REVUE DE STATISTIQUE APPLIQUÉE

A. GUILLOU
P. WILLEMS

Application de la théorie des valeurs extrêmes en hydrologie

Revue de statistique appliquée, tome 54, n° 2 (2006), p. 5-31
<http://www.numdam.org/item?id=RSA_2006__54_2_5_0>

Rev. Statistique Appliquée, 2006, **LIV** (2), 5-31

APPLICATION DE LA THÉORIE DES VALEURS EXTRÊMES EN HYDROLOGIE

A. GUILLOU[1], P. WILLEMS[2]

(1) *Université Paris VI, L.S.T.A., Boîte 158, 175 rue du Chevaleret, 75013 Paris, France*
(2) *Laboratoire d'Hydraulique, Katholieke Universiteit Leuven, Kasteelpark Arenberg 40, 3001 Heverlee, Belgium*

RÉSUMÉ

Les distributions des valeurs extrêmes généralisées (GEV) et de Pareto généralisées (GPD) sont très utilisées dans l'analyse des valeurs extrêmes en hydrologie. La calibration de ces distributions au-delà d'un seuil pose cependant un certain nombre de problèmes et en particulier celui du biais dans les queues de distributions. Une méthode est proposée pour quantifier ce biais, méthode basée sur la calibration d'une fonction à variations lentes. Tenir compte de cette fonction permet entre autre d'obtenir des estimateurs de quantiles plus précis. Par ailleurs, il est nécessaire de supposer que les observations proviennent d'une seule et même distribution. Cette hypothèse est satisfaite à condition que l'on puisse garantir une influence faible des inondations et des modifications artificielles telles que les structures régularisant le courant, les réservoirs, les digues, ... Dans le cas contraire, l'analyse devra être effectuée sur la base de données censurées, en se basant uniquement sur les observations non influencées par ces divers phénomènes physiques. Les distributions des points « non inondés » et des points « inondés » seront alors toutes deux nécessaires. Nous présenterons une approche permettant d'appréhender ce type de problème et nous illustrerons son efficacité sur des données réelles de débits horaires pour la rivière Molenbeek à Erpe-Mere (Belgique).

Mots-clés : *Analyse des valeurs extrêmes, réduction de biais, censure, écoulements, débits, inondations*

ABSTRACT

The Generalized Extreme-Value (GEV) and Generalized Pareto Distributions (GPD) are most commonly used in hydrological extreme-value analysis. Calibration of these distributions above a threshold level leads, however, to a bias in the asymptotic properties of the extreme-value distribution's tail. A methodology is proposed to quantify this bias based on the calibration of a so-called slowly varying function. This function allows to conduct bias corrections, and to increase the accuracy of quantile estimations. Moreover, it is necessary to assume that the observations come from the same distribution. This assumption is satisfied if the flooding or man-made influences (such as flow regulating structures, reservoir, dikes, etc.) are small. In the other case, the extreme-value analysis has to be carried out using truncation : the analysis has to consider the existence of higher events, but should not take their values into consideration for

the calibration of the extreme-value distribution. The distributions of the « non-flooded » points and « flooded » points will be both useful. We describe a technique taking this phenomenon into account and we illustrate its efficiency using hourly discharges for the river Molenbeek at Erpe-Mere (Belgium).

Keywords : *Extreme-value analysis, bias reduction, censoring, rainfall-runoff, discharges, floods*

1. Introduction

Le modèle sur lequel se base toute la théorie des valeurs extrêmes est donné par le résultat de Gnedenko (1943) qui décrit les limites possibles de la loi du maximum de n variables aléatoires indépendantes et identiquement distribuées (i.i.d.), convenablement normalisé. Nous supposerons donc toujours que la loi qui régit le phénomène auquel nous nous intéressons est dans le domaine d'attraction d'une loi des extrêmes (GEV) H_γ, où γ est un paramètre réel. Si nous considérons un échantillon $X,\ X_1, ...,\ X_n$ de même loi F, cela signifie qu'il existe deux suites normalisantes (α_n) (dans $\mathbb{R}$) et (σ_n) (dans $\mathbb{R}_+$) telles que

$$\forall x \in \mathbb{R} \quad \lim_{n\to\infty} F^n(\sigma_n x + \alpha_n) = H_\gamma(x), \tag{1}$$

où

$$H_\gamma(x) = \begin{cases} \exp(-(1+\gamma x)^{-\frac{1}{\gamma}}) & \text{pour tout } x \text{ tel que } 1+\gamma x > 0, \text{ si } \gamma \neq 0, \\ \exp(-\exp(-x)) & \text{pour tout } x \in \mathbb{R}, \text{ si } \gamma = 0. \end{cases}$$

Ce résultat implique de façon évidente que le comportement de la queue de distribution dépend d'un unique paramètre, noté γ, et appelé *indice des valeurs extrêmes*. Le signe de ce paramètre est un indicateur essentiel sur le comportement de cette queue. En effet, trois comportements sont possibles : quand $\gamma < 0$, la distribution de X est bornée et on dit que l'on est dans le domaine de Weibull; quand $\gamma = 0$, la distribution de X présente une décroissance de type exponentiel dans la queue de distribution, on dit que l'on est dans le domaine de Gumbel; et enfin, le domaine de Fréchet, correspondant à une distribution de X non bornée et une décroissance de type polynômial, est le domaine $\gamma > 0$.

Dans tous les cas cette approche basée sur les distributions GEV est appropriée si les données consistent en des maximas annuels ou périodiques. Cependant elle a été fortement critiquée dans le sens où l'utilisation d'un seul maxima conduit à une perte d'information contenue dans les autres grandes valeurs de l'échantillon. Le problème a donc été résolu en considérant plusieurs grandes valeurs au lieu de la plus grande : i.e. en considérant toutes les valeurs au-delà d'un seuil donné. Les différences entre ces valeurs et le seuil donné s'appellent les *excès au-delà d'un seuil*. Une discussion sur la comparaison entre cette méthode et celle basée sur les distributions GEV a été

proposée dans Rasmussen *et al.* (1994). On suppose typiquement que ces excès ont une loi de Pareto généralisée, notée GPD (γ, σ), dont la distribution est donnée par :

$$G_{\gamma,\,\sigma}(x) = \begin{cases} 1-\left(1+\frac{\gamma}{\sigma}x\right)^{-1/\gamma} & \text{si } \gamma \neq 0,\ \sigma > 0, \\ \\ 1-\exp(-x/\sigma) & \text{si } \gamma = 0,\ \sigma > 0, \end{cases} \tag{2}$$

où $x \in [0, \infty[$ si $\gamma \geqslant 0$ et $x \in [0, -\sigma/\gamma[$ si $\gamma < 0$.

Disposant d'un ensemble d'observations $(X_1, ..., X_n)$, l'analyste cherche à appréhender le type de comportement des données extrêmes. Dans ce but, il doit répondre à la question suivante :

dans quel domaine d'attraction est-il raisonnable de se placer ?

Pour cela, il aura recours à des outils statistiques d'estimation, mais aussi à des outils d'exploration afin d'avoir une visualisation du comportement asymptotique de la loi. Nous allons succinctement, dans la Section 2, présenter ces représentations graphiques ainsi que les principales techniques d'estimation de l'indice γ. Ces estimateurs étant biaisés, nous exhiberons une technique de réduction de biais.

Ceci dit quand on veut mettre en oeuvre ces techniques sur des données réelles, on est souvent confronté à un certain nombre de difficultées. En particulier, une hypothèse sous-jacente à cette théorie, ainsi qu'à beaucoup d'autres théories statistiques, est l'indépendance des observations. Il est clair, en pratique, que cette hypothèse est souvent irréaliste, en particulier dans le domaine de l'hydrologie qui sera le domaine d'application que l'on considèrera dans cet article. Nous présenterons donc, dans la Section 3, une méthodologie, couramment utilisée dans les applications hydrologiques, qui permet d'appréhender ce type de problème.

Par ailleurs, il est nécessaire de supposer que les observations proviennent d'une seule et même distribution. Cette hypothèse est satisfaite à condition que l'on puisse garantir une influence faible des inondations et des modifications artificielles telles que les structures régularisant le courant, les réservoirs, les digues, ... Dans le cas contraire, l'analyse des valeurs extrêmes devra être effectuée sur la base de données censurées, en se basant uniquement sur les observations non influencées par ces divers phénomènes physiques. Les distributions des points « non inondés » et des points « inondés » seront alors toutes deux nécessaires.

Toutes les techniques présentées et développées dans cet article seront appliquées en Section 3 à des données hydrologiques que nous aurons au préalable décrites au début de cette même section.

2. Estimation d'indices extrêmes

2.1. Représentations graphiques

(a) Le «Pareto quantile plot». Le domaine de Fréchet ($\gamma > 0$) a été le plus largement étudié dans la littérature dans la mesure où il englobe un grand nombre d'applications pratiques. Dans ce domaine, les distributions F ont la propriété suivante

$$1 - F(x) = x^{-\frac{1}{\gamma}}\, \ell_F(x), \tag{3}$$

avec ℓ_F une fonction à variations lentes à l'infini. Elle satisfait donc la convergence suivante

$$\forall\, \lambda > 0, \quad \frac{\ell_F(\lambda x)}{\ell_F(x)} \to 1, \quad \text{quand} \quad x \to \infty. \tag{4}$$

En pratique, il est souvent plus commode, non pas de travailler sur la fonction F elle-même, mais sur la fonction queue définie par

$$U(x) = \inf\Big\{ y : F(y) \geqslant 1 - \frac{1}{x} \Big\}.$$

Dans ce cas, supposer (3) est équivalent à supposer que

$$U(x) = x^{\gamma}\, \ell_U(x), \tag{5}$$

avec ℓ_U également une fonction à variations lentes à l'infini. Notons que cette fonction queue est directement liée à la notion de période de retour, qui sera ultérieurement définie.

