificial but illustrative example. For the symmetric model (left panel), it can be seen that the hydrograph is (in mean) symmetric. Thus, such models are not suited to represent successive observations of flood events. On the contrary, the asymmetric logistic model seems to perform better in capturing the natural flood hydrograph asymmetry. It is important to notice that all Markovian models could not reproduce a particular "real" hydrograph but only reproduce it in mean. Consequently, if in mean the asymmetric model seems to be well suited, it does not imply that all simulated hydrographs are realistic.

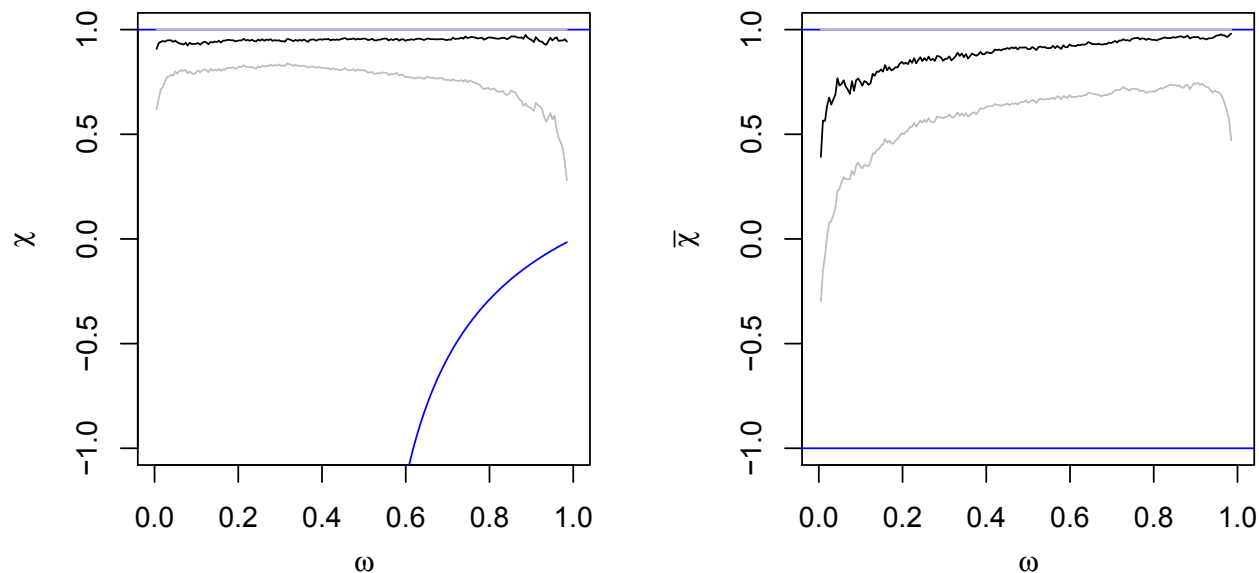


Figure 4: Plot of the  $\chi$  and  $\bar{\chi}$  statistics and the related 95% confidence intervals for the Somme river at Abbeville (E6470910). The solid blue lines are the theoretical bounds.

Figure 9 depicts simulated normalized hydrographs for all asymmetric Markovian models and station H1501010. It can be seen that the mean normalized hydrographs are quite similar for the three asymmetric Markovian models. Thus, the better performance for the *amix* model seems not to be related to the shape of the simulated hydrographs.

Figure 10 represents the Pickands' dependence function  $A(\omega)$  [Pickands, 1981] for all the gaging stations and the three asymmetric Markovian models. One major specificity of the mixed models is that these models can not account for perfect dependence cases. In particular, the Pickands' dependence functions for the mixed models satisfy  $A(0.5) \geq 0.75$  while  $A(0.5) \in [0.5, 1]$  for the logistic and negative logistic models. From Figure 10, it can be seen that only few stations have a dependence function that could not be modeled by the *amix* model. Therefore, the dependence range limitation of the *amix* model does not seem too restrictive.

In this section, the effect of the extremal dependence structure has been assessed. It has been established that the symmetric models are hydrologically inconsistent as they could not reproduce the flood event asymmetry. In addition, for all the quantiles analyzed, the asymmetric mixed model is the most accurate for flood peak estimations. Therefore, in the remainder of this section, only the *amix* model will be compared to conventional POT estimators.

## 4.2 Comparison between *amix* and conventional POT estimators

In this section, the performance of the *amix* estimator is compared to the estimators usually used in flood frequency analysis. For this purpose, the quantile estimates derived from the Maximum Likelihood Estimator (**MLE**), the Unbiased and Biased Probability Weighted moments estimators [Hosking and Wallis, 1987] (**PWU** and **PWB** respectively) are considered.

Figure 11 depicts the *nbias* densities for the *amix*, *MLE*, *PWU* and *PWB* estimators related to the  $Q_5$ ,  $Q_{10}$  and  $Q_{20}$  estimations with a record length of 5 years. It can be seen that *amix* is the most accurate model for all quantiles. Indeed, the *amix* *nbias* densities are the most sharp with a mode close to 0. Focusing only on "classical" estimators (e.g. *MLE*, *PWU* and *PWB*), there is no estimator that perform better than any other anytime. These two results advocate the use of the *amix* model.

Table 3 shows the performance of each estimator to estimate  $Q_{50}$  as the record length increases. It can be seen that the *amix* model performs better than the conventional estimators for the whole range of record lengths analyzed. On one hand, *amix* has the same bias than the conventional estimators. Thus, the *amix* dependence structure seems to be suited to estimate flood quantile estimates. On the other hand, because of its smaller variance, *amix* is more accurate than *MLE*, *PWU* and *PWB* estimators. This smaller variance is mainly a result of all of the exceedances (not only cluster maxima) being used in the inference procedure. Consequently, the *amix* model has a smaller *nmse* - around half of the conventional models ones.

Figure 12 shows the evolution of the *nmse* as the return period increases for the *amix*, *MLE*, *PWU* and *PWB* models. This figure corroborates the conclusions drawn from Figure 11 and Table 3. It can be seen that the *amix* model has the smallest *nmse*, independently of the return period and the record length. In addition, the *amix* becomes increasingly more efficient as the return period increases - mostly for return periods greater than



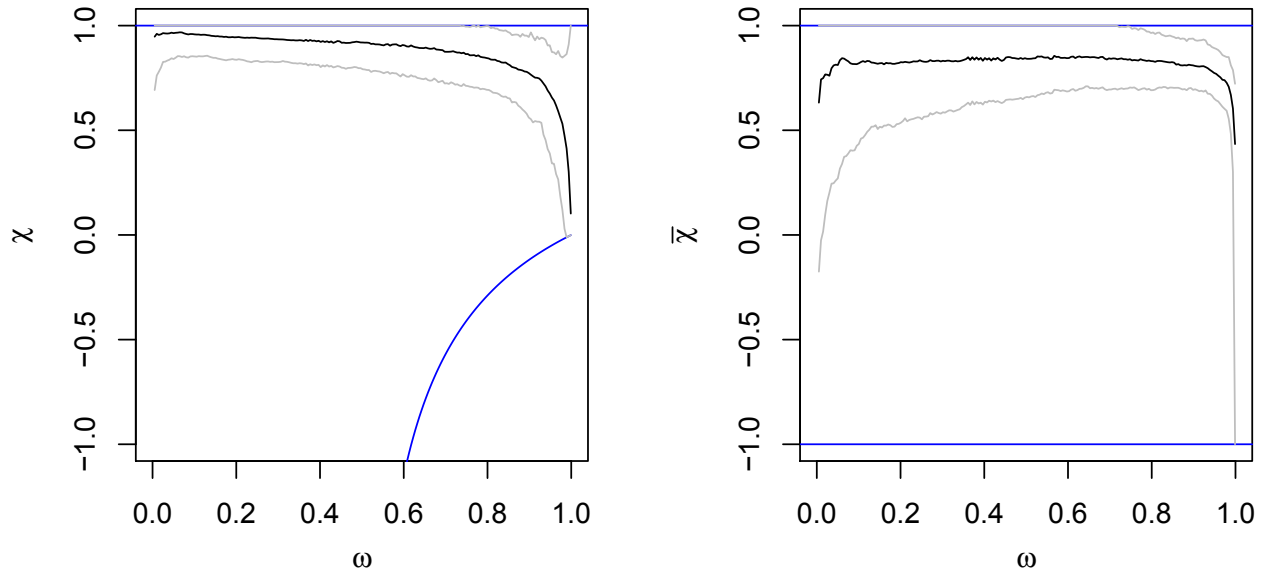


Figure 5: Plot of the  $\chi$  and  $\bar{\chi}$  statistics and the related 95% intervals for the Moselle river at Noiregoux (A4200630). The solid blue lines are the theoretical bounds.

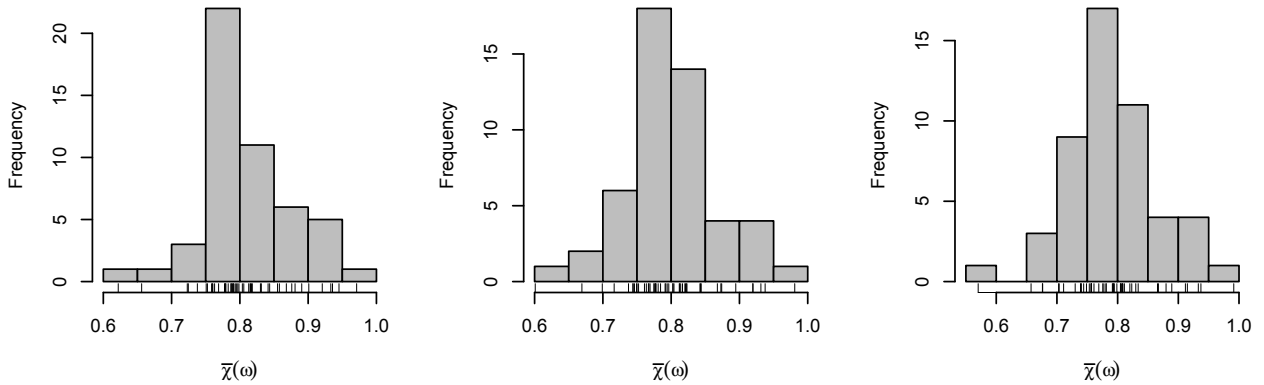


Figure 6: Histograms of the  $\bar{\chi}(\omega)$  statistics for different  $\omega$  values. Left panel:  $\omega = 0.98$ , middle panel:  $\omega = 0.985$  and right panel:  $\omega = 0.99$ .

20 years. While the conventional estimators present an erratic *nmse* behavior as the return period increases, the *amix* model is the only one that has a smooth evolution. To conclude, these results confirm that the *amix* model clearly improves flood peak quantile estimates - especially for large return periods.

## 5 Inference on Other Flood Characteristics

As all exceedances are modeled using a first order Markov chain, it is possible to infer other quantities than flood peaks - e.g. volume and duration. For example, in Section 4.1, stochastic hydrographs were derived from our Markovian models. In this section, the ability of these Markovian models to reproduce the flood duration is analyzed. For this purpose, the most severe flood hydrographs within each year are considered and normalized by their peak values. Consequently, from this observed normalized hydrograph set, two flood characteristics derived from a data set of hydrographs [Robson and Reed, 1999; Sauquet et al., 2008] are considered: (a) the duration  $d_{mean}$  above 0.5 of the normalized hydrograph set mean and (b) the median  $d_{med}$  of the durations above 0.5 of each normalized hydrograph.

### 5.1 Global Performance

Figure 13 plots the flood durations  $d_{mean}$  and  $d_{med}$  biases derived from the three asymmetric Markovian models in function of their empirical estimates. It can be seen that no model leads to accurate flood duration estimations.

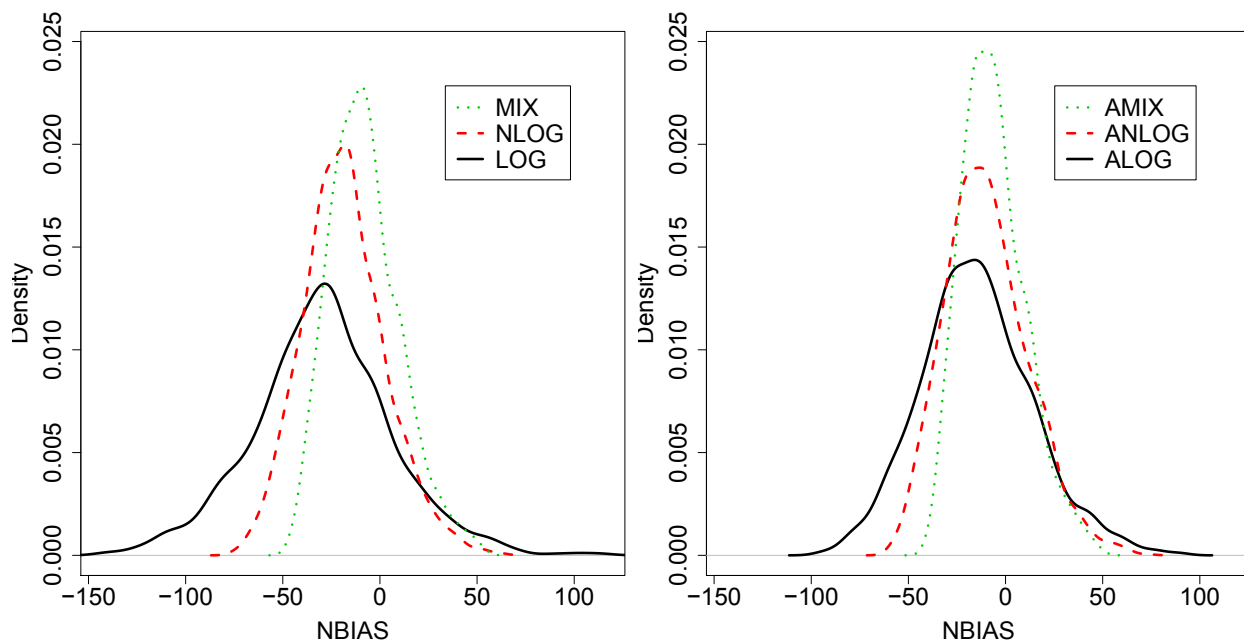


Figure 7: Densities of the normalized biases of  $Q_{20}$  estimates for the symmetric Markovian models (left panel) and the asymmetric ones (right panel). Target site record length: 5 years.

In addition, the extremal dependence structure has a clear impact on these estimations. In particular, the *anlog* and *amix* models seem to underestimate the flood durations, while the *alog* model leads to overestimations. Consequently, two different conclusions can be drawn. First, as large durations are poorly estimated, higher order Markov chains may be of interest. However, this is a considerable task as higher dimensional multivariate extreme value distributions often lead to numerical problems. Instead of considering higher order, another alternative may be to change daily observations for  $d$ -day observations - where  $d$  is larger than 1. Second, it is overwhelming that the extremal dependence structure affects the flood duration estimations. As noticed in Section 2.1, there is no finite parametrization for the extremal dependence structure  $V$  - see Equation (3). Consequently, it seems reasonable to suppose that one suited for flood hydrograph estimation may exist.

Figure 14 depicts the observed normalized mean hydrographs and the ones predicted by the three asymmetric Markovian models. For the J0621610 station (left panel), the normalized hydrograph is well estimated by the three models; whereas for the L0400610 station (right panel), the normalized hydrograph is poorly predicted. This result confirms the inability of the three Markovian models to reproduce long flood events with daily data and a first order Markov chain.

Figure 15 represents the biases related to each value of the normalized mean hydrograph. In addition, to help estimator comparison, the *nmse* is reported at the right side. It can be seen that the *alog* model dramatically overestimates the hydrograph rising limb while giving reasonable estimations for the falling phase. The *anlog* model slightly overestimates the rising part while strongly underestimates the falling one. The *amix* model always leads to underestimations - this is more pronounced for the falling limb. However, despite these different behaviors, these three estimators seems to have a similar performance - in terms of *nmse*.

Figure 16 represents the spatial distribution of the *nmse* on the normalized mean hydrograph estimation for each Markovian model. It seems that there is a specific spatial distribution. In particular, the worst cases are related to the middle part of France. In addition, for different extremal dependence structures, the best *nmse* values correspond to different spatial locations. The *alog* model is more accurate for the extreme north part of France; the *anlog* model is more efficient for the east of France; while the *amix* model performs best in the middle of France. Consequently, as at a global scale no model is accurate to estimate the normalized mean hydrograph, it is worth trying to identify which catchment types are related to the best estimations.