Si nous considérons la statistique d'ordre $X_{1,n} \leqslant \ldots \leqslant X_{n,n}$ associée à notre échantillon initial, le «Pareto quantile plot», correspondant au graphe de $(\log \frac{n+1}{j}, \log X_{n-j+1,n})$, est une représentation très utile pour visualiser graphiquement si des données sont distribuées selon une loi du domaine de Fréchet ou non. En effet, de (5), il découle que

$$\log U(x) = \gamma \log x + \log \ell_U(x) = \gamma \log x \Big(1 + \frac{\log \ell_U(x)}{\gamma \log x} \Big). \tag{6}$$

En utilisant les propriétés des fonctions à variations lentes, il est immédiat que $\frac{\log \ell_U(x)}{\log x} \to 0$ $(x \to \infty)$ ce qui implique que

$$\log U(x) \sim \gamma \log x \quad (x \to \infty). \tag{7}$$

En remplaçant la fonction queue par sa version empirique $\hat{U}_n$ et en remarquant que $\hat{U}_n\Big(\frac{n+1}{j}\Big) = X_{n-j+1,n}$, nous obtenons finalement l'équivalence suivante :

$$\log X_{n-j+1,n} \sim \gamma \log\Big(\frac{n+1}{j}\Big) \quad \text{quand} \quad \Big(\frac{n+1}{j}\Big) \to \infty.$$

En d'autres termes, le «Pareto quantile plot» sera approximativement linéaire, avec une pente γ, pour les petites valeurs de j, c'est-à-dire les points extrêmes.

(b) «*L'exponential quantile plot*» est similaire au graphe précédent mais concerne cette fois-ci le domaine de Gumbel ($\gamma = 0$). Il consiste tout simplement à remplacer $\log x$ par x en ordonnée du graphe précédent. Dans ce cas, la pente asymptotique dans «l'exponential quantile plot» est égale au paramètre σ.

Nous ne discuterons pas ici de représentations graphiques dans le domaine de Weibull ($\gamma < 0$) puisqu'on garde en mémoire que l'objectif principal de cet article est l'application de la théorie des valeurs extrêmes en hydrologie, et que, comme nous l'expliquerons ultérieurement, dans ce type d'applications ce domaine est très peu fréquent, les variables hydrologiques, telles que les précipitations ou les écoulements, n'étant pas bornées.

Une approche permettant d'éviter le choix a priori du domaine d'attraction a été proposée par Beirlant *et al.* (1996). Elle consiste à utiliser un «quantile plot» généralisé, défini comme le graphe $(\log \frac{n+1}{j}, \log UH_{j,n})$ avec $UH_{j,n}$ de la forme

$$UH_{j,n} = X_{n-j,n}\Big(j^{-1}\sum_{i=1}^{j} \log X_{n-i+1,n} - \log X_{n-j,n}\Big).$$

Suivant la courbure de ce graphe, on peut déduire dans quel domaine d'attraction on se situe : si pour les points extrêmes on voit apparaître une droite de pente positive, on est alors dans le domaine de Fréchet, si par contre on est plutôt constant, on est alors dans le domaine de Gumbel; le cas d'une décroissance linéaire signifie que l'on appartient au domaine de Weibull.

2.2. Estimateurs classiques de γ

Il existe beaucoup d'estimateurs de l'indice proposés dans la littérature. Les plus utilisés en hydrologie sont sans aucun doute les estimateurs des moments, du maximum de vraisemblance (*cf.* e.g. Smith, 1987) et des moments pondérés (*cf.* e.g. Hosking et Wallis, 1987; Rasmussen, 2001). On peut citer par exemple comme référence classique sur l'analyse des valeurs extrêmes les ouvrages de Coles (2001) et Embrechts *et al.* (1997) qui font le point sur les différentes techniques existantes. Grâce aux représentations graphiques introduites dans la section précédente, nous allons présenter d'autres estimateurs que nous étudierons plus en détail ci-après.

(a) Dans le domaine de Fréchet. Comme nous venons de le signaler, le comportement linéaire dans les points extrêmes a lieu avec une pente γ. Autrement dit, on peut facilement construire des estimateurs de l'indice à partir de ce graphe. Cette linéarité apparaît au-delà d'un point $(\log \frac{n+1}{k}, \log X_{n-k+1,n})$. Deux approches sont donc possibles pour la construction de tels estimateurs : soit en forçant la droite à passer par ce point, ce que l'on appelera par la suite «avec contrainte»; soit simplement par moindres carrés, donc «sans contrainte».

Dans le cas «avec contrainte», Csörgő *et al.* (1985) ont proposé les estimateurs à noyau $K_{k,n}$ définis de la façon suivante :

$$K_{k,n} = \frac{\sum_{j=1}^{k} \frac{j}{k} K\left(\frac{j}{k}\right) (\log X_{n-j+1,n} - \log X_{n-j,n})}{\sum_{j=1}^{k} \frac{1}{k} K\left(\frac{j}{k}\right)}$$

où K représente un noyau d'intégrale égale à 1.

Suivant le choix de ce noyau, différents estimateurs peuvent en résulter, le plus connu étant l'estimateur de Hill (1975), correspondant au cas particulier $K(x) = \mathbb{1}_{(0,1]}(x)$, qui peut donc se réécrire simplement comme

$$H_{k,n} = \frac{1}{k} \sum_{j=1}^{k} j\Big(\log X_{n-j+1,n} - \log X_{n-j,n} \Big) = \frac{1}{k} \sum_{j=1}^{k} \log X_{n-j+1,n} - \log X_{n-k,n}.$$

Dans le cas «sans contrainte», en utilisant une approche par moindres carrés, Schultze et Steinebach (1996) et indépendamment Kratz et Resnick (1996) ont introduit l'estimateur Zipf définit par :

$$Z_{k,n} = \frac{\sum_{j=1}^{k} \log \frac{k+1}{j} \log X_{n-j+1,n} - \frac{1}{k} \sum_{j=1}^{k} \log \frac{k+1}{j} \sum_{j=1}^{k} \log X_{n-j+1,n}}{\sum_{j=1}^{k} \log^2 \frac{k+1}{j} - \frac{1}{k} \Big(\sum_{j=1}^{k} \log \frac{k+1}{j} \Big)^2}.$$

Pour établir la consistance (en probabilité ou presque sûrement) de ces estimateurs, deux types d'hypothèses sont nécessaires : l'une portant sur le paramètre k, dont nous discuterons ultérieurement le choix pratique, et l'autre sur la fonction à variations lentes ℓ_U du modèle (5). Concernant ce dernier point, il est en effet nécessaire de faire une hypothèse du second ordre, qui spécifie la vitesse de convergence vers 1 dans (4) :

HYPOTHÈSE $(R_{\ell_U}(b, \rho))$. – *Il existe une constante réelle* $\rho < 0$ *et une fonction* b *vérifiant* $b(x) \to 0$ *quand* $x \to \infty$*, telles que pour tout* $\lambda \geqslant 1$*,*

$$\log \frac{\ell_U(\lambda x)}{\ell_U(x)} \sim b(x) \, \frac{\lambda^\rho - 1}{\rho}, \quad \textit{quand} \quad x \to \infty. \tag{8}$$

Signalons que cette condition est satisfaite pour la plupart des distributions du domaine de Fréchet et que plus la convergence vers 1 dans (4) est rapide, plus la linéarité dans le «Pareto quantile plot» apparaît vite. En d'autres termes, plus le paramètre ρ de l'hypothèse (8) est proche de 0, et plus l'estimation de γ est difficile.

Cette hypothèse a été spécifiée un peu plus par Hall (1982) qui a introduit une classe, appelée la *classe de Hall*, de la forme :

$$\ell_U(x) = M_1\Big(1 + M_2 x^{\rho}(1 + o(1))\Big), \tag{9}$$

où $M_1, M_2 > 0$ et ρ est défini comme dans (8).

Si on cherche à obtenir non plus une consistance, mais une normalité asymptotique, une condition additionnelle est nécessaire pour neutraliser le biais. Elle fait intervenir la fonction $b(.)$ de (8). Mais avant de donner son expression explicite, remarquons qu'un paramètre important en pratique est le choix du nombre k de statistiques d'ordre à utiliser. Ce problème a été longuement abordé dans la littérature (*cf.* Beirlant *et al.*, 1996; Drees et Kaufmann, 1998; ...). L'issue est importante : l'extrême volatilité du graphe $\{(k, \hat{\gamma}_{k,n}) : 1 \leqslant k \leqslant n\}$, où $\hat{\gamma}_{k,n}$ désigne n'importe quel estimateur introduit précédemment, rend difficile l'utilisation de l'estimateur en pratique si aucune indication sur le choix de k n'est donnée. Une minimisation de l'erreur en moyenne quadratique est en général souvent donnée comme critère, puisque, à cause de la fonction à variations lentes dans le modèle (5), le biais est petit pour k petit alors que la variance décroît quand k augmente.