For our data set, this is a considerable task. No standard statistical technique lead to reasonable results. In particular, the principal component analysis, hierarchical classification, sliced inverse regression lead to no conclusion about which catchment types are more suited for our models. Only a regression approach gives some first guidelines. For this purpose, a regression between the *nbias* on the  $d_{mean}$  estimation for each asymmetric model and some geomorphologic and hydrologic indices are performed. The effect of the drainage area, an index of catchment slope derived from the hypsometric curve [Roche, 1963], the Base Flow Index (**BFI**) [Tallaksen and Van Lanen, 2004, Section 5.3.3] are considered. In addition, one raingage representative of each basin was selected. To

## 7. Modélisation de Tous les Excès Selon une Structure de Dépendance Extrême

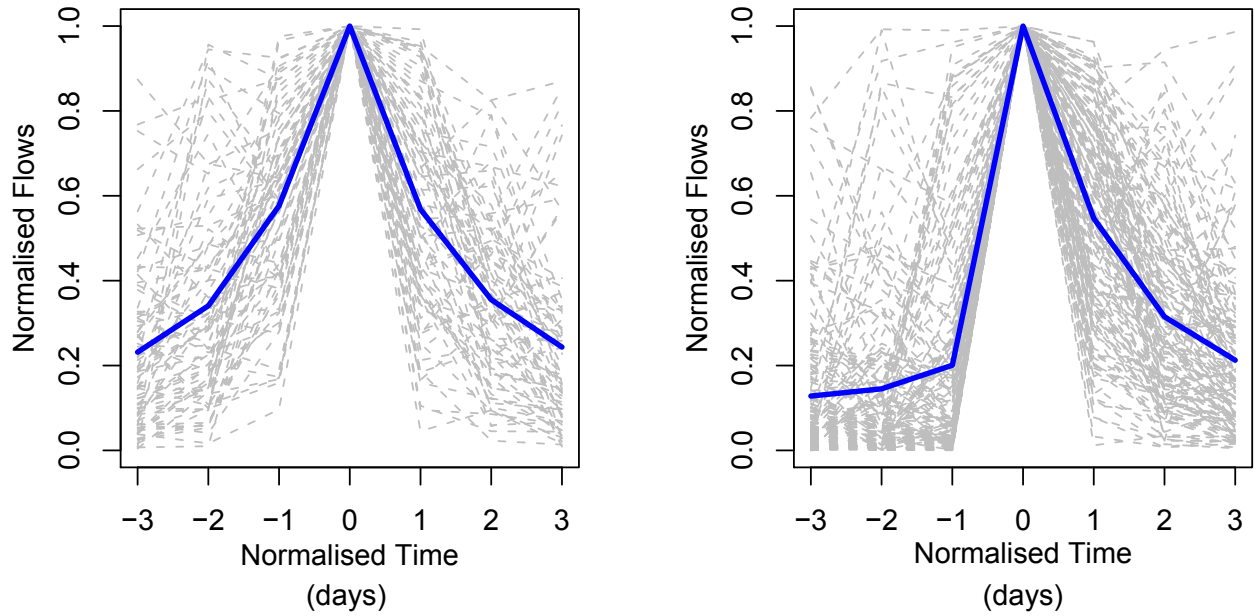


Figure 8: Synthetic hydrographs derived from the *log* (left panel) and the *alog* (right panel) models. Solid bold blue line corresponds to the mean hydrograph.

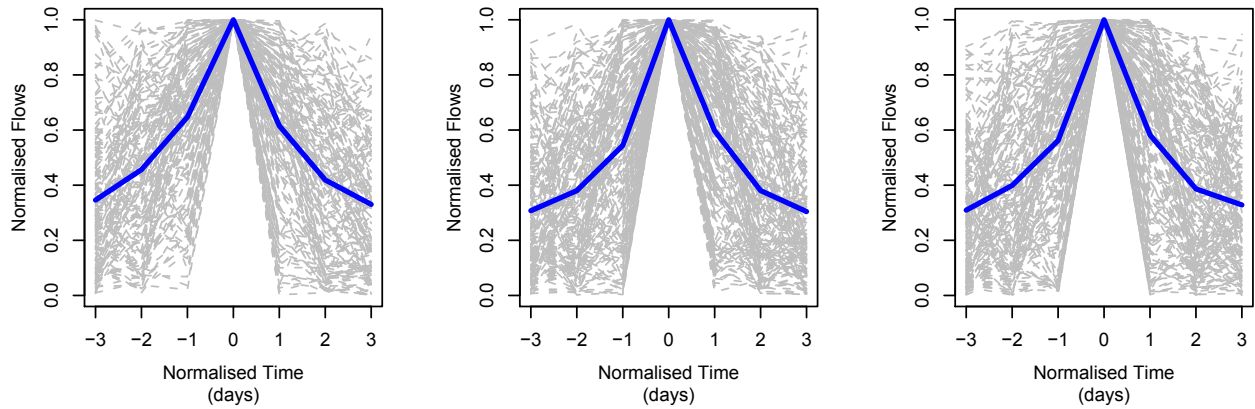


Figure 9: Synthetic hydrographs derived from the *alog* (left panel), the *anlog* (middle panel) and the *amix* (right panel) models. Solid bold blue line corresponds to the mean hydrograph. Station H1501010.

characterize rainfall events persistence [Koutsoyiannis et al., 1998], we derived the exponent  $\beta(T)$  of the power law

$$I(d, T) = \alpha(T)d^{\beta(T)} \quad (17)$$

where  $I(d, T)$  is the  $T$ -year  $d$ -day rainfall. For our study,  $T$  is fixed to 10 years. However, for our data only the  $BFI$  variable seems to be explicative enough. For the *alog* model, no parametrization (including non-parametric methods such as generalized additive models) was able to predict the  $nbias$  accurately.

$$nbias(d_{mean}; anlog) = 0.89 - 2.19BFI, \quad R^2 = 0.40 \quad (18)$$

$$nbias(d_{mean}; amix) = 0.49 - 1.74BFI, \quad R^2 = 0.43 \quad (19)$$

From equations (18) and (19), the  $BFI$  variable explains around 40% of the variance. Despite the fact that a large variance proportion is not taken into account, the  $BFI$  is clearly related to the  $d_{mean}$  estimation performance. These equations indicate that the *anlog* (resp. *amix*) model is more accurate to reproduce the  $d_{mean}$  variable for gaging stations with a  $BFI$  around 0.4 (resp. 0.28). These  $BFI$  values correspond to catchments with moderate up to flash flow regimes respectively. These results corroborate the ones derived from Figure 15: the first order Markovian models with a 1-day lag conditioning are not suited for long flood duration estimations. Consequently, while no physiographic characteristic is related to the *alog* performance; it is suggested, for such

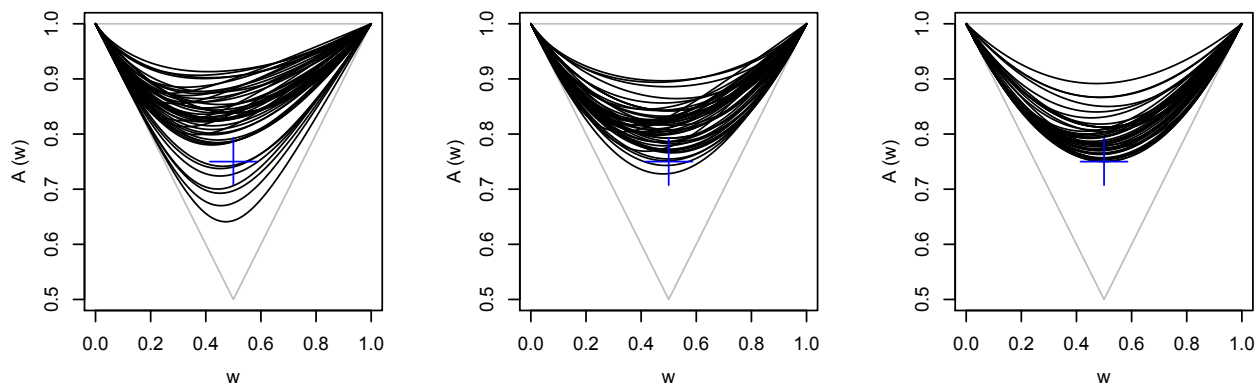


Figure 10: Representation of the Pickands' dependence functions for the 50 gaging stations. Left panel : *alog*, middle panel: *anlog* and right panel: *amix*. “+” represents the theoretical dependence bound for the *amix* model.

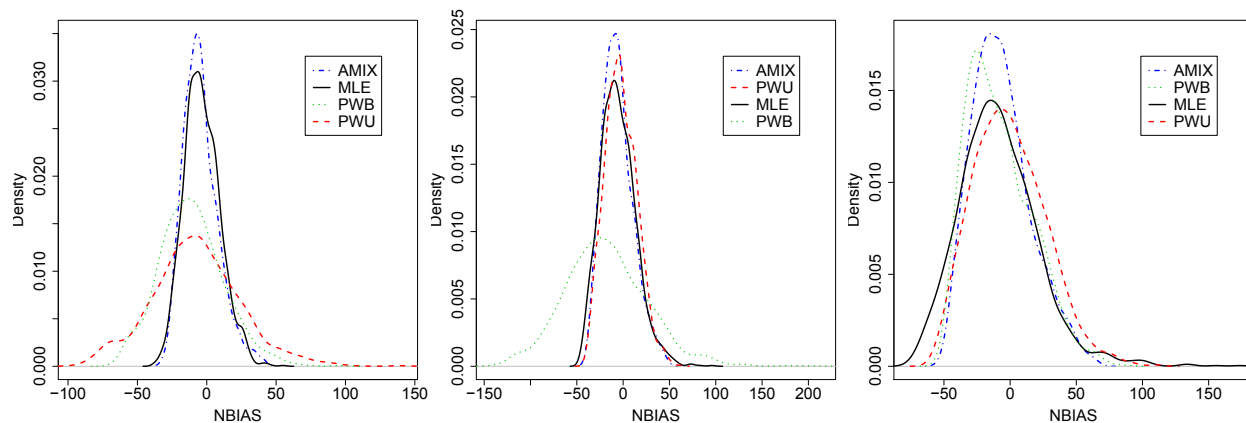


Figure 11: Densities of the normalized biases for the *amix* model and the *MLE*, *PWU*, and *PWB* estimators for quantiles  $Q_5$  (left panel),  $Q_{10}$  (middle panel) and  $Q_{20}$  (right panel). Record length: 5 years.

1-day lag conditioning, to use the *anlog* and *amix* models for quick basins. However, the current results show that the flood duration estimation must be considered with care and lots of improvements still remain to be done. In particular, conditioning the flow  $Y_t$  by the previous value  $Y_{t-1}$  may not be relevant for all of our gaging stations. For this purpose, several guidelines are presented within the conclusion to improve the flood duration estimations.

## 6 Conclusion

Despite that univariate EVT is widely applied in environmental sciences, its multivariate extension is rarely considered. This work tries to promote the use of the multivariate EVT in hydrology. In this work, the bivariate case was considered as the dependence between two successive observations was modeled by a first order Markov chain. This innovative approach has two main advantages for practitioners as: (a) the number of data to be inferred increases considerably and (b) other features can be estimated - flood duration, volume.

In this study, a comparison between six different extremal dependence structures (including both symmetric and asymmetric forms) has been performed. Results show that an asymmetric dependence structure is more relevant. From a hydrological point of view, this asymmetry is rational as flood hydrographs are asymmetric. In particular, for our data, the asymmetric mixed model gives the most accurate flood peak estimations and clearly improves flood peak estimations compared to conventional estimators independently of the return period considered.

The ability of these Markovian models to estimate the flood duration was carried out. It has been shown that, at first sight, no dependence structure is able to reproduce the flood hydrograph accurately. However, it seems that the *anlog* and *amix* models may be better suited for dealing with moderate up to flash flow regimes. These results depend strongly on the conditioning term (i.e.  $\Pr[Y_t \leq y_t | Y_{t-\delta} = y_{t-\delta}]$ ) of the first order Markov chain and on the auto-correlation within the time series. In our application,  $\delta = 1$  and daily time step was considered.

More general conclusions can be drawn. The weakness of the proposed models to derive consistent flood hydrographs may not be related to the daily time step but to the inadequacy between the conditioning term and the flood dynamics. To ensure better results, higher order Markov chains may be of interest. However, as numerical

## 7. Modélisation de Tous les Excès Selon une Structure de Dépendance Extrême

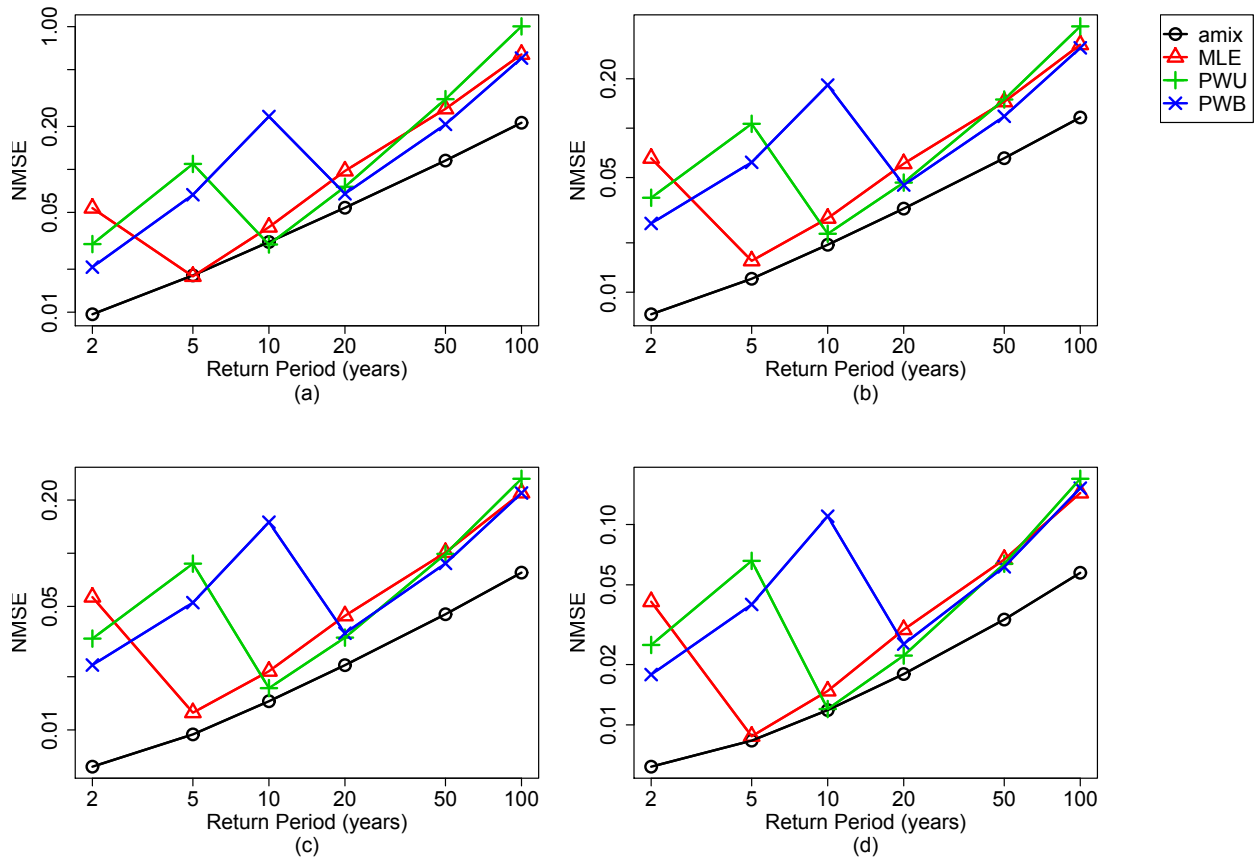


Figure 12: Evolution of the  $nmse$  as the return period increases for the  $amix$ ,  $MLE$ ,  $PWU$  and  $PWB$  estimators. Record length: (a) 5 years, (b) 10 years, (c) 15 years and (d) 20 years.

problems may arise, another alternative may be to still consider a first order chain but to change the “conditioning lag value”  $\delta$ . In particular, for some basins, it may be more relevant to condition the Markov chain with a larger but well suited lag value.