(b) Généralisations aux autres domaines. Les estimateurs précédemment cités concernent tous le domaine de Fréchet, aussi la question légitime maintenant est de savoir ce qui se passe dans les autres domaines. Si on ne fait aucune hypothèse sur le domaine, on a vu que l'on pouvait utiliser le «quantile plot» généralisé. En utilisant ce graphe et les mêmes méthodes de régression que précédemment, nous pouvons obtenir des estimateurs similaires à ceux mentionnés, en remplaçant tout simplement $X_{n-j+1,n}$ par $UH_{j,n}$. Une étude comparative de ces différents estimateurs a été effectuée dans Beirlant, Dierckx *et al.* (2005). L'intérêt de ces techniques réside dans le fait qu'elles ne reposent pas sur l'appartenance ou non à une classe spécifique. Les techniques proposées jusqu'alors dans la littérature hydrologique reposent toutes sur cette hypothèse ou alors sur un modèle particulier (*cf.* e.g. Durrans et Tomic, 2001).

Signalons que tous ces estimateurs de γ peuvent aussi être utilisés comme estimateurs de σ quand on utilise «l'exponential quantile plot» défini dans la Section 2.1.

2.3. Estimation de quantiles ou de périodes de retour

Dans la Section 2.2, nous avons vu qu'un estimateur de γ (resp. de σ) pouvait être obtenu en estimant la pente du Pareto (resp. de «l'exponential») «quantile plot». Ceci dit l'estimation de ce paramètre n'est souvent qu'un objectif intermédiaire, l'objectif réel étant plutôt l'estimation d'un quantile extrême ou d'une queue de probabilité, suivant le type d'application considérée. En hydrologie par exemple, on s'intéresse particulièrement à l'estimation d'une période de retour T associée à une probabilité p $(= 1/10, 1/100, ...)$ représentant la probabilité d'excès au-delà d'un niveau de retour x_p. Cette période s'interprête comme une moyenne du temps ou du nombre d'années séparant un événement de grandeur donnée (x_p) d'un second événement d'une grandeur égale ou supérieure. À titre d'exemple, on peut ainsi définir :

– la crue dite décennale comme la valeur du débit dépassé en moyenne une fois tous les dix ans,
– la crue dite centenaire comme la valeur du débit dépassé en moyenne une fois tous les cent ans.

D'un point de vue mathématique, suivant qu'on utilise l'approche GPD ou GEV, deux définitions de la période de retour sont possibles :

- si on utilise l'approche GPD : on utilise le fait que la loi des excès au-delà d'un seuil u peut être uniformément approchée par une loi GPD (γ, σ). En prenant pour seuil $X_{n-k+1,n}$, on a

$$T = \frac{n}{k+1} \frac{1}{1 - G_{\gamma,\sigma}(x_p)}$$

- si on utilise l'approche GEV : on a directement

$$T = \frac{1}{1 - H_\gamma(x_p)}.$$

Dans le but d'estimer cette période de retour, si on utilise l'approche GPD, on peut à nouveau utiliser les représentations graphiques. En effet, comme nous l'avons expliqué précédemment, les «quantiles plots» sont approximativement linéaires de pente γ au-delà du seuil $X_{n-k+1,n}$. L'équation de cette droite passant par le point $(\log \frac{n}{k+1}, \log X_{n-k+1,n})$ de pente γ est donnée par

$$y = \log X_{n-k+1,n} + \gamma \left(x - \log \left(\frac{n}{k+1} \right) \right).$$

Nous avons donc la relation suivante entre la période de retour T et le débit X :

- pour $\gamma > 0$:

$$\log(X) = \log X_{n-k+1,n} + \hat{\gamma} \left(\log(T) - \log \left(\frac{n}{k+1} \right) \right), \qquad (10)$$

- pour $\gamma = 0$:

$$X = X_{n-k+1,n} + \hat{\sigma} \left(\log(T) - \log \left(\frac{n}{k+1} \right) \right). \qquad (11)$$

Les estimateurs de γ et de σ de la section précédente peuvent être utilisés.

Dans le cas de l'approche GEV, on peut estimer l'indice des valeurs extrêmes γ par l'un des estimateurs de la section précédente et obtenir ainsi un estimateur de la période de retour.

2.4. Correction de biais

Comme nous l'avons déjà vu dans la Section 2.1, la pente dans les «quantiles plots» est approximée par les différents estimateurs à un facteur de nuisance près : la fonction à variations lentes. En effet, l'approximation (7) est obtenue en négligeant la fonction à variations lentes.

Il est donc légitime de penser que si nous arrivons à estimer l'effet de cette fonction à variations lentes, nous pourrons réduire le biais de nos estimateurs. Pour cela nous allons nous placer sous le modèle de Hall défini dans (9), avec $M_1 = 1$ de façon à garantir un comportement de loi extrême dans la queue et donc pas de biais dans cette partie extrême ($\xi(p) = \log \ell_U(1/p) \to 0$ quand $p \to 0$). Nous avons alors l'approximation suivante :

$$\xi(p) \simeq M_2\, p^{-\rho} \quad (\text{quand} \quad p \to 0). \tag{12}$$

La pente asymptotique γ peut aussi être calibrée en utilisant les techniques de régression présentées dans la Section 2.2 après translation des observations dans le «quantile plot» (*cf.* (6)) :

$$\log U\Big(\frac{1}{p}\Big) - \log \ell_U\Big(\frac{1}{p}\Big) = -\gamma \log p. \tag{13}$$

Pour chaque ensemble de paramètres sélectionnés dans le modèle $\xi(p)$, la translation peut être faite et la pente asymptotique déterminée en se basant sur une régression des observations translatées. Le bien fondé de l'ensemble des paramètres sélectionnés peut être testé en traçant le graphe des différences entre la droite de régression et les valeurs observées de $\log U\Big(\frac{1}{p}\Big)$. Avec le modèle (12), on devrait être proche d'une régression linéaire en traçant $\log(\xi(p))$ en fonction de $-\log p$:

$$\log \xi(p) = \log M_2 - \rho \log(p). \tag{14}$$

Cela signifie que le paramètre ρ de la fonction à variations lentes peut être obtenu par régression dans le graphe de $\log(\xi(p))$ en fonction de $-\log(p)$, après avoir sélectionné les valeurs des paramètres M_2 et γ. À nouveau les différentes méthodes de régression peuvent être utilisées. Différents estimateurs du paramètre ρ, du même type que ceux de γ, peuvent être ainsi obtenus. En général, deux paramètres doivent être optimisés tandis que le 3ème peut être calculé (soit γ en se basant sur (13), soit ρ en utilisant (14)). L'optimisation peut être faite en minimisant l'erreur en moyenne quadratique (MSE) de la régression (soit pour γ ou pour ρ).

Après calibration de la fonction à variations lentes, une correction du biais peut être appliquée à l'estimation de quantile :

- pour $\gamma > 0$:

$$\log(X) = \log X_{n-k+1,n} + \hat{\gamma}\left(\log(T) - \log\left(\frac{n}{k+1}\right)\right) - \log\left(\hat{\ell}_U\left(\frac{1}{p}\right)\right), \tag{15}$$

- pour $\gamma = 0$:

$$X = X_{n-k+1,n} + \hat{\sigma}\left(\log(T) - \log\left(\frac{n}{k+1}\right)\right) - \hat{\ell}_U\left(\frac{1}{p}\right). \qquad (16)$$

L'estimateur de γ à biais corrigé doit être utilisé comme $\hat{\gamma}$ ou $\hat{\sigma}$ dans (15) et (16). L'estimateur $\hat{\ell}_U\left(\frac{1}{p}\right)$ a été obtenu en utilisant le modèle sur la fonction à variations lentes avec M_2 et ρ remplacés par leurs valeurs calibrées.

2.5. Données censurées

Le problème de l'analyse des valeurs extrêmes en présence de censure a été jusqu'à présent quasiment ignoré dans la littérature. Il faut savoir que, de façon générale, ce problème est très complexe, tous les outils de la statistique classique devant être modifiés. Le contexte de données censurées à droite, qui est typiquement celui que l'on peut utiliser dans les applications hydrologiques, est le suivant : on suppose disposer de deux échantillons

$$X_1, ..., X_n \sim^{i.i.d.} F$$
$$Y_1, ..., Y_n \sim^{i.i.d.} G$$

où F et G sont inconnues et les deux échantillons supposés indépendants. Dire que l'on est en présence de données censurées, revient à dire que l'on n'observe pas les X_i, mais seulement les couples $(Z_i, \delta_i)_{i=1,...,n}$, où

$$Z_i = \min(X_i, Y_i) \quad \text{et} \quad \delta_i = \mathbb{1}_{X_i \leqslant Y_i},$$

δ_i représentant l'indicateur de censure.

Dans le cas de données non censurées, l'estimateur usuel de F que l'on utilise pour construire nos estimateurs est la fonction de répartition empirique, qui associe la masse $1/n$ à chacune des observations. Dans le cas censuré, cet estimateur est remplacé par l'estimateur de Kaplan Meier (1958) défini comme

$$1 - \hat{F}_n(x) = \prod_{j=1}^{n}\left[1 - \frac{\delta_{j,n}\mathbb{1}_{Z_{j,n} \leqslant x}}{n-j+1}\right],$$

où $Z_{j,n}$ représente la statistique d'ordre associée à $Z_1, ..., Z_n$ et $\delta_{j,n} = \delta_k$ si et seulement si $Z_{j,n} = Z_k$. Bien que complexe à première vue, cet estimateur fonctionne de la façon suivante : il n'affecte aucun poids à une donnée censurée, mais la masse $1/n$ qu'il aurait dû lui affecter si elle était observée est dispatchée équitablement entre les autres plus grandes valeurs non censurées de l'échantillon.

La difficulté bien sûr est que, bien que l'on n'observe pas l'échantillon $X_1, ..., X_n$, on cherche à estimer l'indice γ_1 de F. Signalons de plus que les estimateurs usuels (définis en Section 2.2) de l'indice ne sont plus consistants dans

ce type de contexte. Dans le cas où F et G ont des indices $(\gamma_1, \gamma_2) \in \mathbb{R}_+ \times \mathbb{R}_+^*$ ou $\mathbb{R}_-^* \times \mathbb{R}_-^*$, des estimateurs de γ_1 ont été tout récemment proposés par Beirlant, Guillou *et al.* (2005).

Toutefois, l'utilisation de ces estimateurs adaptés à la censure est remise en cause du fait que, dans les applications hydrologiques, nous n'avons pas d'indépendance entre les deux échantillons X_i et Y_i. Il serait donc utile dans l'avenir de développer des techniques dans ce type de contexte. Une méthode possible pour aborder ce problème consiste à utiliser l'approche POT («Peaks-Over-Thresholds») qui repose sur les résultats de Balkema et de Haan (1974) ainsi que ceux de Pickands (1975) selon lesquels les (N_t) excès absolus $E_j = Z_j - t > 0$ au-delà d'un seuil t peuvent être approchés par une loi de Pareto généralisée. Reste alors à adapter la vraisemblance en tenant compte de la censure

$$\prod_{j=1}^{N_t} \Big[g_{\gamma,\sigma}(E_j) \Big]^{\delta_j} \Big[1 - G_{\gamma,\sigma}(E_j) \Big]^{1-\delta_j},$$

où $g_{\gamma,\sigma}$ est la densité associée à la GPD (γ, σ) définie dans (2).

Si on arrive à modéliser la dépendance entre X et Y par exemple par un modèle de régression pour le (ou les) paramètre(s) γ et σ (*cf.* l'exemple dans Beirlant, Guillou *et al.*, 2005), il suffit alors de maximiser la vraisemblance. Ceci dit cette modélisation est en dehors du cadre de cet article, mais fera l'objet d'un travail ultérieur.

Nous avons donc, dans ce papier, utilisé non pas des techniques de censure au sens classique du terme, mais des techniques de données tronquées. Ce problème avait déjà été abordé dans un cadre un peu différent pour des applications actuarielles par Beirlant et Guillou (2001). Quand on veut tronquer une distribution (par exemple pour calculer la distribution des écoulements en se basant sur les débits non inondés), les estimateurs de la pente dans le «quantile plot» définis dans la Section 2.2 doivent être modifiés. Par exemple si ℓ représente le rang correspondant au niveau de troncature, la pente peut être calculée sur la base de l'estimateur modifié suivant

$$\hat{\gamma}_{\ell,k} := \frac{\sum_{j=\ell+1}^{k} w_j \log \frac{k+1}{j} \Big(\log X_{n-j+1,n} - \log X_{n-k,n} \Big)}{\sum_{j=\ell+1}^{k} w_j \log^2 \frac{k+1}{j}},$$

où les w_j sont des facteurs poids convenables.

Nous nous proposons d'appliquer ces différentes techniques d'estimation à des données hydrologiques que nous allons dans un premier temps décrire dans la section suivante.

3. Présentation des données hydrologiques

L'analyse des valeurs extrêmes a beaucoup d'applications en hydrologie (*cf.* e.g. Rosbjerg et Madsen (2004) pour une récente vision d'ensemble). Elle est utilisée pour évaluer les risques d'inondation, pour l'élaboration de cartes d'inondations, ...

Comme nous allons l'expliquer dans la section suivante, l'analyse est le plus souvent basée sur les débits de rivière.

3.1. Données brutes

En règle générale, les chroniques enregistrées sont souvent des chroniques de hauteurs d'eau qui sont ensuite converties en débit grâce à ce que l'on appelle la courbe de tarage. La courbe de tarage donne, pour une section d'une rivière, la relation entre la hauteur du niveau d'eau relevé sur une échelle limnigraphique et le débit. Le débit est un volume par unité de temps. Cependant, dans le cas d'une rivière, il n'est bien sûr pas possible de mesurer directement les volumes écoulés. Par contre, on peut mesurer la vitesse de l'écoulement en différents points de la section de la rivière. Le débit est alors obtenu en intégrant ces vitesses sur toute la section. Le débit calculé est quasi instantané, puisque les vitesses ont été calculées sur de courts instants, durant lesquels les grandeurs mesurées sont supposées constantes. La courbe de tarage est alors obtenue en regroupant sur un même graphe les mesures de débit instantané effectuées pour différentes hauteurs.

Les gestionnaires de chaque station de jaugeage doivent définir la courbe de tarage avec le plus grand soin, afin de limiter au maximum les incertitudes de mesure. En effet, plusieurs problèmes se présentent, en particulier celui de l'extrapolation : lors de très fortes crues, le plus souvent il n'est pas possible d'effectuer une mesure de débit correspondant à la hauteur relevée. Le calcul du débit est alors effectué en extrapolant la courbe de tarage, ce qui a pour conséquence une augmentation des incertitudes. Nous reviendrons sur ce point ultérieurement.

Les différentes méthodes d'estimation présentées en Section 2 seront appliquées à des débits horaires concernant la rivière Molenbeek, située dans le bassin de la rivière Dender en Flandre (Belgique). Des observations de débits limnigraphiques sont disponibles dans la station Erpe-Mere qui a une superficie de 47 km^2. La Figure 1 illustre l'influence des inondations sur la courbe de tarage pour la rivière Molenbeek à Erpe-Mere avant 1996. Cette influence a été étudiée dans Willems *et al.* (2002) et Rombauts et Willems (2003) au moyen d'un modèle hydraulique pour la rivière et les zones inondables avoisinantes. Il est clair que les inondations affectent la courbe de tarage qui « tombe » à cause de l'accumulation de l'eau dans la zone inondable. Très souvent, seulement un nombre limité d'observations est disponible au-delà du seuil d'inondation. Dans certains cas même, aucune observation n'a été faite pendant cette période. L'extrapolation est alors d'autant plus imprécise et peut conduire à un indice des valeurs extrêmes (à tort) négatif (pour plus de détail, nous référons à Willems, 2004).

Dans les études de modélisation des inondations, une description précise des débits de la rivière et des périodes de retour est indispensable. Le calcul d'une période de retour peut être basé sur une série temporelle des débits de rivière ou des mesures du niveau d'eau. Il peut également être basé sur des résultats de simulations découlant d'un modèle hydrologique et/ou hydraulique. Les résultats de simulations découlant de modèles mathématiques sont utiles dans le cas où aucune mesure de l'écoulement de la rivière n'est disponible (pour les secteurs non mesurés, pour les endroits plus en amont ou en aval de la station mesurée). Dans tous les cas, une distinction claire doit

être faite entre le débit des rivières (ou niveau d'eau) et l'écoulement naturel après précipitation (appelé ultérieurement l'écoulement). Ce dernier est un écoulement arrivant dans la rivière et peut être considéré en amont d'un endroit donné de la rivière. Le débit de la rivière (en un point donné) est la transformation de l'écoulement (en amont de ce point) après qu'il ait été acheminé à travers le réseau de la rivière. L'acheminement aura un effet d'aplatissement sur les hydrogrammes (graphes du débit en fonction du temps) et peut être fortement influencé par les modifications artificielles telles que les structures régularisant le courant, les réservoirs, les digues, ... Pour les plus forts débits, les inondations auront aussi une importance. La Figure 2 (a) donne un exemple des différences entre le débit des rivières d'une part et l'écoulement d'autre part pour la rivière Molenbeek à Erpe-Mere en Belgique. De cette figure, il est clair que les pics les plus élevés (approximativement $6\,m^3/s$) sont aplatis par l'écoulement dans la rivière. Le phénomène peut physiquement s'expliquer ici de façon conceptuelle par l'accumulation de l'eau dans la zone inondable de la rivière dès que le débit dépasse la capacité du lit de la rivière (en réalité de nombreux phénomènes physiques sont impliqués, mais une explication détaillée de ce phénomène est en dehors du cadre de cet article). L'eau stockée ne contribuera pas au débit de la rivière et pour cette raison le débit de la rivière sera plus petit que l'écoulement. Une illustration schématique de ce concept est donnée dans la Figure 2 (b).

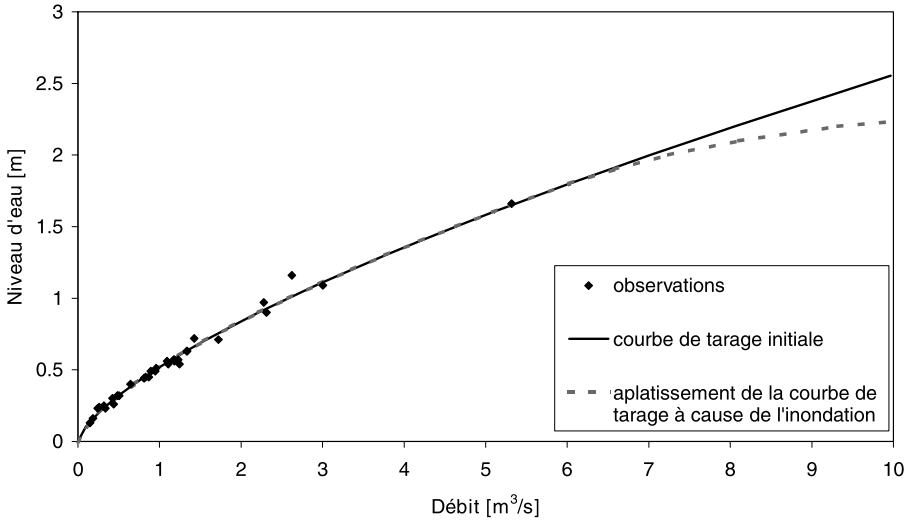
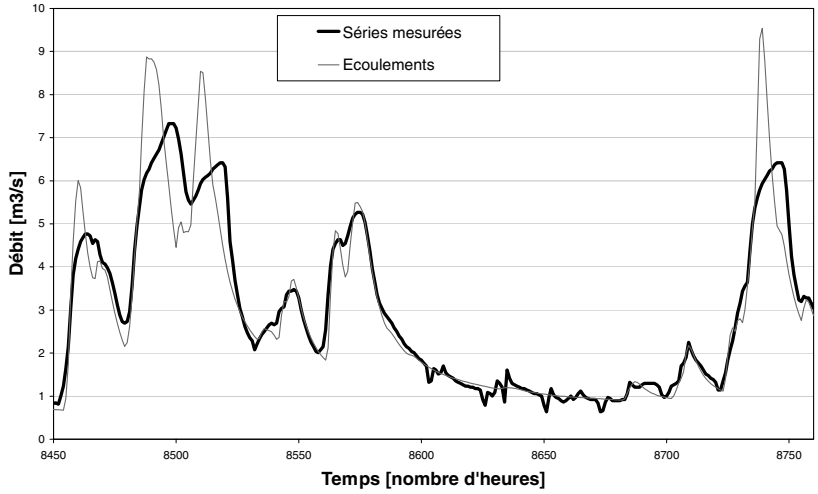