Another option to improve the proposed models for flood hydrograph estimation is to use a more suited dependence function  $V$ . As there is no finite parametrization for the extremal dependence structure, it seems reasonable that one suited for flood hydrographs may exist. In this work, the results show that the  $anlog$  model is more able to reproduce the hydrograph rising part, while the  $alog$  is better the falling phase. Define

$$V(z_1, z_2) = \alpha V_1(z_1, z_2) + \beta V_2(z_1, z_2)$$

where  $V_1$  (resp.  $V_2$ ) is the extremal dependence function for the  $alog$  (resp.  $anlog$ ) model and  $\alpha$  and  $\beta$  are real constants such as  $\alpha + \beta = 1$ . By definition,  $V$  is a new extremal dependence function. In particular,  $V$  may combine the accuracy of the  $alog$  and  $anlog$  models for both the rising and falling part of the flood hydrograph. Another alternative may be to look at non-parametric Pickands’ dependence function estimators [Capéraà et al., 1997] but that will require techniques to simulate Markov chains from these non-parametric estimations.

All statistical analysis were performed within the R Development Core Team [2007] framework. In particular, the POT package [Ribatet, 2007] integrates the tools that were developed to carry out the modeling effort presented in this paper. This package is available, free of charge, at the website <http://www.R-project.org>, section CRAN, Packages.

## Acknowledgments

The authors wish to thank the French HYDRO database for providing the data. Benjamin Renard is acknowledged for criticizing thoroughly the data analyzed in this study.

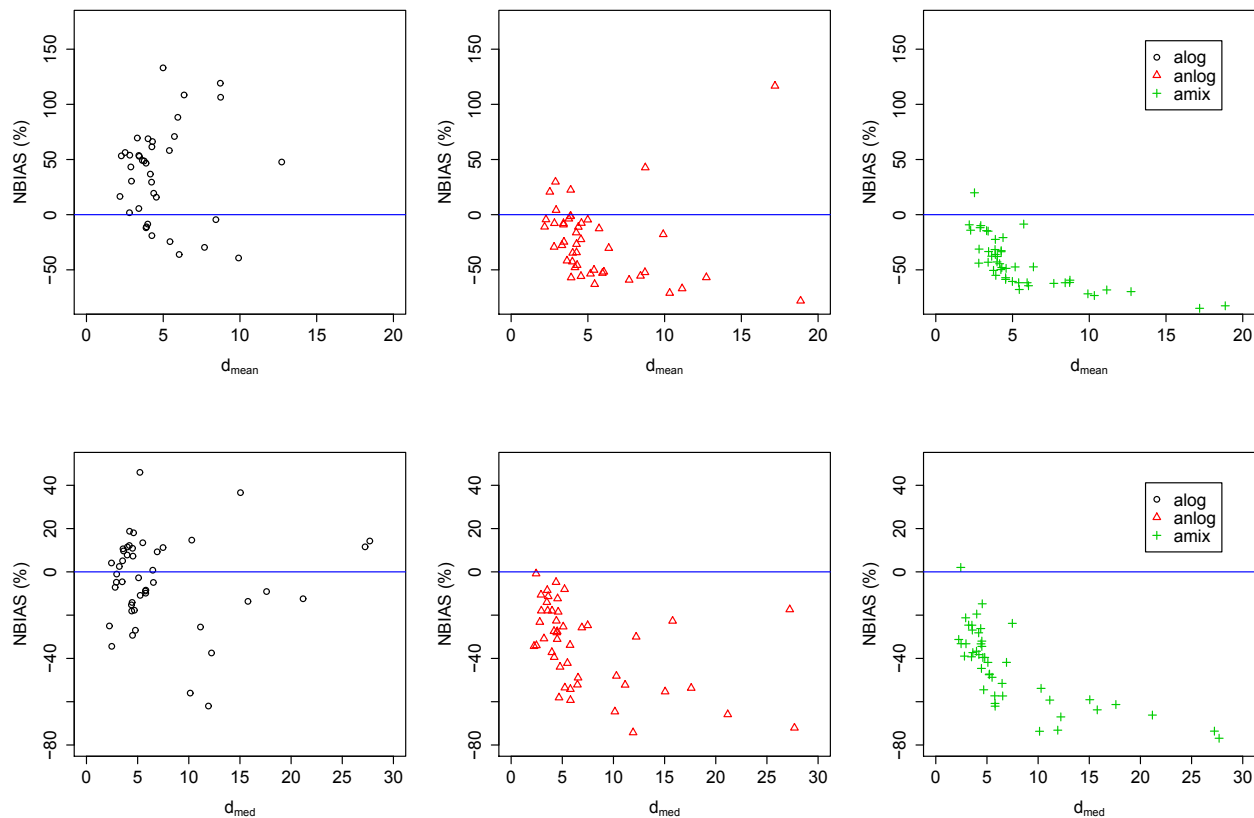


Figure 13:  $d_{mean}$  and  $d_{med}$  normalized biases in function of the theoretical values for the three asymmetric Markovian models.

## A Parametrization for the Extremal Dependence

This annex presents some useful results for the six extremal dependence models that have been considered in this work. As first order Markov chains were used, only the bivariate results are described.

## 7. Modélisation de Tous les Excès Selon une Structure de Dépendance Extrême

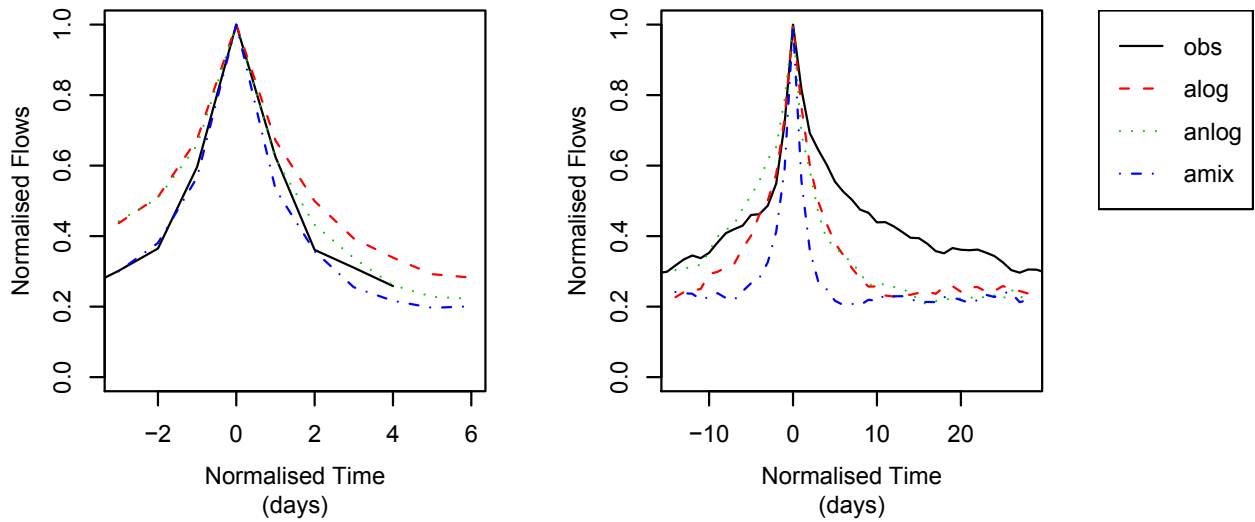


Figure 14: Observed and simulated normalized mean hydrographs for the J0621610 (left panel) and the L0400610 (right panel) stations.

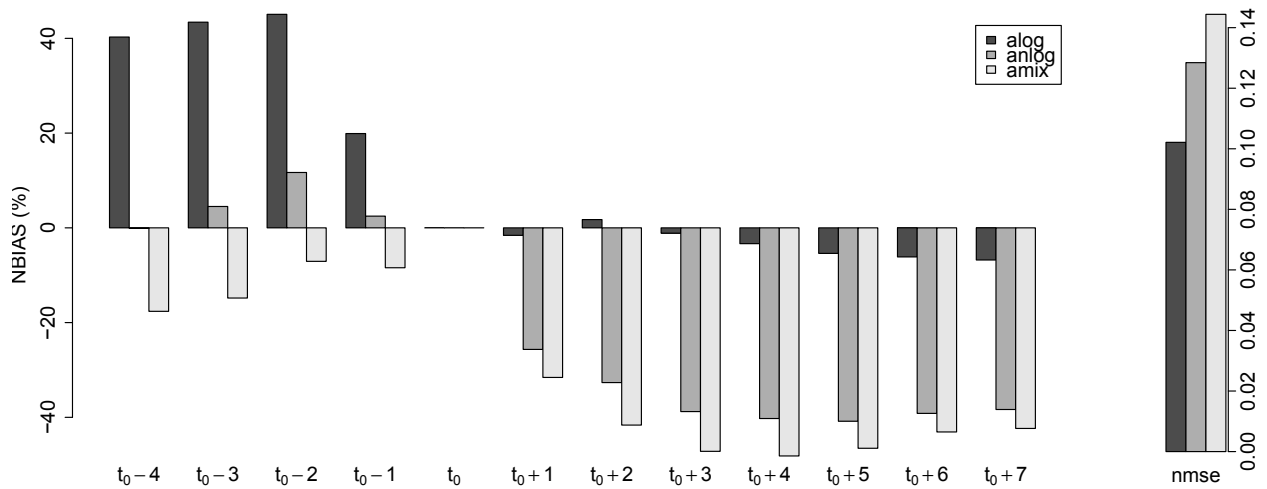


Figure 15: Evolution of the biases for the normalized mean hydrograph estimations in function of the distance of the flood peak time.

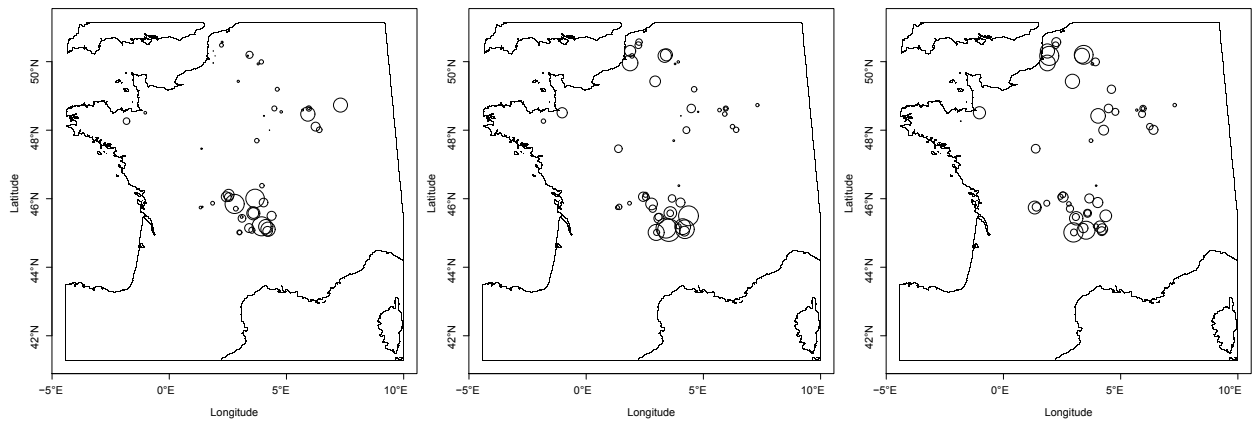


Figure 16: *nmse* spatial distribution according for the three Markovian models. Left panel: *alog*, middle panel: *anlog* and right panel: *amix*. The radius is proportional to the *nmse* value.

Table 1:  $\chi(\omega)$  statistics for all stations.  $\omega = 0.98, 0.985, 0.99$ .

Stations	$\omega = 0.98$		$\omega = 0.985$		$\omega = 0.99$	
	$\chi(\omega)$	95% C.I.	$\chi(\omega)$	95% C.I.	$\chi(\omega)$	95% C.I.
A3472010	0.67	(-0.02, 1.00)	0.60	(-0.02, 1.00)	0.57	(-0.01, 1.00)
A4200630	0.53	( 0.21, 0.81)	0.45	( 0.07, 0.77)	0.38	(-0.01, 0.76)
A4250640	0.55	( 0.27, 0.82)	0.49	( 0.18, 0.76)	0.41	( 0.02, 0.71)
A5431010	0.44	(-0.02, 1.00)	0.44	(-0.02, 1.00)	0.41	(-0.01, 1.00)
A5730610	0.59	( 0.25, 0.94)	0.56	( 0.20, 0.90)	0.50	( 0.07, 0.97)
A6941010	0.62	( 0.22, 0.99)	0.60	( 0.16, 1.00)	0.56	( 0.06, 1.00)
A6941015	0.63	( 0.29, 0.95)	0.60	( 0.20, 0.96)	0.58	( 0.17, 0.98)
D0137010	0.39	( 0.04, 0.69)	0.33	(-0.02, 0.67)	0.28	(-0.01, 0.69)
D0156510	0.59	( 0.25, 0.88)	0.55	( 0.20, 0.86)	0.53	( 0.14, 0.92)
E1727510	0.62	( 0.18, 0.91)	0.59	( 0.16, 0.93)	0.47	(-0.01, 0.89)
E1766010	0.63	( 0.23, 0.98)	0.59	( 0.17, 0.96)	0.54	( 0.09, 0.96)
E3511220	0.59	( 0.10, 1.00)	0.53	(-0.02, 1.00)	0.50	(-0.01, 0.99)
E4035710	0.77	( 0.02, 1.00)	0.68	(-0.02, 1.00)	0.60	(-0.01, 1.00)
E5400310	0.88	( 0.30, 1.00)	0.89	( 0.29, 1.00)	0.83	( 0.13, 1.00)
E5505720	0.91	( 0.24, 1.00)	0.87	( 0.09, 1.00)	0.86	( 0.02, 1.00)
E6470910	0.96	( 0.40, 1.00)	0.94	( 0.25, 1.00)	0.98	( 0.00, 1.00)
H0400010	0.84	( 0.12, 1.00)	0.83	( 0.02, 1.00)	0.78	(-0.01, 1.00)
H1501010	0.82	( 0.36, 1.00)	0.90	( 0.39, 1.00)	0.84	( 0.26, 1.00)
H2342010	0.68	( 0.31, 1.00)	0.67	( 0.25, 1.00)	0.60	( 0.11, 1.00)
H5071010	0.75	( 0.30, 1.00)	0.76	( 0.22, 1.00)	0.75	( 0.15, 1.00)
H5172010	0.80	( 0.47, 1.00)	0.77	( 0.42, 1.00)	0.73	( 0.30, 1.00)
H6201010	0.69	( 0.29, 1.00)	0.69	( 0.14, 1.00)	0.69	( 0.08, 1.00)
H7401010	0.85	( 0.46, 1.00)	0.85	( 0.38, 1.00)	0.81	( 0.27, 1.00)
I9221010	0.67	( 0.23, 1.00)	0.66	( 0.19, 1.00)	0.59	( 0.04, 1.00)
J0621610	0.61	( 0.25, 0.92)	0.58	( 0.20, 0.94)	0.51	( 0.08, 0.91)
K0433010	0.59	( 0.22, 0.91)	0.54	( 0.15, 0.89)	0.45	( 0.00, 0.85)
K0454010	0.71	( 0.37, 1.00)	0.67	( 0.24, 1.00)	0.65	( 0.14, 1.00)
K0523010	0.62	(-0.02, 1.00)	0.58	(-0.02, 1.00)	0.53	(-0.01, 1.00)
K0550010	0.61	( 0.22, 0.94)	0.57	( 0.15, 0.94)	0.54	( 0.07, 1.00)
K0673310	0.67	( 0.24, 1.00)	0.65	( 0.18, 1.00)	0.66	( 0.07, 1.00)
K0910010	0.65	(-0.02, 1.00)	0.61	(-0.02, 1.00)	0.58	(-0.01, 1.00)
K1391810	0.68	( 0.27, 1.00)	0.64	( 0.16, 0.98)	0.60	( 0.06, 0.96)
K1503010	0.69	( 0.38, 0.98)	0.67	( 0.30, 0.98)	0.64	( 0.23, 1.00)
K2330810	0.68	( 0.29, 1.00)	0.66	( 0.22, 1.00)	0.62	( 0.09, 1.00)
K2363010	0.65	( 0.26, 0.98)	0.66	( 0.16, 1.00)	0.61	( 0.01, 1.00)
K2514010	0.61	( 0.24, 1.00)	0.61	( 0.21, 1.00)	0.58	( 0.12, 1.00)
K2523010	0.53	(-0.02, 1.00)	0.53	(-0.02, 1.00)	0.51	(-0.01, 1.00)
K2654010	0.68	( 0.37, 1.00)	0.68	( 0.31, 1.00)	0.60	( 0.10, 1.00)
K2674010	0.60	( 0.25, 0.89)	0.58	( 0.22, 0.94)	0.54	( 0.08, 0.95)
K2871910	0.62	( 0.26, 0.95)	0.57	( 0.15, 0.94)	0.56	( 0.10, 0.97)
K2884010	0.62	( 0.25, 1.00)	0.57	( 0.17, 0.97)	0.59	( 0.16, 1.00)
K3222010	0.56	( 0.21, 0.90)	0.53	( 0.18, 0.93)	0.46	( 0.11, 0.89)
K3292020	0.59	( 0.27, 0.91)	0.57	( 0.17, 0.91)	0.48	( 0.07, 0.90)
K4470010	0.76	( 0.39, 1.00)	0.77	( 0.40, 1.00)	0.73	( 0.27, 1.00)
K5090910	0.64	( 0.27, 0.93)	0.64	( 0.26, 0.96)	0.58	( 0.12, 0.98)
K5183010	0.57	( 0.14, 0.91)	0.56	( 0.15, 0.96)	0.53	( 0.06, 0.97)
K5200910	0.63	( 0.24, 0.93)	0.62	( 0.20, 0.95)	0.56	( 0.11, 0.97)
L0140610	0.73	( 0.23, 1.00)	0.66	( 0.15, 1.00)	0.58	(-0.01, 1.00)
L0231510	0.59	( 0.16, 0.91)	0.55	( 0.11, 0.92)	0.53	(-0.01, 0.92)
L0400610	0.74	(-0.02, 1.00)	0.65	(-0.02, 1.00)	0.61	(-0.01, 1.00)



## 7. Modélisation de Tous les Excès Selon une Structure de Dépendance Extrême

Table 2: Several characteristics of the Markovian estimators on  $Q_{50}$  estimation as the record length increases.