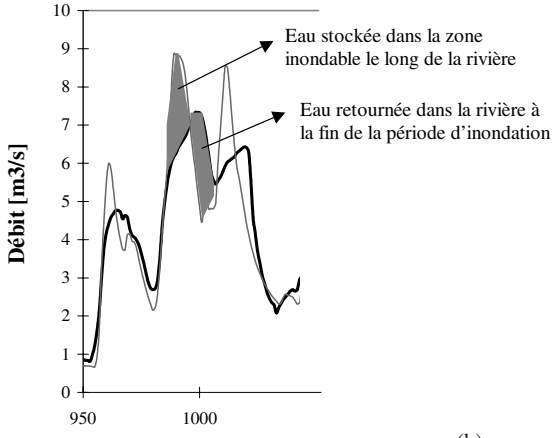
FIGURE 1
Influence des inondations sur la courbe de tarage pour la rivière Molenbeek à Erpe-Mere avant 1996

Dans la mesure où les débits disponibles sont horaires, parfois seulement journaliers, il est clair qu'il est peu réaliste de considérer la série comme formée de variables aléatoires indépendantes. Or si cette hypothèse, sous-jacente à de nombreuses théories statistiques et en particulier à celle relative à la théorie des valeurs extrêmes, n'est pas satisfaite, des résultats aberrants peuvent en résulter. La plupart des travaux théoriques dans l'approche POT (*cf.* e.g. Borgman, 1963; Shane

et Lynn, 1964; Bernier, 1967; Todorovic, 1970) suppose simplement qu'une série de pic est disponible. Bien que cette hypothèse puisse être appropriée dans les études théoriques, les praticiens ont besoin de lignes de conduite pour l'obtention de cette série. Nous nous proposons donc, dans la section suivante, de décrire le critère de sélection que nous utiliserons dans toute la suite.



(a)



(b)

FIGURE 2

(a) Comparaison des débits horaires de la rivière et de l'écoulement pour la période d'inondation de décembre 1993 concernant la rivière Molenbeek à Erpe-Mere (Belgique);
(b) Explication schématique de l'influence de l'inondation sur les hydrogrammes

3.2. Sélection d'observations indépendantes

Des pics d'inondations consécutifs seront considérés comme indépendants si l'intervalle de temps entre ces deux pics dépasse un temps critique et si le débit entre ces deux événements passe en dessous d'un niveau proche du débit de base, qui peut s'interpréter comme un débit « normal » au sens où il s'agit ni d'une période d'inondation, ni d'une période de sécheresse. Différents critères de sélection ont été proposés dans la littérature (*cf.* e.g. USWRC, 1976; Lang *et al.*, 1999; Claps et Laio, 2003). Nous décrivons ci-dessous plus en détail celui que nous allons utiliser et nous illustrons dans la Figure 3 (a), les valeurs indépendantes sélectionnées à partir de la série de débits horaires de Molenbeek à Erpe-Mere.

Plus spécifiquement, deux pics consécutifs seront considérés comme indépendants quand le temps p entre ces deux pics est plus long que le facteur de récession k, et quand le débit minimum entre ces deux pics est plus petit qu'une fraction f du pic du débit. En utilisant ce critère, on peut tester si la période est suffisamment sèche, sans avoir à utiliser des données de précipitation. Un critère supplémentaire est dans ce cas nécessaire pour éviter que des petits pics ne soient sélectionnés : il faut donc supposer que q_{max} soit supérieur à une valeur limite notée q_{lim}. La méthode doit donc vérifier les 3 critères suivants (*cf.* Figure 3 (b) pour la définition des paramètres) :

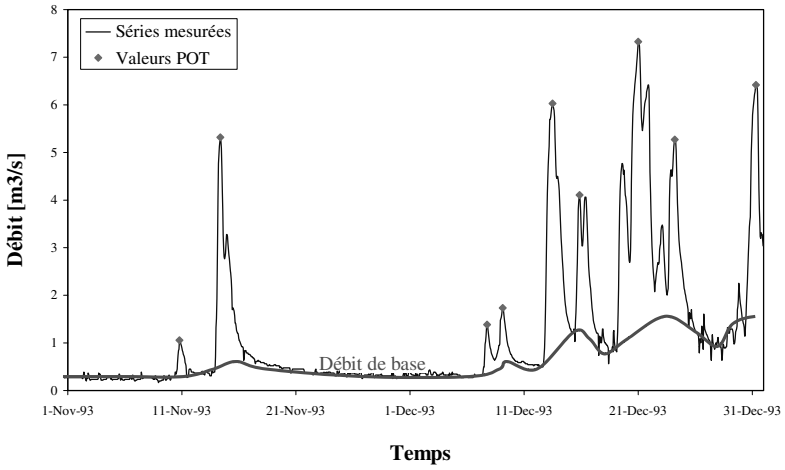
$$p > k, \quad \frac{q_{min}}{q_{max}} < f, \quad q_{max} > q_{lim}.$$

Une question légitime que l'on peut se poser est de savoir quel est l'impact du choix de f. Pour cela, dans la Figure 4, nous représentons les différences entre les valeurs POT sélectionnées pour différentes valeurs de f. Il est clair que pour un critère d'indépendance plus strict (en sélectionnant moins de valeurs POT, ce qui correspond à utiliser une valeur plus petite pour f), la distribution se courbe plus et un seuil plus haut est nécessaire. En conséquence, la calibration de la distribution des valeurs extrêmes devient moins précise. Cela conduit à une difficulté dans le choix du critère d'indépendance : un critère plus strict conduit à des extrêmes plus indépendants (ce qui est préférable pour l'application de la théorie des valeurs extrêmes), mais détériore la précision dans la calibration de la distribution des valeurs extrêmes. De l'analyse de la Figure 4, il est clair aussi que pour un critère d'indépendance plus faible ($f = 0.7$) un comportement asymptotique linéaire de la queue est valide. Cela conduit à la conclusion pour cette station que le désavantage occasionné par un critère d'indépendance moins strict est moins important que l'avantage d'utiliser un tel critère. Puisque des conclusions similaires ont été faites pour d'autres jeux de données hydrologiques (*cf.* e.g. Claps et Laio, 2003; Willems *et al.*, 2005), on peut conclure que pour de telles données, il n'est pas nécessaire de considérer un critère très strict. Dans la plupart des cas, il est préférable de sélectionner un grand nombre d'extrêmes POT de la série et donc de disposer de plus d'extrêmes indépendants.

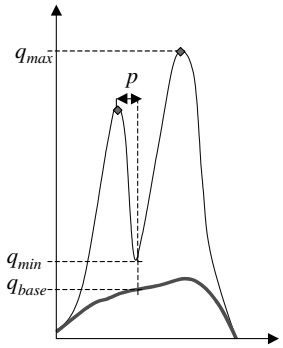
Cette technique d'extraction de variables aléatoires indépendantes a été utilisée, de façon plus ou moins modifiée suivant le type d'applications, dans de nombreux domaines, comme dans les analyses économiques, en data mining, ... Il est clair qu'un tel critère d'indépendance influence le nombre d'extrêmes et aussi l'interprétation de la période de retour d'un événement extrême. Il n'est pas complètement en accord avec

«l'indépendance statistique» qui est sous jacente à la théorie des valeurs extrêmes. Dans la plupart des cas pratiques cependant le critère subjectif aboutira à une grande indépendance, et ceci garantira asymptotiquement l'existence d'une distribution GPD dans la queue de distribution.

La dépendance entre des événements POT consécutifs peut aussi être calculée plus explicitement pour les différents critères de sélection utilisés (par exemple en calculant la dépendance de la série dans les extrêmes sélectionnés). Rosbjerg (1987) a montré comment cette information sur la dépendance de la série pouvait être utilisée pour modifier le modèle GPD présenté en Section 1. On peut citer également l'article de Ferro et Segers (2003) intéressant dans ce contexte.



(a)



(b)

FIGURE 3

(a) Valeurs indépendantes POT sélectionnées à partir de la série de débits horaires de la rivière Molenbeek à Erpe-Mere (Belgique);
(b) Paramètres utilisés dans le critère de sélection des valeurs POT indépendantes

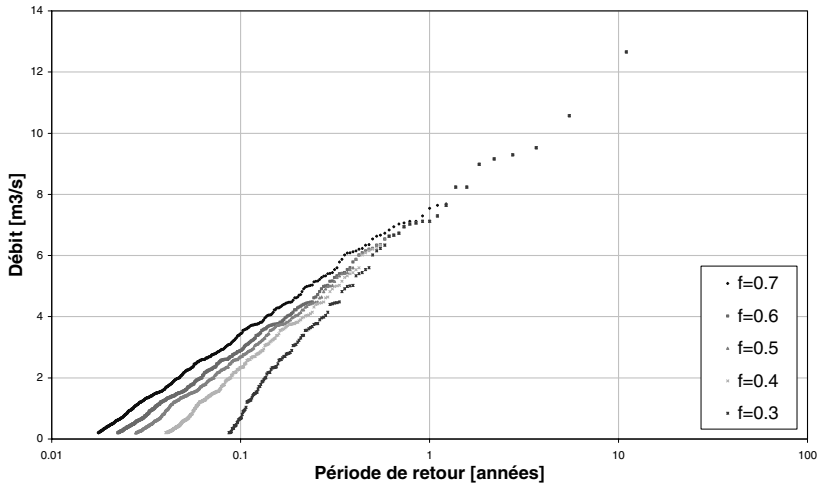


FIGURE 4
Comparaison des extrêmes POT empiriques sélectionnés de la série de débits horaires de la rivière Molenbeek à Erpe-Mere (Belgique) de 1986 à 1996 pour différents niveaux d'indépendance ($f = 0.3, 0.4, ..., 0.7$)

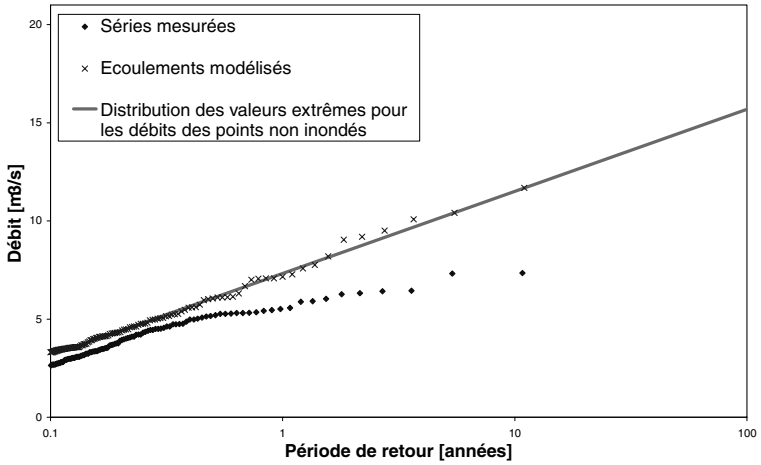
3.3. Applications de la théorie des valeurs extrêmes aux données hydrologiques

Dans les études hydrologiques, les distributions des valeurs extrêmes des débits d'écoulement et des débits de rivière sont toutes deux intéressantes. Alors que la théorie des valeurs extrêmes est souvent valide pour les débits d'écoulement, ce n'est pas souvent le cas pour les débits de rivière. Les influences discontinues et artificielles sur les débits de rivière sont en effet très variables et dépendent fortement du niveau des débits. L'influence des inondations ou des structures hydrauliques n'est par exemple présente que pour les débits les plus élevés. Pour cette raison, les plus petits et les plus forts débits ne sont pas i.i.d. Ce phénomène peut être clairement observé sur la Figure 5 (a).

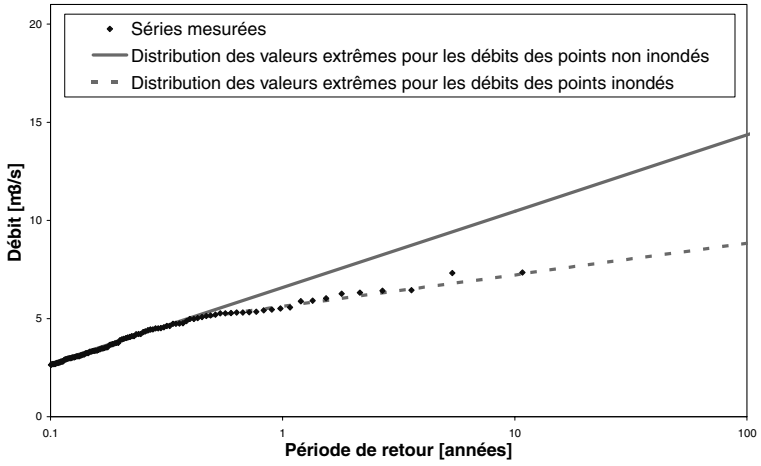
Alors que les débits horaires d'écoulement POT semblent suivre la même loi exponentielle, ce n'est pas le cas des débits de rivière. En amont de la station Erpe-Mere, les inondations commencent à avoir lieu pour un débit supérieur à 5.3 m^3/s et c'est pour cette raison que pour les débits supérieurs à ce seuil, nous voyons une certaine courbure apparaître dans le graphe. De cette analyse il est clair que les débits doivent être décomposés en au moins deux sous-populations. Les débits POT plus petits et plus grands que 5.3 m^3/s suivent une loi exponentielle avec une pente différente dans la Figure 5 (a) et la Figure 5 (b).

Dans la Figure 5 (a), la pente (droite) de la loi exponentielle pour les points non inondés est plus ou moins parallèle à celle pour les débits d'écoulement. Le petit décalage est expliqué par la propagation des débits le long de la rivière. Il est bien connu en hydrologie que cet acheminement va aplatir les hydrogrammes. Pour les points inondés, il n'est pas aussi clair que la distribution exponentielle puisse être utilisée pour les débits les plus élevés. Pour les débits supérieurs à 8 m^3/s (en

dehors des mesures de débits disponibles), des inondations supplémentaires peuvent se produire et peuvent causer soit une décroissance supplémentaire dans la pente de la distribution exponentielle dans le «quantile plot», soit une croissance de la pente due au fait que la zone inondable soit pleine.



(a)



(b)

FIGURE 5

(a) Comparaison de la distribution empirique pour les débits horaires POT et les débits d'écoulement en amont pour la rivière Molenbeek à Erpe-Mere (Belgique) de 1986 à 1996;
(b) Distribution des valeurs extrêmes décomposée en deux sous-populations (points inondés et non inondés) pour les débits horaires de la rivière Molenbeek à Erpe-Mere (Belgique) de 1986 à 1996

Dans tous les cas, la première chose à faire est une distinction entre les domaines $\gamma > 0$, $\gamma = 0$ et $\gamma < 0$. Signalons cependant que le domaine $\gamma < 0$ est très rare en hydrologie puisque cela revient à dire que la distribution a une limite finie. Les données de précipitations montrent le plus souvent une distribution avec un indice nul ou positif strict (*cf.* e.g. Buishand, 1989; Harremoës et Mikkelsen, 1995). Par conséquent, les débits de rivières et les écoulements n'apparaissent pas avec des limites supérieures, sauf s'il y a eu des influences humaines qui les limitent ou des inondations qui réduisent les débits de pointes. En se basant sur notre expérience pour de nombreuses rivières on peut aboutir à la conclusion que le cas $\gamma < 0$ est souvent observé pour les débits de rivière soit à cause de l'influence des inondations, soit à cause d'une mauvaise extrapolation le long de la courbe de tarage. Des techniques d'estimation traditionnelles (en hydrologie en particulier) pour les paramètres de la GPD peuvent être utilisées. On peut citer par exemple les méthodes des moments, du maximum de vraisemblance, des moments pondérés, des L-moments, ... Ces méthodes ont cependant le désavantage de ne pas se concentrer sur la forme de la queue de distribution. Par conséquent, la variance de ces estimateurs peut être élevée, comme cela a été signalé par exemple dans Rosbjerg *et al.* (1992). Des tests ont également été récemment proposés pour identifier le domaine dans lequel on se trouve (*cf.* e.g. Chaouche et Bacro, 2004). Les estimateurs que nous avons proposés en Section 2, valables sans restriction sur le domaine, peuvent être utilisés. Dans le cas de Molenbeek, les valeurs POT indépendantes suivent une distribution des valeurs extrêmes exponentielle (indice des valeurs extrêmes $\gamma = 0$). Ceci est illustré dans la Figure 6 où la pente de «l'exponential quantile plot»est asymptotiquement constante, et pourrait également être corroboré par la pente du «Pareto quantile plot» qui décroit alors continûment vers 0 ou celle du «quantile plot» généralisé qui fluctue autour de 0. La conclusion de tous ces graphes est donc la même : l'indice des valeurs extrêmes pour Molenbeek est nul. Concernant l'estimation du paramètre σ dans «l'exponential quantile plot», les différents estimateurs de la pente peuvent à nouveau être utilisés. La Figure 6 montre les différentes estimations pour des seuils k allant de 1 au nombre maximal de valeurs POT extraites (623). Pour des seuils élevés, le calcul de la pente est seulement basé sur un nombre limité d'observations. De ce fait l'incertitude sur l'estimation est grande (une MSE élevée pour approximativement 50 observations) et l'estimation de la pente présente alors de grandes fluctuations. Pour les niveaux les plus bas, la MSE est plus petite et les estimations de la pente plus stables. C'est la gamme de seuils dans laquelle doit être sélectionné le seuil optimal (de préférence au rang ayant la MSE la plus faible, ici 617).

Nous avons également appliqué la technique de correction de biais à cette rivière pour différents seuils f d'indépendance de façon à illustrer la qualité et la robustesse de ces techniques.