Model	5 years			10 years			15 years			20 years		
	<i>nbias</i>	<i>sd</i>	<i>nmse</i>	<i>nbias</i>	<i>sd</i>	<i>nmse</i>	<i>nbias</i>	<i>sd</i>	<i>nmse</i>	<i>nbias</i>	<i>sd</i>	<i>nmse</i>
<i>log</i>	-0.35	0.74	0.66	-0.32	0.56	0.42	-0.30	0.48	0.32	-0.28	0.42	0.25
<i>nlog</i>	-0.21	0.44	0.24	-0.20	0.34	0.15	-0.18	0.29	0.12	-0.18	0.25	0.09
<i>mix</i>	-0.08	0.37	0.14	-0.07	0.28	0.08	-0.06	0.23	0.06	-0.05	0.20	0.04
<i>alog</i>	-0.15	0.62	0.41	-0.13	0.47	0.24	-0.11	0.40	0.17	-0.10	0.34	0.13
<i>anlog</i>	-0.10	0.44	0.21	-0.09	0.33	0.12	-0.08	0.28	0.09	-0.08	0.24	0.06
<i>amix</i>	-0.06	0.33	0.12	-0.05	0.25	0.06	-0.04	0.21	0.05	-0.03	0.18	0.03

Table 3: Several characteristics of the *amix*, *MLE*, *PWU* and *PWB* estimators for  $Q_{50}$  estimation as the record length increases.

Model	5 years			10 years			15 years			20 years		
	<i>nbias</i>	<i>sd</i>	<i>nmse</i>	<i>nbias</i>	<i>sd</i>	<i>nmse</i>	<i>nbias</i>	<i>sd</i>	<i>nmse</i>	<i>nbias</i>	<i>sd</i>	<i>nmse</i>
<i>amix</i>	-0.06	0.33	0.12	-0.05	0.25	0.07	-0.04	0.21	0.05	-0.04	0.18	0.03
<i>MLE</i>	-0.13	0.50	0.27	-0.14	0.36	0.14	-0.13	0.29	0.10	-0.11	0.23	0.07
<i>PWU</i>	0.08	0.55	0.31	-0.01	0.39	0.15	-0.03	0.31	0.10	-0.03	0.25	0.06
<i>PWB</i>	-0.07	0.45	0.21	-0.10	0.33	0.12	-0.11	0.27	0.09	-0.10	0.23	0.06

Table 4: Partial and mixed partial derivatives, definition domain, total independent and perfect dependent cases for each extremal symmetric dependence function  $V$ .

Model	Symmetric Models		
	<i>log</i>	<i>nlog</i>	<i>mix</i>
$V(x, y)$	$(x^{-1/\alpha} + y^{-1/\alpha})^\alpha$	$\frac{1}{x} + \frac{1}{y} - (x^\alpha + y^\alpha)^{-1/\alpha}$	$\frac{1}{x} + \frac{1}{y} - \frac{\alpha}{x+y}$
$V_1(x, y)$	$-x^{-\frac{1}{\alpha}-1} V(x, y)^{\frac{\alpha-1}{\alpha}}$	$-\frac{1}{x^2} + x^{\alpha-1} (x^\alpha + y^\alpha)^{-\frac{1}{\alpha}-1}$	$-\frac{1}{x^2} + \frac{\alpha}{(x+y)^2}$
$V_2(x, y)$	$-y^{-\frac{1}{\alpha}-1} V(x, y)^{\frac{\alpha-1}{\alpha}}$	$-\frac{1}{y^2} + y^{\alpha-1} (x^\alpha + y^\alpha)^{-\frac{1}{\alpha}-1}$	$-\frac{1}{y^2} + \frac{\alpha}{(x+y)^2}$
$V_{12}(x, y)$	$-(xy)^{-\frac{1}{\alpha}-1} \frac{1-\alpha}{\alpha} V(x, y)^{\frac{\alpha-2}{\alpha}}$	$-(\alpha+1)(xy)^{\alpha-1} (x^\alpha + y^\alpha)^{-\frac{1}{\alpha}-2}$	$-\frac{2\alpha}{(x+y)^3}$
$A(w)$	$\left[(1-w)^{\frac{1}{\alpha}} + w^{\frac{1}{\alpha}}\right]^\alpha$	$1 - [(1-w)^{-\alpha} + w^{-\alpha}]^{-\frac{1}{\alpha}}$	$1 - w(1-w)^\alpha$
Independence	$\alpha = 1$	$\alpha \rightarrow 0$	$\alpha = 0$
Total dependence	$\alpha \rightarrow 0$	$\alpha \rightarrow +\infty$	Never reached
Constraint	$0 < \alpha \leq 1$	$\alpha > 0$	$0 \leq \alpha \leq 1$

Table 5: Partial and mixed partial derivatives, definition domain, total independent and perfect dependent cases for each extremal asymmetric dependence function  $V$ .

Model	Asymmetric Models		
	<i>alog</i>	<i>anlog</i>	<i>amix</i>
$V(x, y)$	$\frac{1-\theta_1}{x} + \frac{1-\theta_2}{y} + \left[ \left( \frac{x}{\theta_1} \right)^{-1/\alpha} + \left( \frac{y}{\theta_2} \right)^{-1/\alpha} \right]^\alpha$	$\frac{1}{x} + \frac{1}{y} - \left[ \left( \frac{x}{\theta_1} \right)^\alpha + \left( \frac{y}{\theta_2} \right)^\alpha \right]^{-1/\alpha}$	$\frac{1}{x} + \frac{1}{y} - \frac{(2\alpha+\theta)x+(\alpha+\theta)y}{(x+y)^2}$
$V_1(x, y)$	$-\frac{1-\theta_1}{x^2} - \theta_1^{\frac{1}{\alpha}} x^{-\frac{1}{\alpha}-1} \left[ \left( \frac{x}{\theta_1} \right)^{-1/\alpha} + \left( \frac{y}{\theta_2} \right)^{-1/\alpha} \right]^{\alpha-1}$	$-\frac{1}{x^2} + \theta_1^{-\alpha} x^{\alpha-1} \left[ \left( \frac{x}{\theta_1} \right)^\alpha + \left( \frac{y}{\theta_2} \right)^\alpha \right]^{-1/\alpha-1}$	$-\frac{1}{x^2} - \frac{2\alpha+\theta}{(x+y)^2} + 2\frac{(2\alpha+\theta)x+(\alpha+\theta)y}{(x+y)^3}$
$V_2(x, y)$	$-\frac{1-\theta_2}{y^2} - \theta_2^{\frac{1}{\alpha}} y^{-\frac{1}{\alpha}-1} \left[ \left( \frac{x}{\theta_1} \right)^{-1/\alpha} + \left( \frac{y}{\theta_2} \right)^{-1/\alpha} \right]^{\alpha-1}$	$-\frac{1}{y^2} + \theta_2^{-\alpha} y^{\alpha-1} \left[ \left( \frac{x}{\theta_1} \right)^\alpha + \left( \frac{y}{\theta_2} \right)^\alpha \right]^{-1/\alpha-1}$	$-\frac{1}{y^2} - \frac{\alpha+\theta}{(x+y)^2} + 2\frac{(2\alpha+\theta)x+(\alpha+\theta)y}{(x+y)^3}$
$V_{12}(x, y)$	$\frac{\alpha-1}{\alpha} (\theta_1 \theta_2)^{\frac{1}{\alpha}} (xy)^{-\frac{1}{\alpha}-1} \left[ \left( \frac{x}{\theta_1} \right)^{-1/\alpha} + \left( \frac{y}{\theta_2} \right)^{-1/\alpha} \right]^{\alpha-2}$	$-(\alpha+1)(\theta_1 \theta_2)^{-\alpha} (xy)^{\alpha-1} \left[ \left( \frac{x}{\theta_1} \right)^\alpha + \left( \frac{y}{\theta_2} \right)^\alpha \right]^{-1/\alpha-2}$	$\frac{6\alpha+4\theta}{(x+y)^3} - 6\frac{(2\alpha+\theta)x+(\alpha+\theta)y}{(x+y)^4}$
$A(w)$	$(1-\theta_1)(1-w) + (1-\theta_2)w + \left[ (1-w)^{\frac{1}{\alpha}} \theta_1^{\frac{1}{\alpha}} + w^{\frac{1}{\alpha}} \theta_2^{\frac{1}{\alpha}} \right]^\alpha$	$1 - \left[ \left( \frac{1-w}{\theta_1} \right)^{-\alpha} + \left( \frac{w}{\theta_2} \right)^{-\alpha} \right]^{-\frac{1}{\alpha}}$	$\theta w^3 + \alpha w^2 - (\alpha + \theta)w + 1$
Independence	$\alpha = 1$ or $\theta_1 = 0$ or $\theta_2 = 0$	$\alpha \rightarrow 1$ or $\theta_1 \rightarrow 0$ or $\theta_2 \rightarrow 0$	$\alpha = \theta = 0$
Total dependence	$\alpha \rightarrow 0$	$\alpha \rightarrow +\infty$	Never reached
Constraint	$0 < \alpha \leq 1, 0 \leq \theta_1, \theta_2 \leq 1$	$\alpha > 0, 0 < \theta_1, \theta_2 \leq 1$	$\alpha \geq 0, \alpha + 2\theta \leq 1, \alpha + 3\theta \geq 0$

### References

- P. Bortot and S. Coles. The multivariate Gaussian tail model: An application to oceanographic data. *Journal of the Royal Statistical Society. Series C: Applied Statistics*, 49(1):31–49, 2000.
- P. Bortot and J.A. Tawn. Models for the extremes of Markov chains. *Biometrika*, 85(4):851–867, 1998. ISSN 00063444.
- P. Capéraà, A.-L. Fougères, and C. Genest. A nonparametric estimation procedure for bivariate extreme value copulas. *Biometrika*, 84(3):567–577, 1997. ISSN 00063444.
- S. Coles. *An Introduction to Statistical Modelling of Extreme Values*. Springer Series in Statistics. Springer Series in Statistics, London, 2001.
- J.M. Cunderlik and T.B.M.J. Ouarda. Regional flood-duration-frequency modeling in the changing environment. *Journal of Hydrology*, 318(1-4):276–291, 2006.
- M. Falk and R. Michel. Testing for tail independence in extreme value models. *Annal. Inst. Stat. Math.*, 58(2):261–290, 2006. ISSN 00203157.
- L. Fawcett. *Statistical Methodology for the Estimation of Environmental Extremes*. PhD thesis, University of Newcastle upon Tyne, 2005.
- L. Fawcett and D. Walshaw. Markov chain models for extreme wind speeds. *Environmetrics*, 17(8):795–809, 2006. ISSN 11804009.
- C.A.T. Ferro and J. Segers. Inference for clusters of extreme values. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 65(2):545–556, 2003. ISSN 13697412.
- J. Galambos. Order statistics of samples from multivariate distributions. *Journal of the American Statistical Association*, 9:674–680, 1975.
- E.J. Gumbel. Bivariate exponential distributions. *Journal of the American Statistical Association*, 55(292):698–707, 1960.
- J.R.M. Hosking and J.R. Wallis. Parameter and Quantile Estimation for the Generalized Pareto Distribution. *Technometrics*, 29(3):339–349, 1987.
- H. Joe. Families of min-stable multivariate exponential and multivariate extreme value distributions. *Statist. Probab. Lett.*, 9:75–82, 1990.
- T.R. Kjeldsen and D. Jones. Estimation of an index flood using data transfer in the UK. *Hydrol. Sci. J.*, 52(1):86–98, 2007. ISSN 02626667.
- T.R. Kjeldsen and D.A. Jones. Prediction uncertainty in a median-based index flood method using L moments. *Water Resources Research*, 42(7):–, 2006. ISSN 00431397.
- D. Koutsoyiannis, D. Kozonis, and A. Manetas. A mathematical framework for studying rainfall intensity-duration-frequency relationships. *Journal of Hydrology*, 206(1-2):118–135, 1998.
- A.W. Ledford and J.A. Tawn. Statistics for near independence in multivariate extreme values. *Biometrika*, 83:169–187, 1996.
- A.W. Ledford and J.A. Tawn. Diagnostics for dependence within time series extremes. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 65(2):521–543, 2003.
- G. Lindgren and H. Rootzen. Extreme values: theory and technical applications. *Scandinavian journal of statistics*, 14(4):241–279, 1987.
- J. Pickands. Multivariate Extreme Value Distributions. In *Proceedings 43rd Session International Statistical Institute*, 1981.
- R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2007. URL <http://www.R-project.org>. ISBN 3-900051-07-0.
- D.S. Reis Jr. and J.R. Stedinger. Bayesian MCMC flood frequency analysis with historical information. *Journal of Hydrology*, 313(1-2):97–116, 2005. ISSN 00221694.

- 
- S.I. Resnick. *Extreme Values, Regular Variation and Point Processes*. New-York: Springer-Verlag, 1987.
- M. Ribatet. POT: Modelling Peaks Over a Threshold. *R News*, 7(1):34–36, April 2007.
- M. Ribatet, E. Sauquet, J.-M. Grésillon, and T.B.M.J. Ouarda. A regional Bayesian POT model for flood frequency analysis. *Stochastic Environmental Research and Risk Assessment (SERRA)*, 21(4):327–339, 2007a.
- M. Ribatet, E. Sauquet, J.-M. Grésillon, and T.B.M.J. Ouarda. Usefulness of the Reversible Jump Markov Chain Monte Carlo Model in Regional Flood Frequency Analysis. *Water Resources Research*, 43(8):W08403, 2007b. doi: 10.1029/2006WR005525.
- A.J. Robson and D.W. Reed. *Flood Estimation Handbook*, volume 3. Institute of Hydrology, Wallingford, 1999.
- M. Roche. *Hydrologie de surface*. Gauthier-Villars, Paris, 1963.
- E. Sauquet, M.-H. Ramos, L. Chapel, and P. Bernardara. Accepted in stream flow scaling properties: investigating characteristic scales from different statistical approaches. *Hydrological Processes*, 2008. doi: 10.1002/hyp.6952.
- L. Tallaksen and H. Van Lanen. *Hydrological Drought: Processes and Estimation Methods for Streamflow and Groundwater*, volume 48. Elsevier, 2004.
- J.A. Tawn. Bivariate extreme value theory: Models and estimation. *Biometrika*, 75(3):397–415, 1988.
- A. Werritty, J.L. Paine, N. Macdonald, J.S. Rowan, and L.J. McEwen. Use of multi-proxy flood records to improve estimates of flood risk: Lower River Tay, Scotland. *Catena*, 66(1-2):107–119, 2006. ISSN 03418162.

## 7.5 Conclusion et Critique de ce Modèle

Ce modèle [Ribatet *et al.*, 2007a] s'est révélé très efficace pour l'estimation des débits de pointe mais aussi très prometteur pour de futurs développements. Son originalité réside dans une utilisation complète des données dont peut disposer un hydrologue pour son analyse, notamment pour le contexte des stations partiellement jaugées.

Ce nouveau modèle a été validé sur un jeu de 50 longues séries à pas de temps journalier. Pour cette étude, six structures de dépendance extrême ont été étudiées : les dépendances logistique, logistique négative et mixte dans leurs versions symétriques et asymétriques.

Pour l'estimation des débits de pointe, il a été montré que les versions asymétriques sont plus pertinentes. Ce point est loin d'être surprenant puisque les hydrogrammes de crue sont également asymétriques. Ainsi, les modèles symétriques qui supposent que la modélisation du couple  $(y_{i-1}, y_i)$  est identique à celle de  $(y_i, y_{i-1})$  sont logiquement moins pertinents.

Parmi les modèles asymétriques, le modèle mixte montre les meilleures performances pour l'estimation des débits de pointe. Les résultats sont sans équivoque, comparé aux estimateurs locaux classiquement utilisés en hydrologie, le modèle mixte asymétrique est nettement plus performant. Ceci est encore bien plus vrai lorsque les périodes de retour considérées deviennent importantes - *i.e.*  $T \geq 20$  ans. En particulier, pour le quantile  $Q_{100}$ , l'erreur quadratique moyenne est divisée par trois par rapport aux estimateurs purement locaux traditionnellement utilisés.

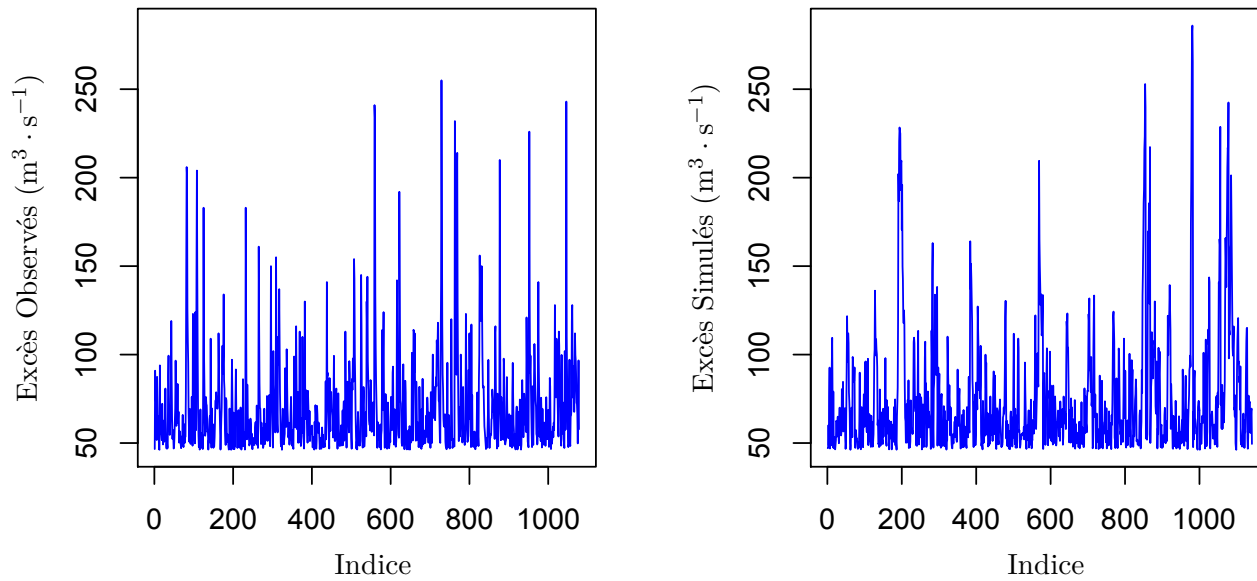
Comme notre modèle correspond à une chaîne de Markov, il est possible, une fois ce modèle ajusté, de simuler des chroniques continues. Ainsi, il nous a paru intéressant d'estimer des durées caractéristiques de crue. Ce modèle, appliqué tel qu'il a été présenté dans ce chapitre, semble pouvoir estimer raisonnablement une durée caractéristique de crue pour des bassins présentant des dynamiques rapides à modérés *i.e.* ayant un indice BFI [Tallaksen et Van Lanen, 2004] compris entre 0.28 et 0.4. Les résultats obtenus sont toutefois quelque peu décevant pour les autres bassins. Néanmoins, il n'existe pas à notre connaissance de méthode alternative pour obtenir in fine des hydrogrammes entièrement probabilisés. Par hydrogramme probabilisé, nous entendons que les observations successives de l'hydrogramme proviennent d'une loi de probabilité.

Enfin, au vu des résultats obtenus, ce modèle peut bien entendu être amélioré pour l'estimation des hydrogrammes de crue. En particulier, nous proposons à la fin de l'article quelques pistes. Ce sont pour les dynamiques lentes que notre modèle est le moins performant. La modélisation par une chaîne de Markov d'ordre un pourrait donc être trop pauvre en information et des ordres supérieurs pourraient être utilisés avec profit. Néanmoins, les problèmes numériques résultant de la complexité des distributions des valeurs extrêmes limitent leur utilisation aux dimensions inférieures ou égales à trois.

Une alternative plus facile à mettre en oeuvre serait de considérer un conditionnement adapté à notre chronique de débit. Dans la version présentée ici, le conditionnement correspond à la valeur précédente - *i.e.*  $\Pr[Y_t \leq y_t | Y_{t-1} = y_{t-1}]$ . Pour des bassins à dynamique lente, il serait probablement avantageux de considérer un conditionnement  $\Pr[Y_t \leq y_t | Y_{t-\delta} = y_{t-\delta}]$ ,  $\delta > 1$ . Par exemple, une valeur pour  $\delta$  pourrait être fonction de la durée de Socose [CTGREF, 1980]. Toutefois, sauf à redéfinir artificiellement le pas de temps de la chronique disponible, nous n'aurions plus affaire à des chaînes de Markov et la définition de la vraisemblance mais aussi la simulation de séries temporelles pourrait alors devenir problématique.

Une autre approche pour améliorer l'estimation de l'hydrogramme de crue serait de considérer d'autres structures extrêmes de dépendance. La MEVT nous apprend qu'il n'existe pas de paramétrisation finie pour ces structures, il paraît donc raisonnable de penser qu'il en existe de plus adaptées à la modélisation temporelle des débits que celles que nous avons considérées.

Ces pistes de recherche devraient être selon nous engagées car elles pourraient conduire à des outils puissants. Notamment, il serait possible de modéliser aléatoirement une durée caractéristique de crue contrairement au modèle QdF. Il serait également possible d'obtenir la distribution des



**Fig. 7.8:** Valeurs supérieures à  $46 \text{ m}^3 \cdot \text{s}^{-1}$  observées (à gauche) et simulés (à droite). La Sioule à St-Priest-des-Champs.

formes d'hydrogramme, sans contrainte de formulations mathématiques, pour un débit de pointe spécifique. Ce pourrait donc être une réelle avancée pour l'analyse fréquentielle des débits.

Toutefois, le modèle proposé n'est pas non plus sans défaut. En particulier, du fait de la complexité de la MEVT, la structure de dépendance imposée est utilisée comme une boîte noire. En effet, il est difficile de traduire la structure sous forme d'hypothèses physiques. Cela pourrait être un frein pour l'utilisation opérationnelle de ce type d'approches par rapport à des modèles dont les hypothèses hydrologiques sont plus explicites.



# Conclusions et Perspectives

# 8

**D**URANT notre travail de recherche, nous avons essayé de proposer des approches innovantes pour l'analyse fréquentielle des crues. Notre travail a porté sur le développement de nouvelles techniques statistiques. Par exemple, nous avons proposé l'utilisation des chaînes de Markov à sauts réversibles pour l'estimation des forts quantiles de crue. La théorie multivariée des valeurs extrêmes a également été la base de certains de nos développements.

Notre travail de recherche consistait initialement à tester la faisabilité et la performance des techniques d'estimation régionale. Partant du constat que le modèle régional de référence, *i.e.* l'indice de crue, utilisait l'information disponible au sein d'une région de manière peu rationnelle, nous nous sommes attaché à construire de nouvelles approches utilisant l'information de manière plus sensée. Ceci nous a conduit à développer deux modèles régionaux Bayésiens ; modèles distinguant l'information disponible au site cible de celle en provenance des autres sites de la région.

L'information hydrologique la plus précieuse réside dans les mesures contemporaines au site d'étude. Ainsi, l'information historique, tout comme l'information pluviométrique, doit être considérée comme de l'information additionnelle. En conséquence, elle doit être de moindre importance que les dernières mesures de débit. En effet, il n'existera jamais de modèle idéal permettant d'intégrer parfaitement cette information supplémentaire. Dès lors, en donnant trop de poids à ce type d'information extérieure, nous prenons le risque de biaiser nos estimations mais aussi de réduire faussement les incertitudes sur nos estimations. En effet, le but final n'est pas d'avoir l'intervalle de confiance le plus resserré possible pour  $Q_{100}$  mais d'obtenir l'incertitude réelle sur l'estimation de  $Q_{100}$ . En ce sens, nous avons essayé de proposer de nouveaux estimateurs, plus performants bien sûr, mais toujours *justes* quant aux incertitudes.

Le premier modèle développé fut un modèle régional Bayésien pour l'estimation des quantiles de crue. Il a permis de relaxer les hypothèses trop rigides imposées par l'indice de crue. Les équations déterministes imposées par ce modèle ont été assouplies à l'aide d'une loi *a priori* établie à partir de l'information régionale disponible. La relaxation des hypothèses de l'indice de crue a permis d'obtenir un modèle plus robuste face au degré d'hétérogénéité de la région et ainsi d'élargir son champ d'application. Il a également été montré qu'il ne semble pas opportun de travailler avec des régions comportant un faible nombre de stations mais fortement homogènes. En revanche, les régions comportant un nombre plus important de stations (supérieures à 14 pour notre application française) semblent plus adéquates pour l'estimation des quantiles, quitte à perdre en homogénéité.

Après avoir étudié la performance de ce premier modèle sur des données simulées, nous avons montré que, dans sa version initiale, ce modèle ne pouvait être utilisé avec efficacité que pour des périodes de retour inférieures à 20 ans. Nous avons donc suggéré un deuxième estimateur autorisant l'estimation pour des périodes de retour plus élevées. L'idée de ce nouveau modèle consiste à intégrer, au sein de la loi *a priori*, une mixture de deux lois GPD. La première étant une GPD conventionnelle et la deuxième ayant un paramètre de forme fixe représentant la cohérence régionale des sites de la région. Cette approche a été validée sur un grand nombre de régions générées stochastiquement et a montré sa supériorité face au modèle de l'indice de crue, y compris lorsque  $T > 20$  ans. Nous avons également montré que les approches Bayésiennes semblent plus robustes face aux erreurs



d'estimation de l'indice de crue, erreurs d'estimation non contrôlables.

Enfin, dans un troisième temps, nous avons proposé un modèle s'éloignant de notre première thématique de recherche. En effet, nous sommes passé d'une modélisation régionale à une modélisation purement locale. Ce dernier modèle, n'utilisant que l'information au site cible, est apparu comme une étape logique. En effet, avant de considérer une source d'information supplémentaire, ne devons nous pas nous questionner quant à l'utilisation de nos données au site d'étude? Les approches purement locales classiques sont très destructrices en information puisqu'elles réduisent une série temporelle quasiment continue à un échantillon restreint de maxima annuels ou de valeurs supérieures à un seuil supposées indépendantes.

A partir des derniers résultats concernant l'EVT, nous avons suggéré l'utilisation d'un modèle permettant la modélisation de tous les excès (nécessairement dépendants) au dessus d'un seuil de la chronique de débit. Pour notre étude, trois structures de dépendance extrême ont été considérées - dans leurs versions symétriques et asymétriques. Nous avons montré que les versions asymétriques sont les plus performantes puisqu'elles permettent de reproduire l'asymétrie observée des hydrogrammes. Plus particulièrement, une structure de dépendance précise s'est révélée plus pertinente : le modèle de dépendance mixte asymétrique. Avec cette structure, et sur un jeu de 50 stations françaises, nous avons obtenu des estimations 3 fois meilleures, en termes d'erreur quadratique moyenne, que celles issues des approches locales classiques.

De plus, puisque la dépendance temporelle est prise en compte, il est désormais possible de simuler des hydrogrammes de crue ; et donc d'estimer les volumes et durées associées. Deux structures de dépendance particulières ont été jugées plus aptes à reproduire la dynamique des crues. Toutefois, en l'état actuel, ce modèle a montré quelques difficultés pour des bassins à dynamiques lentes - *i.e.* supérieures à 5 jours. Ce type d'approche reste néanmoins prometteur. Nous avons proposé au sein de l'article initiateur quelques voies possibles pour parvenir à des modèles plus convaincants sur le problème de l'estimation des hydrogrammes de crues.

A terme, ce modèle pourrait, par exemple, permettre d'estimer la distribution de probabilité jointe du volume et d'une durée caractéristique de crue à débit de pointe fixé. L'estimation régionale pourrait également bénéficier de ces améliorations. En effet, il serait tout à fait possible de coupler les avancées statistiques présentes dans le dernier modèle au sein de modèles régionaux - notamment à sauts réversibles. Néanmoins, des étapes plus accessibles à court terme consisteraient à introduire la notion de dépendance dans nos modèles. Pour nos deux premiers modèles régionaux, il existe deux types de dépendance : (a) celle existant entre les paramètres du modèle et (b) celle présente entre les sites de la région homogène. Ces deux types de dépendance entraînent des conséquences bien différentes. La première réduit *l'espace de liberté* de la loi *a priori* et donc la variabilité de nos estimations. En revanche, pour la seconde, la redondance d'information potentielle existant entre plusieurs sites de la région homogène serait prise en compte. Ainsi, la *précision* de la loi *a priori* serait alors plus *honnête* et les incertitudes liées à nos estimations plus *vraies*.

Enfin, par construction, les trois modèles développés durant ce travail de recherche restent largement flexibles et peuvent sans difficulté s'adapter à d'autres contextes que ceux étudiés ici. Par exemple, les modèles développés aux Chapitres 5 et 6 pourraient utiliser l'information pluviométrique disponible sur le bassin afin de définir la loi *a priori*. La prise en compte de la non stationnarité est également accessible. Il suffirait par exemple, pour chacun de nos trois modèles, d'intégrer au sein de la vraisemblance une tendance fonction du temps. De même, par la flexibilité des chaînes de Markov à sauts réversibles, il serait possible d'établir des sauts entre un modèle stationnaire et non stationnaire. Cette approche aurait l'avantage de ne pas supposer initialement la présence d'une non stationnarité au sein des chroniques.

# Annexe : Éléments théoriques sur les L-moments

Les moments pondérés, introduit par [Greenwood \*et al.\* \[1979\]](#), présentent une alternative aux moments classiques et sont en lien avec les statistiques d'ordre. Ainsi, le moment pondéré  $M_{p,r,s}$  d'une variable aléatoire  $X$  ayant pour fonction de répartition  $F$  est donné par :

$$M_{p,r,s} = \mathbb{E} [X^p F(X)^r (1 - F(X))^s] \quad (\text{A.1})$$

Certains moments pondérés sont particulièrement intéressants pour l'inférence statistique puisqu'ils sont une caractérisation des paramètres de position, d'échelle et de forme. En particulier, les moments pondérés  $\alpha_r = M_{1,0,r}$  et  $\beta_r = M_{1,r,0}$  sont souvent utilisés [\[Hosking et Wallis, 1995\]](#).

$$\alpha_r = \int_0^1 x(u) (1-u)^r du, \quad \beta_r = \int_0^1 x(u) u^r du \quad (\text{A.2})$$

où la fonction quantile est notée  $x(u)$ .