Comme nous l'avons indiqué en Section 2.4, les estimateurs précédents de la pente sont biaisés à cause de la fonction à variation lente. Le biais dépend fortement du niveau f d'indépendance considéré. Différents niveaux d'indépendance conduisent à différentes pentes asymptotiques (*cf.* Figure 7 (a)). L'estimateur de la pente asymptotique est d'autant plus grand que f est petit. Les valeurs POT translatées dans «l'exponential quantile plot» sont exhibées en Figure 7 (b). La correction de biais dans l'estimation de la pente basée sur la translation est illustrée en Figure 8 (a). Il est clair, au vu de cette figure, que l'estimation est beaucoup plus stable après correction du biais. Finalement, les valeurs POT translatées sont représentées dans

la Figure 8 (b) pour différents niveaux d'indépendance. Les différences sont faibles. Après translation, des estimateurs de la pente asymptotique proches sont obtenus pour différents niveaux d'indépendance (*cf.* Figure 7 (a)). Il est clair que sans la correction de biais, des estimateurs plus élevés sont obtenus.

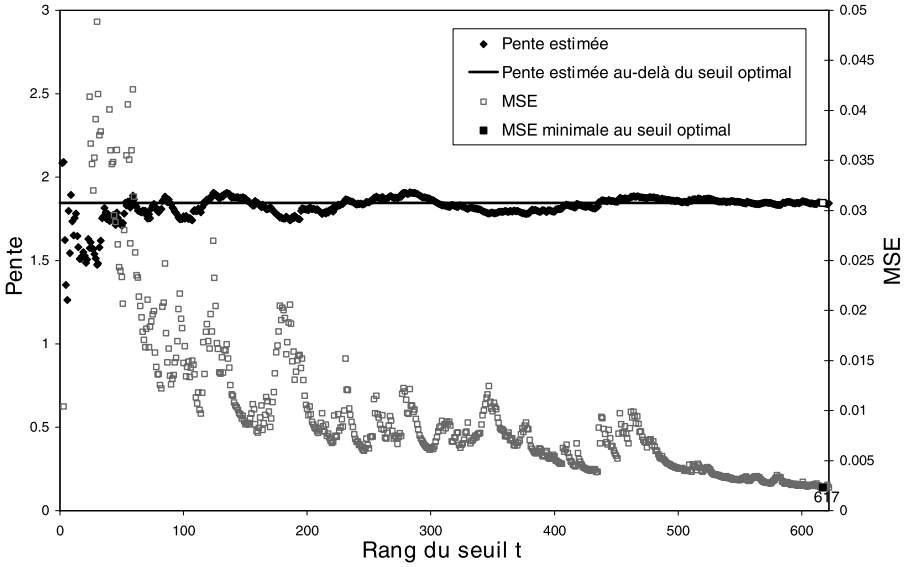
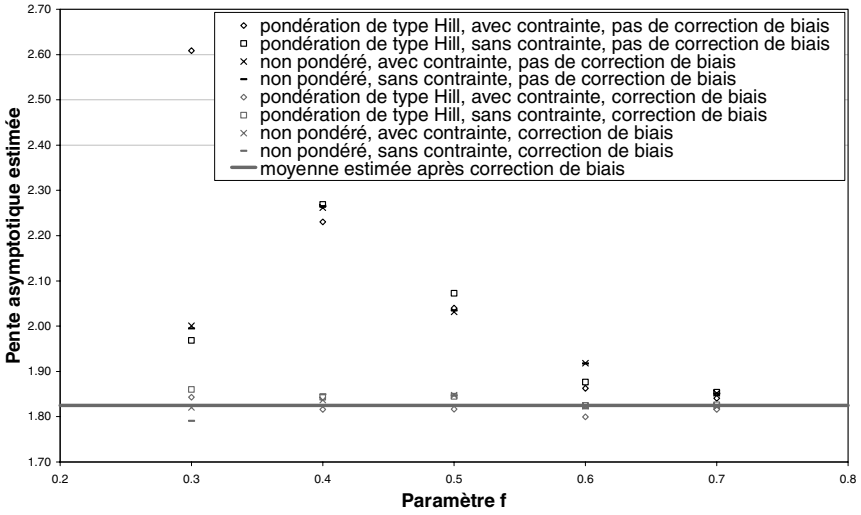


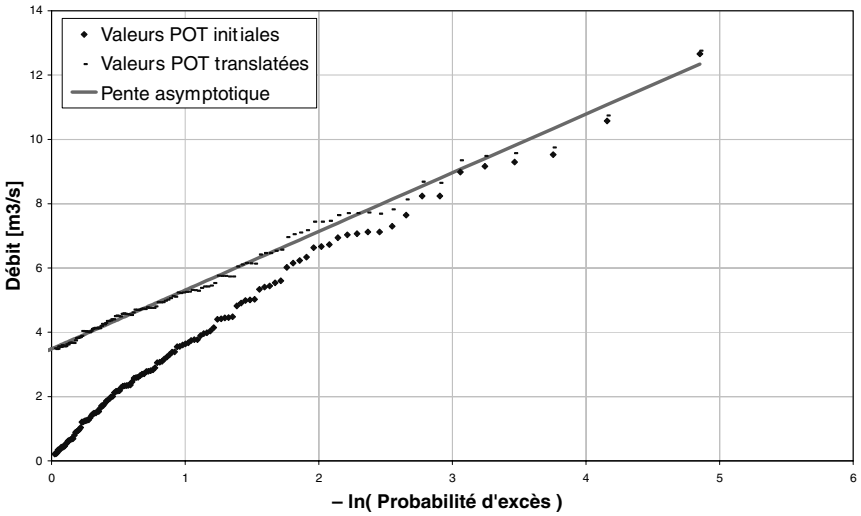
FIGURE 6
Estimation de type Hill de la pente de «l'exponential quantile plot» en fonction du seuil considéré pour les débits horaires de la rivière Molenbeek à Erpe-Mere (Belgique) de 1986 à 1996 pour une valeur de $f = 0.3$

Signalons de plus que la correction de biais peut également être appliquée à l'estimateur UH et conduirait à un résultat similaire : après la correction de biais, la pente asymptotique devient de plus en plus constante dans la queue et l'estimateur UH est, après correction de biais, d'autant plus proche de 0.

De plus, comme nous l'avons indiqué, l'analyse des valeurs extrêmes peut être fortement erronée pour certains débits de rivière à cause de la nature discontinue (par paliers) des influences artificielles. Pour les débits d'écoulement, ce problème est moins important puisque l'hydrologie est beaucoup plus naturelle. Cependant, le modèle statistique de l'écoulement naturel est difficile à construire puisque les mesures des écoulements ne sont pas réellement disponibles. Elles doivent être calculées à partir des débits de rivière non inondée (ce qui est fait sur la base de la Figure 5 (b)) ou à partir de la modélisation des débits (comme cela est fait dans la Figure 5 (a)). Dans le premier cas, l'analyse des valeurs extrêmes doit être faite en utilisant la censure : l'analyse doit considérer les événements les plus hauts («avec influence des inondations»), mais ne doit pas tenir compte des valeurs des débits pour la calibration de la distribution des valeurs extrêmes de l'écoulement naturel.



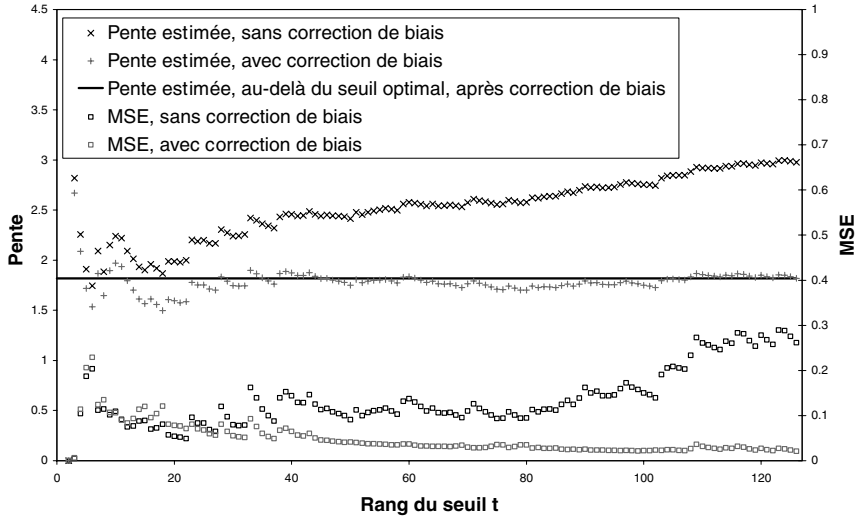
(a)



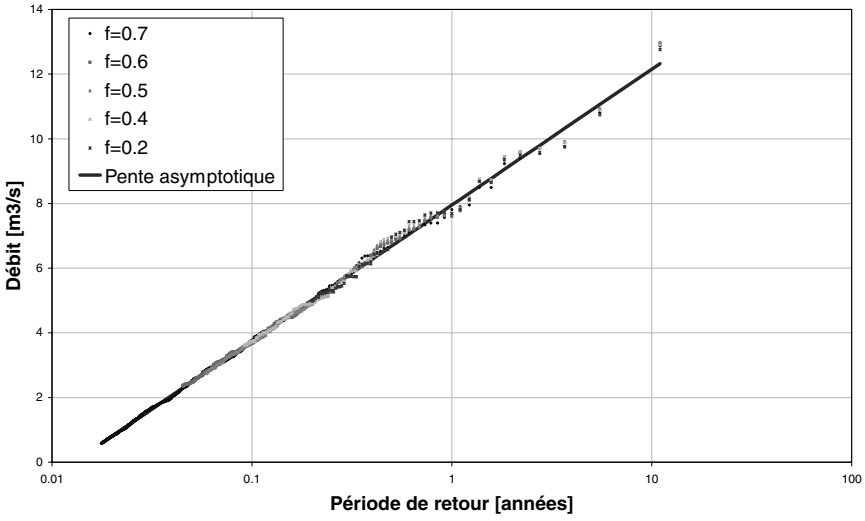
(b)