Certaines combinaisons linéaires de  $\alpha_r$  et  $\beta_r$  permettent d'estimer les paramètres de position, d'échelle et de forme. Ces combinaisons linéaires, introduites initialement par [Silito \[1969\]](#) et reprises plus tard par [Hosking \[1990\]](#), sont définies selon la relation suivante :

$$\lambda_r = (-1)^r \sum_{k=0}^r p_{r,k}^* \alpha_k = \sum_{k=0}^r p_{r,k}^* \beta_k, \quad (\text{A.3})$$

où  $p_{r,k}^* = \frac{(-1)^{r-k} (r+k)!}{(k!)^2 (r-k)!}$ .

Ainsi, les quatre premiers  $L$ -moments sont donnés par :

$$\ell_1 = \alpha_0 = \beta_0, \quad (\text{A.4a})$$

$$\lambda_2 = \alpha_0 - 2\alpha_1 = 2\beta_1 - \beta_0, \quad (\text{A.4b})$$

$$\lambda_3 = \alpha_0 - 6\alpha_1 + 6\alpha_2 = 6\beta_2 - 6\beta_1 + \beta_0, \quad (\text{A.4c})$$

$$\lambda_4 = \alpha_0 - 12\alpha_1 + 30\alpha_2 - 20\alpha_3 = 20\beta_3 - 30\beta_2 + 12\beta_1 - \beta_0. \quad (\text{A.4d})$$

Comme pour les moments classiques, il est souvent préférable de travailler avec des moments adimensionnels comme le coefficient de variation, d'aplatissement ou d'asymétrie. Les  $L$ -moments autorisent également leur utilisation. Ainsi,

$$\tau = \frac{\lambda_2}{\ell_1}, \quad \tau_3 = \frac{\lambda_3}{\lambda_2}, \quad \tau_4 = \frac{\lambda_4}{\lambda_2} \quad (\text{A.5})$$

correspondent respectivement aux  $L$ -CV ( $L$ -Coefficient de Variation),  $L$ -Skewness ( $L$ -Coefficient d'asymétrie) et  $L$ -Kurtosis ( $L$ -Coefficient d'aplatissement).

L'estimation par les  $L$ -moments est basée sur la même approche que celle des moments classiques ; c'est à dire que les moments théoriques seront supposés égaux aux moments empiriques :

*(L-) moments empiriques = (L-) moments théoriques*

Hosking [1990] ont montré que les  $L$ -moments empiriques pouvaient être obtenus à partir de l'estimateur suivant :

$$\hat{\beta}_r = \frac{1}{n} \sum_{j=r+1}^n \frac{(j-1)(j-2)\cdots(j-r)}{(n-1)(n-2)\cdots(n-r)} x_{j:n} \quad (\text{A.6})$$

où  $\{x_{1:n}, x_{2:n}, \dots, x_{n:n}\}$  correspond à l'échantillon ordonné classé par ordre croissant.

Cette première méthode d'estimation est appelé dans la littérature *estimation par Moments Pondérés non biaisée (PWU)*. Une autre approche consiste à utiliser les *plotting position* pour estimer les  $L$ -moments empiriques. Hosking et Wallis [1995] ont montré que les  $L$ -moments empiriques pouvaient également être obtenus à partir de la relation :

$$\hat{\alpha}_r = \frac{1}{n} \sum_{i=1}^n (1 - p_{j:n})^r x_{j:n} \quad (\text{A.7})$$

où  $p_{j:n} = (j + \gamma) / (n + \delta)$  et  $-1 < \gamma < \delta$ . Hosking et Wallis [1995] préconisent la paramétrisation  $\gamma = -0.35$  et  $\delta = 0$  pour les lois de Wakeby, GEV et GPD. L'estimation à partir de l'équation (A.7) est appelée *estimation par la méthode des Moment Pondérés Biaisés (PWB)*. Néanmoins, cette approche comporte quelque inconvénients puisque l'invariance des rapports des  $L$ -moments n'est plus conservée par translation sur le paramètre de position - *i.e.* ajout d'une constante sur tout l'échantillon.

# Annexe : Méthodes d'Ajustement de la GPD



Dans cette annexe, nous allons regrouper les détails théoriques et algorithmiques permettant d'ajuster la GPD à un échantillon. Pour ce faire, supposons que la variable aléatoire  $X$  est distribuée selon une GPD ayant pour paramètre de position  $u$  (non estimé), d'échelle  $\sigma$  et de forme  $\xi$ . Supposons également que nous disposons d'un échantillon  $\mathbf{x} = \{x_1, \dots, x_n\}$  de taille  $n$ . Les abréviations des méthodes sont celles utilisées dans le mémoire.

## B.1 La Méthode des Moments (MOM)

Les moments de la GPD sont obtenus grâce à la formule générale :

$$\mathbb{E} \left[ \left( 1 + \xi \frac{X - u}{\sigma} \right)^r \right] = \frac{1}{1 - r\xi}, \quad 1 - r\xi > 0 \quad (\text{B.1})$$

Notons de plus que le moment d'ordre  $r$  existe si et seulement si  $\xi < 1/r$ . Ainsi, sous l'hypothèse que ces moments existent, la moyenne, la variance, l'aplatissement  $\gamma$  et l'asymétrie  $\kappa$  vérifient :

$$\mathbb{E}[X] = u + \frac{\sigma}{1 - \xi} \quad (\text{B.2a})$$

$$\text{Var}[X] = \frac{\sigma^2}{(1 - 2\xi)(1 - \xi)^2} \quad (\text{B.2b})$$

$$\gamma = \frac{2(1 + \xi)\sqrt{1 + 2\xi}}{1 - 3\xi} \quad (\text{B.2c})$$

$$\kappa = \frac{3(1 - 2\xi)(3 + \xi + 2\xi^2)}{(1 - 3\xi)(1 - 4\xi)} - 3 \quad (\text{B.2d})$$

En conséquence, l'estimation des paramètres d'échelle et de forme par la méthode des moments est donnée par les équations :

$$\hat{\sigma} = \frac{1}{2}\bar{x} \left( \frac{\bar{x}^2}{s^2} + 1 \right), \quad \hat{\xi} = -\frac{1}{2} \left( \frac{\bar{x}^2}{s^2} - 1 \right) \quad (\text{B.3})$$

où  $\bar{x}$  et  $s^2$  correspondent respectivement à la moyenne et la variance empirique de l'échantillon  $\mathbf{x}$ . Pour  $\xi = 0$  (*i.e.*  $X$  suit une loi Exponentielle), l'estimateur des moments est identique à celui du maximum de vraisemblance.

Lorsque  $\xi < 1/4$ ,  $\hat{\sigma}$  et  $\hat{\xi}$  sont asymptotiquement distribués selon une loi Normale donnée par :

$$n\text{Var} \begin{bmatrix} \hat{\sigma} \\ \hat{\xi} \end{bmatrix} = \frac{(1 - \xi)^2}{(1 - 2\xi)(1 - 3\xi)(1 - 4\xi)} \begin{bmatrix} 2\sigma^2(1 - 6\xi + 12\xi^2) & \sigma(2\xi - 1)(1 - 4\xi + 12\xi^2) \\ \sigma(2\xi - 1)(1 - 4\xi + 12\xi^2) & (1 - 2\xi)^2(1 - \xi + 6\xi^2) \end{bmatrix} \quad (\text{B.4})$$

## B.2 La Méthode des Moments Pondérés (PWU & PWB)

Supposons pour cette section que notre échantillon  $\mathbf{x}$  est ordonné de sorte que  $x_1 \leq x_2 \leq \dots \leq x_n$ . En reprenant les notations des  $L$ -moments introduites lors de l'Annexe A, l'estimation des paramètres de la GPD est obtenue à partir de la relation suivante :

$$\alpha_r = \mathbb{E}[X(1 - F(x))^r] = \frac{\sigma}{(r+1)(r+1-\xi)} \quad (\text{B.5})$$

En conséquence, les paramètres d'échelle et de forme de la GPD vérifient les relations suivantes :

$$\sigma = \frac{2\alpha_0\alpha_1}{\alpha_0 - 2\alpha_1}, \quad \xi = \frac{\alpha_0}{2\alpha_1 - \alpha_0} + 2, \quad \xi < 1 \quad (\text{B.6})$$

Ainsi les estimateurs  $\hat{\sigma}$  et  $\hat{\xi}$  des paramètres d'échelle et de forme sont obtenus en remplaçant  $\alpha_r$  par  $\hat{\alpha}_r$ , avec :

$$\hat{\alpha}_r = \frac{1}{n} \sum_{i=1}^n \frac{(n-i)(n-i-1)\dots(n-i-r+1)}{(n-1)(n-2)\dots(n-r)} x_i, \quad \text{pour les PWU} \quad (\text{B.7a})$$

$$\hat{\alpha}_r = \frac{1}{n} \sum_{i=1}^n (1 - p_{i:n})^r x_j, \quad \text{pour les PWB} \quad (\text{B.7b})$$

Hosking *et al.* [1985] ont montré que pour  $\xi < 0.5$ , les estimateurs  $\hat{\sigma}$  et  $\hat{\xi}$  sont asymptotiquement distribués selon une loi Normale vérifiant :

$$n\text{Var} \begin{bmatrix} \hat{\sigma} \\ \hat{\xi} \end{bmatrix} = \frac{1}{(1-2\xi)(3-2\xi)} \begin{bmatrix} \sigma^2(7-18\xi+11\xi^2-2\xi^3) & -\sigma(2-\xi)(2-6\xi+7\xi^2-2\xi^3) \\ -\sigma(2-\xi)(2-6\xi+7\xi^2-2\xi^3) & (1-\xi)(2-\xi)^2(1-\xi+2\xi^2) \end{bmatrix} \quad (\text{B.8})$$

## B.3 La Méthode du Maximum de Vraisemblance (MLE)

Hormis pour le cas  $\xi = 0$ , l'estimateur du maximum de vraisemblance ne possède pas de solution analytique. Il faut donc avoir recours à des procédures d'optimisation numérique pour obtenir une estimation des paramètres de la GPD par maximum de vraisemblance.

Bien souvent, il est plus pratique de maximiser la log-vraisemblance plutôt que la vraisemblance ; ainsi l'estimateur du maximum de vraisemblance s'attache à maximiser la quantité suivante :

$$\log \ell(\mathbf{x}; \sigma, \xi) = -n \log \sigma - \left( \frac{1}{\xi} + 1 \right) \sum_{i=1}^n \log \left( 1 + \xi \frac{x_i - u}{\sigma} \right) \quad (\text{B.9})$$

Du fait de problèmes numériques survenant lorsque le paramètre de forme  $\xi$  devient proche de 0, il est nécessaire de passer à la log-vraisemblance d'une loi Exponentielle lorsque  $|\xi| \leq \varepsilon$ ,  $\varepsilon$  petit, *i.e.*

$$\log \ell(\mathbf{x}; \sigma, 0) = -n \log \sigma - \sum_{i=1}^n \frac{x_i - u}{\sigma} \quad (\text{B.10})$$

L'une des propriétés fondamentales de l'estimateur du maximum de vraisemblance est qu'il est asymptotiquement distribué selon une loi Normale, *i.e.*

$$\hat{\theta} \longrightarrow \mathcal{N} \left( \theta, \mathcal{I}_n(\theta)^{-1} \right), \quad n \rightarrow +\infty \quad (\text{B.11})$$

où  $\hat{\theta}$  est l'estimateur par MLE,  $\theta$  la valeur théorique,  $n$  la taille de l'échantillon  $\mathbf{x}$  et  $\mathcal{I}_n(\theta)$  la matrice d'information de Fisher.

Pour la GPD, la matrice (théorique) d'information de Fisher est donnée par :

$$\mathcal{I}_n(\theta) = n \begin{pmatrix} \sigma^{-2}(1+2\xi)^{-1} & \sigma^{-1}(1+\xi)^{-1}(1+2\xi)^{-1} \\ \sigma^{-1}(1+\xi)^{-1}(1+2\xi)^{-1} & 2(1+\xi)^{-1}(1+2\xi)^{-1} \end{pmatrix} \quad (\text{B.12})$$

Rappelons également que, pour  $\xi > -0.5$ , l'estimateur du maximum de vraisemblance est asymptotiquement efficace - *i.e.* il atteint la borne de Cramer-Rao et est donc l'estimateur le plus précis asymptotiquement.

## B.4 La Méthode du « Likelihood Moment » (LME)

Dans cette section, nous décrivons les étapes logiques nécessaires pour utiliser l'estimateur introduit par Zhang [2007]. Comme pour l'estimateur des moments, l'estimateur LME utilise la relation suivante :

$$\mathbb{E}[(1+bY)^r] = \frac{1}{1-r\xi}, \quad 1-r\xi > 0 \quad (\text{B.13})$$

où  $b = -\xi/\sigma$ ,  $Y = X - u$  et  $r$  est une constante. L'équation (B.13), dans sa version empirique, est alors donnée par :

$$\frac{n^{-1} \sum_{i=1}^n (1+by_i)^r - (1-r\xi)^{-1}}{r} = 0 \quad (\text{B.14})$$

Zhang [2007] remarque que l'estimateur MOM est obtenu lorsque  $r$  vaut 1 et 2, alors que l'estimateur MLE correspond aux cas où  $r = -1$  et  $r \rightarrow 0$ . L'estimateur LME est solution de l'équation à une inconnue ( $b$ ) suivante :

$$\frac{1}{n} \sum_{i=1}^n (1+by_i)^p - \frac{1}{1-r} = 0, \quad b < \frac{1}{y_n} \quad (\text{B.15})$$

où  $p = rn / \sum_{i=1}^n \log(1+by_i)$  et  $r < 1$ .

Zhang [2007] montre que le membre de gauche de l'équation (B.15) est une fonction régulière de  $b$  sauf lorsque  $r = 0$  et  $x_1 = x_2 = \dots = x_n$ . De plus, l'équation (B.15) admet une unique solution  $\hat{b}$  sur  $(-\infty, y_n^{-1})$ , si  $r < 0.5$ ,  $r \neq 0$  et  $n > 2$ . Une fois  $\hat{b}$  obtenu, les paramètres d'échelle et de forme de la GPD sont facilement estimés à l'aide des relations :

$$\hat{\sigma} = -\frac{\hat{\xi}}{\hat{b}}, \quad \hat{\xi} = \frac{1}{n} \sum_{i=1}^n \log(1+\hat{b}y_i) \quad (\text{B.16})$$

Lorsque  $r < 0.5$  et  $\xi > -0.5$ , l'estimateur LME est également asymptotiquement distribué selon une loi Normale vérifiant :

$$n \text{Var} \begin{bmatrix} \hat{\sigma} \\ \hat{\xi} \end{bmatrix} = \begin{bmatrix} \sigma^2 \left( 2 + \frac{(r-\xi)^2 + 2\xi}{1-2r} \right) & \sigma \left( 1 + \frac{r^2 + \xi^2 + \xi}{1-2r} \right) \\ \sigma \left( 1 + \frac{r^2 + \xi^2 + \xi}{1-2r} \right) & (1-r) \left( 1 + \frac{2\xi^2 + 2\xi + r}{1-2r} \right) \end{bmatrix} \quad (\text{B.17})$$

Afin d'obtenir une estimation par LME, la constante  $r$  doit être fixée. Zhang [2007] montre que lorsque  $r \approx -\xi$ , l'estimateur LME est proche de l'estimateur maximum de vraisemblance. Toutefois, comme souvent aucune information ne sera disponible sur la valeur du paramètre de forme, il préconise l'utilisation de  $r = -0.5$ .

# Index

- Courbe de tarage, 7
- Crue maximale probable, 17
- Decclusterisation, 71
- Densité de probabilité, 2, 29, 73
- Durée caractéristique de crue, 76, 98
- Estimateur
  - Bayésien, 35, 101
  - local, 10, 52, 71
  - régional, 21, 26, 33, 35, 71, 101, 102
- Extrapolation, 2, 9
- Homogénéité, degré d', 54
- Hydrogramme, 98, 102
- Hyperparamètre, 29, 54, 69
- Incertitudes, 9, 19, 30, 69, 101, 102
- Indice
  - de crue, 25, 26, 33, 34, 50, 51, 53, 71
- Information
  - censurée, 11
  - historique, 11, 101
  - régionale, 35, 53
- Intervalle
  - de confiance, 31, 101
  - de crédibilité, 31, 36, 50
- Loi
  - a posteriori, 29, 30, 33
  - a priori, 29, 32, 33, 35, 50, 53, 102
  - de Fréchet, 3
  - de Gumbel, 3, 9
  - de Weibull Négative, 3
  - Exponentielle, 9
  - GEV, 4, 24
  - GPD, 4, 24, 34, 50
  - Kappa, 23
  - prédictive, 30
  - régionale, 26
- Markov, chaîne de, 73, 98, 101
- Modèle
  - Bayésien, 27, 51
  - du GRADEX, 13, 14
  - hydrologique, 16
  - Pluie-Débit, 12, 13, 16
- QdF, 76, 98
- Paramètre
  - d'échelle, 26, 34, 50
  - de forme, 9, 11, 26, 34, 50, 52
  - de position, 26, 34, 50
- Période de retour, 13, 50, 51, 53, 69, 98
- Queue de distribution, 2
- Région homogène, 21, 33, 35, 50, 51, 102
- Saut réversible, 69, 101, 102
- Stationnarité, 6
- Théorème
  - Central Limite, 3
  - de Bayes, 30, 33
  - de consistance, 35
  - de Normalité asymptotique, 36
  - de Pickands, 4
  - du type extremal, 3, 4
- Théorie
  - Bayésienne, 29
  - des valeurs extrêmes, 3, 26, 98, 102

# Liste des symboles

$\ell$	Fonction vraisemblance, page 12
$\mathcal{V}(x)$	Ensemble des voisinages de $x$ , page 35
$\hat{X}$	Estimateur de $X$ , page 3
$\mathbb{E}[X]$	Espérance de $X$ , page 105
$\mu$ (resp. $u$ )	Paramètre de position de la GEV (resp. GPD), page 4
$\pi$	Densité de probabilité - contexte Bayésien, page 29
$\pi(\theta)$	Loi <i>a priori</i> , page 30
$\Pr[X \leq x]$	Probabilité que $X$ soit inférieur à $x$ , page 3
$\sigma$	Paramètre d'échelle, page 4
$\sigma^{(R)}$	Paramètre d'échelle régional, page 34
$\sigma^{(S)}$	Paramètre d'échelle du site d'étude, page 34
$\tau, \tau_3, \tau_4$	2 <sup>e</sup> , 3 <sup>e</sup> et 4 <sup>e</sup> rapports des $L$ -moments pondérés, page 25
$\xi$	Paramètre de forme, page 4
$\xi^{(R)}$	Paramètre de forme régional, page 34
$\xi^{(S)}$	Paramètre de forme du site d'étude, page 34
$C^{(S)}$	Estimation de l'indice de crue au site cible, page 25
$F$	Fonction de répartition, page 3
$f$	Densité de probabilité, page 2
$H, H_1$	Statistiques d'homogénéité, page 25
$M_n$	Maximum de $n$ <i>v.a.</i> , page 3
$Q^{(R)}$	Fonction de répartition inverse régionale, page 25
$Q^{(S)}$	Fonction de répartition inverse au site cible, page 25
$Q_T$	Quantile de période de retour $T$ , page 52
$T$	Période de retour, page 13
$u^{(R)}$	Paramètre de position régional, page 34
$u^{(S)}$	Paramètre de position du site d'étude, page 34
$X \sim \pi(\theta)$	$X$ est distribué selon $\pi(\theta)$ , page 35
$Y$	Variable aléatoire, page 2





# Bibliographie

- P. ARNAUD : *Modèle de prédétermination de crues basé sur la simulation stochastique des pluies horaires*. Thèse de doctorat, Montpellier II, 1997.
- P. ARNAUD et J. LAVABRE : Coupled rainfall model and discharge model for flood frequency estimation. *Water Resources Research*, 38(6):111–1111, 2002. ISSN 00431397.
- M. BAYAZIT et B. ÖNÖZ : To prewhiten or not to prewhiten in trend analysis? *Hydrol. Sci. J.*, 52(4):611–624, 2007. ISSN 02626667.
- M. BAYES : An Essay towards Solving a Problem in the Doctrine of Chances. *Philosophical Transactions (1683-1775)*, 53:370–418, jan. 1763.
- A. BAYLISS et R. JONES : Peaks-over-threshold flood database : summary statistics and seasonality. *Report - UK Institute of Hydrology*, 121, 1993.
- M. BEABLE et A. MCKERCHAR : Regional flood estimation in New Zealand. *Water and Soil Technical Publication*, 20, 1981.
- L. BERTOLI et U. MOISELLO : A regional model for estimating the probability distribution of routed peak discharges. *Journal of Hydrology*, 153(1-4):103–138, 1994. ISSN 00221694.
- K. BEVEN : How far can we go in distributed hydrological modelling? *Hydrol. Earth Syst. Sci.*, 5(1):1–12, 2001. ISSN 10275606.
- P. BHUNYA, R. BERNDTSSON, C. OJHA et S. MISHRA : Suitability of Gamma, Chi-square, Weibull, and Beta distributions as synthetic unit hydrographs. *Journal of Hydrology*, 334(1-2):28–38, 2007. ISSN 00221694.
- P. BHUNYA, S. MISHRA, C. OJHA et R. BERNDTSSON : Parameter estimation of beta distribution for unit hydrograph derivation. *Journal of Hydrologic Engineering*, 9(4):325–332, 2004. ISSN 10840699.
- P. BLOOMFIELD : Trends in global temperatures. *Clim. Change*, 21:1–16, 1992.
- R. BRÁZDIL, Z. KUNDZEWICZ et G. BENITO : Historical hydrology for studying flood risk in Europe. *Hydrol. Sci. J.*, 51(5):739–764, 2006. ISSN 02626667.
- P. BROHAN, J. KENNEDY, I. HARRIS, S. TETT et P. JONES : Uncertainty estimates in regional and global observed temperature changes : A new data set from 1850. *J. Geophys. Res. D Atmos.*, 111(12), 2006. ISSN 01480227.
- D. BURN et M. HAG ELNUR : Detection of hydrologic trends and variability. *Journal of Hydrology*, 255(1-4):107–122, 2002. ISSN 00221694.
- L. CADAVID, J. OBEYSEKERA et H. SHEN : Flood frequency derivation from kinematic wave. *Journal of Hydrology Engng*, 117(4):489–510, 1991.

- O. CAYLA : Probabilistic calculation of design floods : SPEED. *In International Symposium on Engineering Hydrology*, San Francisco, July 1993. American Society of Civil Engineers.
- F. CHEBANA et T. OUARDA : Multivariate L-moment homogeneity test. *Water Resources Research*, Accepted(x):xx–xx, 2007.
- K. CHOKMANI et T. OUARDA : Physiographical space-based kriging for regional flood frequency estimation at ungauged sites. *Water Resources Research*, 40(12):1–13, 2004.
- S. COLES et M. DIXON : Likelihood-based inference for extreme value models. *Extremes*, 2(1):5–23, 1999.
- S. COLES et F. PAULI : Models and inference for uncertainty in extremal dependence. *Biometrika*, 89(1):183–196, 2002.
- S. COLES et L. PERICCHI : Anticipating catastrophes through extreme value modelling. *Journal of the Royal Statistical Society. Series C : Applied Statistics*, 52(4):405–416, 2003.
- S. COLES et J. TAWN : A Bayesian analysis of extreme rainfall data. *Journal of the Royal Statistical Society. Series C : Applied Statistics*, 45(4):463–478, 1996.
- COMITÉ FRANÇAIS DES GRANDS BARRAGES : Les crues de projet des barrages : Méthode du GRADEX. *In BARRAGES & RÉSERVOIRS*, éd. : 18ème congrès CIGB/ICOLD, vol. 2, p. 96, Novembre 1994.
- COMMITTEE ON SAFETY CRITERIA FOR DAMS : Safety of Dams : Flood and Earthquake Criteria. Rap. tech., National Research Council, Washington, D.C., 1985.
- CTGREF : Synthèse nationale sur les crues des petits bassins, fascicule 2 : la méthode Socose. Rap. tech., Cemagref, 1980.
- J. CUNDERLIK et D. BURN : Site-focused nonparametric test of regional homogeneity based on flood regime. *Journal of Hydrology*, 318(1-4):301–315, 2006a. ISSN 00221694.
- J. CUNDERLIK et D. BURN : Switching the pooling similarity distances : Mahalanobis for Euclidean. *Water Resources Research*, 42(3):-, 2006b. ISSN 00431397.
- J. CUNDERLIK et T. OUARDA : Regional flood-duration-frequency modeling in the changing environment. *Journal of Hydrology*, 318(1-4):276–291, 2006.
- J. CUNDERLIK, T. OUARDA et B. BOBÉE : On the objective identification of flood seasons. *Water Resources Research*, 40(1):W015201–W0152012, 2004. ISSN 00431397.
- T. DALRYMPLE : Flood frequency analysis. *U.S. Geol. Surv. Water Supply Pap.*, 1543 A, 1960.
- A. DAVISON et R. SMITH : Models for Exceedances over High Thresholds (with discussion). *Journal of the Royal Statistical Society. Series B. Statistical methodology*, 52:393–442, 1990.
- C. DAWSON, R. ABRAHART, A. SHAMSELDIN et R. WILBY : Flood estimation at ungauged sites using artificial neural networks. *Journal of Hydrology*, 319(1-4):391–409, 2006. ISSN 00221694.
- C. DE MICHELE et R. ROSSO : A multi-level approach to flood frequency regionalisation. *Hydrology and Earth System Sciences*, 6(2):185–194, 2002.
- W. DICKINSON, P. KELLY et H. WHITELEY : Extremes for rainfall and streamflow, how strong are the links? *Can. Water Resour. J.*, 17(3):224–236, 1992.

- J. DIEBOLT, A. GUILLOU et I. RACHED : A new look at probability-weighted moments estimators. *C. R. Math.*, 338(8):629–634, 2004. ISSN 1631073X (ISSN).
- E. DOUGLAS, R. VOGEL et C. KROLL : Trends in floods and low flows in the United States : Impact of spatial correlation. *Journal of Hydrology*, 240(1-2):90–105, 2000. ISSN 00221694.
- A. DUMAS : Maximization method : Estimation of maximum probable flood (PMF). *La Houille Blanche*, 5:74–79, 2006. ISSN 00186368.
- D. DUPUIS : Exceedances over High Thresholds : A Guide to Threshold Selection. *Extremes*, 1(3):251–261, 1998.
- P. EAGLESON : Dynamics of flood frequency. *Water Resources Research*, 8(4):878–898, 1972.
- K. ENGELAND, H. HISDAL et A. FRIGESSI : Practical extreme value modelling of hydrological floods and droughts : A case study. *Extremes*, 7(1):5–30, 2005. ISSN 13861999.
- M. FALK et R. MICHEL : Testing for tail independence in extreme value models. *Annal. Inst. Stat. Math.*, 58(2):261–290, 2006. ISSN 00203157.
- A. FAVRE, S. EL ADLOUNI, L. PERREAULT, N. THIEMONGE et B. BOBÉE : Multivariate hydrological frequency analysis using copulas. *Water Resources Research*, 40(1):W01101, jan. 2004.
- C. FERRO et J. SEGERS : Inference for clusters of extreme values. *Journal of the Royal Statistical Society. Series B : Statistical Methodology*, 65(2):545–556, 2003. ISSN 13697412.
- R. FISHER et L. TIPPETT : Limiting forms of the frequency distribution of the largest or smallest member of a sample. *In Proceedings of the Cambridge Philosophical Society*, vol. 24, p. 180–190, 1928.
- G. GALÉA et C. PRUDHOMME : Basic notions and useful concepts for understanding the modeling of flood regimes of basins in QdF models. *Rev. Sci. Eau*, 10(1):83–101, 1997. ISSN 09927158.
- R. GARÇON : Prévision opérationnelle des apports de la Durance à Serre-Ponçon à l’aide du modèle MORDOR. *La Houille Blanche*, 5, 1996.
- GIECC : Résumé à l’intention des décideurs. *In Bilan 2007 des changements climatiques : Impacts, adaptation et vulnérabilité. Contribution du Groupe de Travail II au quatrième rapport d’évaluation du Groupe d’experts intergouvernemental sur l’évolution du climat*, p. 19, 2007.
- D. GINGRAS et K. ADAMOWSKI : Homogeneous region delineation based on annual flood generation mechanisms. *Hydrological Sciences Journal*, 38(2):103–121, 1993.
- N. GOEL, R. KUROTHE, B. MATHUR et R. VOGEL : A derived flood frequency distribution for correlated rainfall intensity and duration. *Journal of Hydrology*, 228(1-2):56–67, 2000. ISSN 00221694.
- L. GOTTSCHALK, L. TALLAKSEN et G. PERZYNA : Derivation of low flow distribution functions using recession curves. *Journal of Hydrology*, A94(1-4):239–262, 1997.
- J. GREENWOOD, J. LANDWEHER et N. NATALES : Probability weighted moments : definition and relation to parameters of several distributions expressible in inverse form. *Water Resources Research*, 15:1049–1054, 1979.
- P. GROISMAN et D. EASTERLING : Variability and trends in total precipitation and snowfall over the United States and Canada. *J. Clim.*, 10:184–205, 1994.

- P. GROVER, D. BURN et J. CUNDERLIK : A comparison of index flood estimation procedures for ungauged catchments. *Canadian Journal of Civil Engineering*, 29(5):734–741, 2002. ISSN 03151468.
- P. GUILLOT et D. DUBAND : La méthode du Gradex pour le calcul de la probabilité des crues à partir des pluies. *In Symposium International d'hydrologie*, Fort Collins, 1967.
- E. GUMBEL : Distribution des valeurs extrêmes en plusieurs dimensions. Rap. tech., Publication de l'Institut Statistique de l'Université Paris 9, 1960.
- V. K. GUPTA, O. J. MESA et D. R. DAWDY : Multiscaling theory of flood peaks : Regional quantile analysis. *Water Resources Research*, 30(12):3405–3421, 1994.
- T. HAKTANIR et N. SEZEN : Suitability of two-parameter gamma and three-parameter beta distributions as synthetic unit hydrographs in Anatolia. *Hydrological Sciences Journal/Journal des Sciences Hydrologiques*, 35(2):167–184, 1990.
- J. HANSEN, R. RUEDY, M. SATO, M. IMHOFF, W. LAWRENCE, D. EASTERLING, T. PETERSON et T. KARL : A closer look at United States and global surface temperature change. *J. Geophys. Res. D Atmos.*, 106(D20):23947–23963, 2001. ISSN 01480227.
- J. HOSKING : L-moment Analysis and Estimation of Distribution using Linear Combination of Order Statistics. *Journal of the Royal Statistical Society. Series B : Statistical Methodology*, 52:105–124, 1990.
- J. HOSKING et J. WALLIS : Parameter and Quantile Estimation for the Generalized Pareto Distribution. *Technometrics*, 29(3):339–349, 1987.
- J. HOSKING et J. WALLIS : Some statistics useful in regional frequency analysis. *Water Resources Research*, 29(2):271–281, 1993.
- J. HOSKING et J. WALLIS : A comparison of unbiased and plotting-position estimators of L moments. *Water Resources Research*, 31(8):2019–2025, 1995.
- J. HOSKING et J. WALLIS : *Regional Frequency Analysis*. Cambridge University Press, 1997.
- J. HOSKING, J. WALLIS et E. WOOD : Estimation of the Generalized Extreme Value Distribution by the Method of Probability Weighted Moments. *Technometrics*, 27(3):251–261, 1985.
- P. JAVELLE : *Caractérisation du régime des crues : le modèle débit-durée-fréquence convergent. Approche locale et régionale*. Thèse de doctorat, Institut National Polytechnique de Grenoble, 2001.
- P. JAVELLE, T. OUARDA, M. LANG, B. BOBÉE, G. GALÉA et J.-M. GRÉSILLON : Development of regional flood-duration-frequency curves based on the index-flood method. *Journal of Hydrology*, 258(1-4):249–259, 2002.
- A. JENKINSON : The frequency distribution of the annual maximum (or minimum) values of meteorological events. *Quarterly Journal of the Royal Meteorological Society*, 81:158–172, 1955.
- S. JUÁREZ et W. SCHUCANY : Robust and efficient estimation for the generalized pareto distribution. *Extremes*, 7(3):237–251, 2004. ISSN 13861999.
- T. KARL et R. KNIGHT : Secular trends of precipitation amount, frequency and intensity in the United States. *Bull. Am. Meteorol. Soc.*, 79(2):231–241, 1998.

- R. KATZ, M. PARLANGE et P. NAVEAU : Statistics of extremes in hydrology. *Advances in Water Resources*, 25(8-12):1287–1304, 2002.
- T. KJELDSEN et D. JONES : Estimation of an index flood using data transfer in the UK. *Hydrol. Sci. J.*, 52(1):86–98, 2007. ISSN 02626667.
- T. KJELDSEN, J. SMITHERS et R. SCHULZE : Regional flood frequency analysis in the kwazulu-natal province, South Africa, using the index-flood method. *Journal of Hydrology*, 255(1-4):194–211, 2002.
- D. KOUTSOYANNIS : Statistics of extremes and estimation of extreme rainfall : Theoretical investigation. *Hydrological Sciences Journal*, 49(4):575–590, 2004.
- D. KOUTSOYIANNIS, D. KOZONIS et A. MANETAS : A mathematical framework for studying rainfall intensity-duration-frequency relationships. *Journal of Hydrology*, 206(1-2):118–135, 1998.
- Z. KUNDZEWICZ et A. ROBSON : Change detection in hydrological records - A review of the methodology. *Hydrol. Sci. J.*, 49(1):7–20, 2004. ISSN 02626667.
- M. LANG, T. OUARDA et B. BOBÉE : Towards operational guidelines for over-threshold modeling. *Journal of Hydrology*, 225(3-4):103–117, 1999.
- M. LANG, C. PERRET, E. RENOUF et E. SAUQUET : Incertitudes sur les débits de crue. *La Houille Blanche*, 6:33–41, 2006.
- C. LARSON et B. REICH : Relationship of observed rainfall and runoff recurrence intervals. In E. SCHULZ, V. KOELZER et K. MAHMOOD, édés : *Floods and Droughts*, p. 34–43, Fort Collins, 1972. Water Resources Publications.
- L. LAVE et T. BALVANYOS : Risk analysis and management of dam safety. *Risk Anal.*, 18(4):455–462, 1998. ISSN 02724332.
- M. LEADBETTER : Extremes and local dependence in stationary sequences. *Probability Theory and Related Fields (Historical Archive)*, 65(2):291–306, 1983.
- A. LEDFORD et J. TAWN : Statistics for near independence in multivariate extreme values. *Biometrika*, 83:169–187, 1996.
- H. LINS et J. SLACK : Streamflow trends in the United States. *Geophys. Res. Lett.*, 26:227–230, 1999.
- A. LOUKAS : Flood frequency estimation by a derived distribution procedure. *Journal of Hydrology*, 255(1-4):69–89, 2002. ISSN 00221694.
- H. MADSEN et D. ROSBJERG : Generalized least squares and empirical Bayes estimation in regional partial duration series index-flood modeling. *Water Resources Research*, 33(4):771–781, 1997.
- M. MARGOUM, G. OBERLIN, M. LANG et R. WEINGARTNER : Flood estimation : Agregee concepts. *Hydrologie Continentale*, 9(1):85–100, 1995. ISSN 02461528.
- R. MERZ et G. BLÖSCHL : Flood frequency regionalisation - Spatial proximity vs. catchment attributes. *Journal of Hydrology*, 302(1-4):283–306, 2005. ISSN 00221694.
- R. MERZ, G. BLÖSCHL et U. PLOCK-ELLENA : Applicability of the gradex-method in Austria. *Osterr. Wasser- Abfallwirtsch.*, 51(11-12):291–305, 1999. ISSN 0945358X.

- S. MOSQUERA-MACHADO et S. AHMAD : Flood hazard assessment of Atrato River in Colombia. *Water Resour. Manage.*, 21(3):591–609, 2007. ISSN 09204741.
- M. MOUGHAMIAN, D. McLAUGHLIN et R. BRAS : Estimation of flood frequency : an evaluation of tow derived distribution procedures. *Water Resources Research*, 23(7):1309–1319, 1987.
- A. MULLER : *Comportement asymptotique de la distribution des pluies extrêmes en France*. Thèse de doctorat, Université de Montpellier II, 2006. URL <http://www.lyon.cemagref.fr/doc/these/muller/index.shtml>.
- R. MUTTIAH, R. SRINIVASAN et P. ALLEN : Prediction of two-year peak stream discharges using neural networks. *J. AM. WATER RESOUR. ASSOC.*, 33(3):625–630, 1997. ISSN 1093474X.
- M. NAGHETTINI, K. POTTER et T. ILLANGASEKARE : Estimating the upper tail of flood-peak frequency distributions using hydrometeorological information. *Water Resources Research*, 32(6):1729–1740, 1996. ISSN 00431397.
- R. NAULET, M. LANG, T. OUARDA, D. COEUR, B. BOBÉE, A. RECKING et D. MOUSSAY : Flood frequency analysis on the ardèche river using french documentary sources from the last two centuries. *Journal of Hydrology*, 313(1-2):58–78, 2005. ISSN 00221694.
- P. NORTHROP : Likelihood-based approaches to flood frequency estimation. *Journal of Hydrology*, 292(1-4):96–113, 2004.
- G. ODIVO et G. HEIZ : How long must we wait chef Eric? *The international journal of wildland fire*, 3(;):193–204, 2007.
- J. OLSEN, J. STEDINGER, N. MATALAS et E. STAKHIV : Climate variability and flood frequency estimation for the upper Mississippi and lower Missouri rivers. *Water Resour. Assoc.*, 35(6):1509–1520, 1999.
- H. ONIBON, T. OUARDA, B. BOBÉE, M. BARBET, A. ST-HILAIRE et P. BRUNEAU : Regional frequency analysis of annual maximum daily precipitation in Quebec, Canada [Analyse fréquentielle régionale des précipitations journalières maximales annuelles au Québec, Canada]. *Hydrological Sciences Journal*, 49(4):717–735, 2004.
- T. OUARDA, J. CUNDERLIK, A. ST-HILAIRE, M. BARBET, P. BRUNEAU et B. BOBÉE : Data-based comparison of seasonality-based regional flood frequency methods. *Journal of Hydrology*, 330(1-2):329–339, 2006. ISSN 00221694.
- T. OUARDA, C. GIRARD, G. CAVADIAS et B. BOBÉE : Regional flood frequency estimation with canonical correlation analysis. *Journal of Hydrology*, 254(1-4):157–173, 2001.
- G. PANDEY et V.-T.-V. NGUYEN : A comparative study of regression based methods in regional flood frequency analysis. *Journal of Hydrology*, 225(1-2):92–101, 1999. ISSN 00221694.
- E. PAQUET, J. GAILHARD et R. GARÇON : Evolution of the GRADEX method : Improvement by atmospheric circulation classification and hydrological modelling. *La Houille Blanche*, 5:80–90, 2006. ISSN 00186368.
- E. PARENT et J. BERNIER : Bayesian POT modeling for historical data. *Journal of Hydrology*, 274(1-4):95–108, 2003. ISSN 00221694.
- C. PEARSON : New Zealand regional flood frequency analysis using L-moments. *Journal of Hydrology (New Zealand)*, 30(2):53–64, 1991.

- L. PENG et A. WELSH : Robust Estimation of the Generalized Pareto Distribution. *Extremes*, 4 (1):53–65, 2001.
- J. PICKANDS : Statistical Inference Using Extreme Order Statistics. *Annals of Statistics*, 3:119–131, 1975.
- A. RAHMAN, P. WEINMANN, T. HOANG et E. LAURENSEN : Monte Carlo simulation of flood frequency curves from rainfall. *Journal of Hydrology*, 256(3-4):196–210, 2002. ISSN 00221694.
- T. RAINES et J. VALDES : Estimation of flood frequencies for ungauged catchments. *Journal of Hydrology Engng*, 119(10):1138–1154, 1993.
- D. REED : On the GRADEX method of estimating extreme floods. *Dams & Reservoirs*, p. 17–19, June 1994.
- D. REIS JR. et J. STEDINGER : Bayesian MCMC flood frequency analysis with historical information. *Journal of Hydrology*, 313(1-2):97–116, 2005. ISSN 00221694.
- B. RENARD, V. GARRETA et M. LANG : An application of Bayesian analysis and Markov chain Monte Carlo methods to the estimation of a regional trend in annual maxima. *Water Resources Research*, 42(W12422), 2006a. ISSN 00431397.
- B. RENARD, M. LANG et P. BOIS : Statistical analysis of extreme events in a non-stationary context via a Bayesian framework : case study with peak-over-threshold data. *Stochastic Environmental Research and Risk Assessment (SERRA)*, p. 1–16, 2006b.
- E. RENOUF, M. LANG, E. SAUQUET et A. PAQUIER : Contrôle de la qualité des courbes de tarage de la banque HYDRO pour les débits de crue. Rap. tech., Ministère de l'Écologie et du Développement Durable, Août 2005.
- S. RESNICK : *Extreme Values, Regular Variation and Point Processes*. New-York : Springer-Verlag, 1987.
- M. RIBATET, T. OUARDA, E. SAUQUET et J.-M. GRÉSILLON : Modeling All Exceedances Above a Threshold Using an Extremal Dependence Structure : Inferences on Several Flood Characteristics. *Water Resources Research*, Submitted(x):xx–xx, 2007a.
- M. RIBATET, E. SAUQUET, J.-M. GRÉSILLON et T. OUARDA : A regional Bayesian POT model for flood frequency analysis. *Stochastic Environmental Research and Risk Assessment (SERRA)*, 21 (4):327–339, 2007b.
- M. RIBATET, E. SAUQUET, J.-M. GRÉSILLON et T. OUARDA : Usefulness of the Reversible Jump Markov Chain Monte Carlo Model in Regional Flood Frequency Analysis. *Water Resources Research*, 43(8):W08403, 2007c.
- J. RIBEIRO-CORREA, G. CAVADIAS, B. CLEMENT et J. ROUSSELLE : Identification of hydrological neighborhoods using canonical correlation analysis. *Journal of Hydrology*, 173(1-4):71–89, 1995. ISSN 00221694.
- J. ROBINSON et M. SIVAPALAN : An investigation into the physical causes of scaling and heterogeneity of regional flood frequency. *Water Resources Research*, 33(5):1045–1059, 1997.
- A. ROBSON et D. REED : *Flood Estimation Handbook*, vol. 3. Institute of Hydrology, Wallingford, 1999.



- M. RULLI et R. ROSSO : An integrated simulation method for flash-flood risk assessment : 1. Frequency predictions in the Bisagno River by combining stochastic and deterministic methods. *HESS*, 6:267–283, 2002.
- B. SACKL et H. BERGMANN : A Bivariate frequency model and its application. In V. SINGH, éd. : *Hydrologic Frequency Modeling*, p. 571–582. Reidel Publishing Company, 1987.
- E. SAUQUET, P. JAVELLE et S. LE CLERC : Description des régimes hydrologiques des hautes-eaux : une nouvelle formulation pour l’analyse en débit-durée-fréquence et applications en ingénierie. *INGÉNIERIE - EAT*, 34:3–15, Juin 2003.
- M. SCHAEFER : Regional analyses of precipitation annual maxima in Washington State. *Water Resources Research*, 26(1):119–131, 1990.
- L. SHERMAN : Streamflow from rainfall by the unit-graph method. *Eng. News-Rec.*, 108:501–505, 1932.
- G. SILITO : Derivation of approximants to the inverse distribution function of a continuous population. *Biometrika*, 38:377–382, 1969.
- R. SMITH, J. TAWN et S. COLES : Markov chain models for threshold exceedances. *Biometrika*, 84(2):249–268, 1997. ISSN 00063444.
- L. TALLAKSEN et H. VAN LANEN : *Hydrological Drought : Processes and Estimation Methods for Streamflow and Groundwater*, vol. 48. Elsevier, 2004.
- J. TAWN : Bivariate extreme value theory : Models and estimation. *Biometrika*, 75(3):397–415, 1988.
- A. VIGLIONE, F. LAIO et P. CLAPS : A comparison of homogeneity tests for regional frequency analysis. *Water Resources Research*, 43(3):W03428, 2007. ISSN 00431397.
- R. VON MISES : La distribution de la plus grande de  $n$  valeurs. *American Mathematical Society*, 2 (Selected Papers):271–294, 1954.
- S. WILTSHIRE : Grouping basins for regional flood frequency analysis. *Hydrological Sciences Journal/Journal des Sciences Hydrologiques*, 30(1):151–159, 1985.
- S. WILTSHIRE : Regional flood frequency analysis. i : homogeneity statistics. *Hydrological Sciences Journal/Journal des Sciences Hydrologiques*, 31(3):321–333, 1986.
- Z. YANG et D. HAN : Derivation of unit hydrograph using a transfer function approach. *Water Resources Research*, 42(W01501), 2006. ISSN 00431397.
- S. YUE, T. OUARDA, B. BOBÉE, P. LEGENDRE et P. BRUNEAU : Approach for describing statistical properties of flood hydrograph. *Journal of Hydrologic Engineering*, 7(2):147–153, 2002. ISSN 10840699.
- J. ZHANG : Likelihood Moment Estimation for the Generalized Pareto Distribution. *Australian & New Zealand Journal of Statistics*, 49(1):69–77, 2007.

# Consolidation de l'information hydrologique disponible localement et régionale-ment pour l'estimation probabiliste du régime des crues

## Résumé :

Le praticien, lors de l'étape de prédétermination des débits de crue, est souvent confronté à un jeu de données restreint. Dans notre travail de recherche, nous avons proposé trois nouveaux modèles probabilistes spécialement conçus pour l'estimation des caractéristiques du régime des crues en contexte partiellement jaugé. Parmi ces modèles, deux d'entre eux sont des modèles dits *régionaux*, *i.e.* intégrant de l'information en provenance de stations ayant un comportement réputé similaire à celui du site étudié. Ces modèles, basés sur la théorie Bayésienne, ont montré une grande robustesse au degré d'hétérogénéité des sites appartenant à la région. De même, il est apparu que pour l'estimation des forts quantiles ( $T \geq 50$  ans), l'idée d'un paramètre régional contrôlant l'extrapolation est pertinente mais doit d'être intégrée de manière souple et non imposée au sein de la vraisemblance. L'information la plus précieuse dont le praticien dispose étant celle en provenance du site d'étude, le troisième modèle proposé revient sur l'estimation à partir des seules données contemporaines au site d'étude. Ce nouveau modèle utilise une information plus riche que celle issue d'un échantillonnage classique de *v.a.i.id.* maximales puisque toute la chronique est exploitée. Dès lors, même avec seulement cinq années d'enregistrement et grâce à une modélisation de la dépendance entre les observations successives, la taille des échantillons exploités est alors bien plus importante. Nous avons montré que pour l'estimation des quantiles de crues, ce modèle surpasse très nettement les approches locales classiquement utilisées en hydrologie. Ce résultat est d'autant plus vrai lorsque les périodes de retour deviennent importantes. Enfin, par construction, cette approche permet également d'obtenir une estimation probabiliste de la dynamique des crues.

## Efficient use of local and regional hydrological information for flood frequency analysis

### Abstract:

To define the design flood, practitioners must often deal with only few data available. The aim of this work was to propose new classes of probabilistic models that are more accurate for this kind of applications. In this perspective, we propose three different models: two regional approaches and a fully local one. Unlike fully local models, the regional approaches include information from other gauging stations. Our results show that the proposed regional Bayesian estimators are more robust to the discordancy degree of the sites within the region. In addition, for larger quantile estimation ( $T \geq 50$  years), the concept of a regional parameter which controls the tail behaviour seems to be relevant. However, this concept has to be proposed and not imposed within the likelihood function. It is overwhelmingly clear that the most important information one disposes is the target site one. To this aim, we propose a third model that is fully local, *i.e.*, which only uses the latest recorded data. This new model is innovative as the whole time series is involved in the estimation procedure; not only cluster maxima. Consequently, even with only a five years record length time series, the sample size becomes large. Our results show that, for flood quantile estimations, this model clearly outperforms the estimators conventionally used in hydrology. Furthermore, by definition, this model allows inferences on flood dynamics.

**Mots Clés :** Théorie des Valeurs Extrêmes, Analyse Fréquentielle, Estimation Régionale, Méthodes MCMC

**Discipline :** Hydrologie Statistique