FIGURE 7

(a) Résultat de la calibration pour la pente asymptotique dans la distribution des valeurs extrêmes exponentielle;
(b) Translation des observations dans «l'exponential quantile plot» par moyenne de la fonction à variations lentes pour $f = 0.3$ *pour déterminer la pente asymptotique*



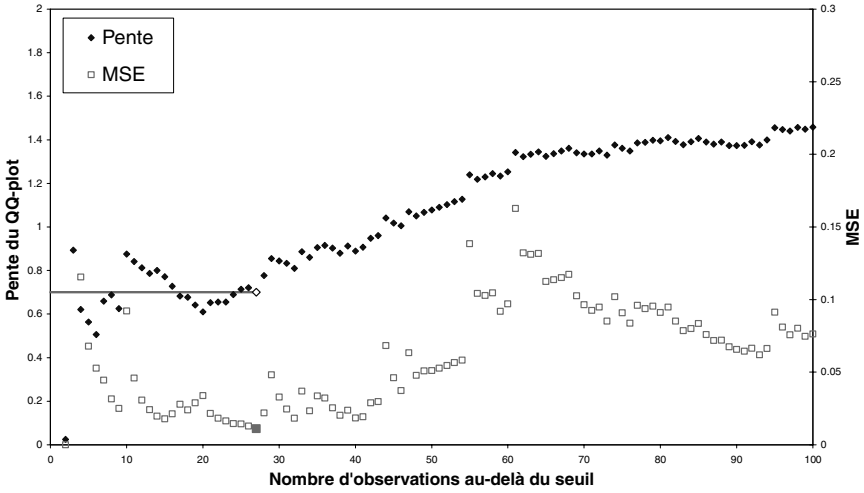
(a)



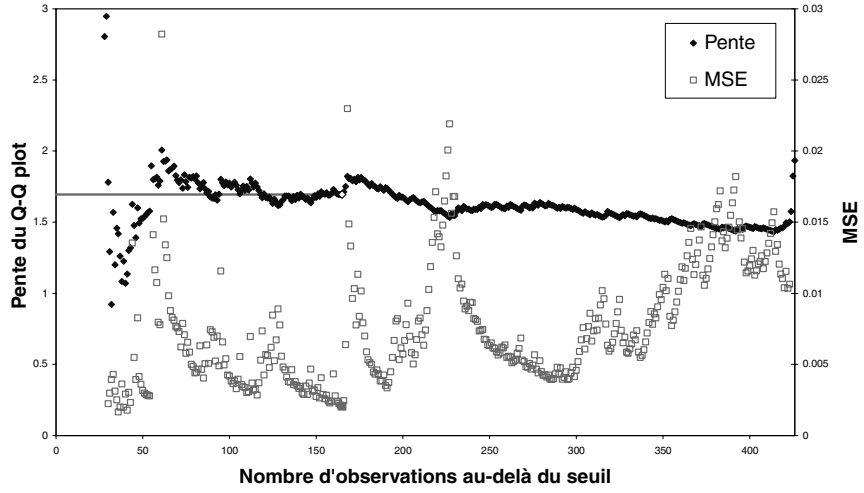
(b)

FIGURE 8

(a) Différence dans l'estimation de la pente asymptotique avec et sans correction de biais basée sur la fonction à variations lentes pour $f = 0.3$*;*
(b) Comparaison des extrêmes POT translatés pour différents niveaux d'indépendance $(f = 0.3, 0.4, ..., 0.7)$



(a)



(b)

FIGURE 9
Estimation de la pente dans «l'exponential quantile plot» (a) avant censure et (b) après censure au seuil $k = 25$ pour les débits horaires de la rivière Molenbeek à Erpe-Mere (Belgique) de 1986 à 1996

Les techniques de troncature présentées en Section 2.5 sont appliquées aux débits de rivière à Molenbeek. Les Figures 9 (a) et 9 (b) montrent les résultats de l'estimation de la pente dans «l'exponential quantile plot» de la Figure 5 (b) (paramètre σ) pour les débits de rivière avant et après censure au seuil d'inondation.

Dans la Figure 9 (a), deux pentes apparaissent clairement pour les seuils plus grands ou plus petits que le seuil d'inondation ($k = 25$). La pente pour les points non inondés (rang supérieur à $k = 25$) ne peut pas cependant être estimée à partir de la Figure 9 (a) puisque elle est biaisée par les points inondés. Par conséquent le calcul est refait en Figure 9 (b) mais cette fois-ci après application de la censure. Le résultat de la calibration finale de la distribution des valeurs extrêmes exponentielle dans ces deux cas est donnée en Figure 5 (b). Signalons enfin, à propos de ce seuil d'inondation, que récemment, Navratil *et al.* (2002) se sont intéressés à ces seuils en les tabulant pour 12 stations en France.

4. Conclusions

Pour la calibration des distributions des valeurs extrêmes, l'analyse de l'indice des valeurs extrêmes (aussi appelé indice de queue) et la détermination du signe de cet indice est de première importance. Cependant, la distribution de la queue, et par conséquent l'exactitude de l'indice des valeurs extrêmes, est fortement affectée par les inondations. Au-dessus du seuil d'inondation, les interventions humaines modélisent la queue de la distribution et cela peut être très irrégulier (tout particulièrement dans les zones où il y a beaucoup de structures de régulation). Pour cette raison, une distinction claire doit être faite entre la distribution de l'écoulement naturel après précipitation et la distribution des débits. Les extrapolations basées sur l'analyse des valeurs extrêmes peuvent seulement être faites pour l'écoulement naturel (où l'hypothèse fondamentale de la théorie des valeurs extrêmes est valide).

Les inondations peuvent aussi biaiser la courbe de tarage des débits d'une rivière mesurés à partir d'une station limnigraphique, comme cela est montré pour la rivière Molenbeek en Belgique. Ce biais doit être éliminé pour éviter qu'une mauvaise distribution des valeurs extrêmes ne soit sélectionnée ou que des calibrations imprécises soient faites. Plusieurs estimateurs ont été présentés pour l'indice des valeurs extrêmes et les paramètres de la GPD. Les estimateurs sont basés sur une régression linéaire dans les « quantiles plots » et sur la minimisation de la MSE pour trouver le seuil optimal. Des techniques de réduction de biais ont été appliquées pour tenir compte de la différence entre la pente dans le « quantile plot » basée sur un nombre limité d'observations et la pente asymptotique dans la distribution des valeurs extrêmes.

Après (a) que le bon type de distribution ait été sélectionné (la bonne décision ait été prise sur le signe de l'indice et l'épaisseur de la queue de distribution), (b) qu'un seuil optimal ait été sélectionné, et (c) que les influences des inondations aient été prises en compte, alors les méthodes traditionnelles d'estimation de loi de probabilité (méthode des moments, du maximum de vraisemblance, des moments pondérés, ...) peuvent être correctement utilisées. Elles conduiront toutes à des résultats similaires (voir aussi Willems, 1998).

Concernant (c), les influences des inondations sont souvent importantes dans l'analyse de la fréquence des valeurs extrêmes et ces influences sont souvent négligées dans les applications pratiques. Pour cette raison, l'estimation de l'indice des valeurs extrêmes peut aussi être sujette à des erreurs. L'influence des inondations conduira

typiquement à des estimateurs de l'indice pouvant être négatifs, alors que les distributions des écoulements ou des débits admettent un indice nul. Il est clair que la qualité des données et la compréhension des facteurs physiques ne devraient pas être sous-évaluées. Comme il est mentionné par Bobée et Rasmussen (1995), l'analyse classique de la fréquence des inondations peut être critiquée en raison d'un manque d'équilibre entre les mathématiques d'une part et la physique d'autre part, la compréhension des phénomènes physiques causant ces événements d'inondation étant souvent négligée. Une discussion critique de ces aspects a été fournie par Klemes (1993) qui dit « si on doit faire plus de lumière sur les probabilités d'extrêmes hydrologiques, alors cela doit provenir de plus d'information sur les phénomènes physiques qui se produisent, mais pas sur des mathématiques ». Notre objectif dans cet article était donc de faire plus un équilibre entre la théorie statistique et la physique des inondations.

Remerciements

Patrick Willems est un postdoctorant de la Fondation Scientifique de Recherches de Flandre (F.W.O.-Vlaanderen). Ce travail est en partie financé par le Ministère de l'Écologie et du Développement Durable dans le cadre du programme Risque Inondation, Rio 2, ainsi que par le BQR 2004 de l'Université Pierre et Marie Curie. Les auteurs remercient le rapporteur pour l'ensemble de ses remarques constructives.

Références

BALKEMA A. et DE HAAN L. (1974), Residual life time at great age, *Ann. Probab.*, **2**, 792-804.

BEIRLANT J., DIERCKX G. et GUILLOU A. (2005), Estimation of the extreme value index and regression on generalized quantile plots, *Bernoulli*, 11, 6, 949-970.

BEIRLANT J. et GUILLOU A. (2001), Pareto index estimation under moderate right censoring, *Scand. Actuarial J.*, **2**, 111-125.

BEIRLANT J., GUILLOU A., DELAFOSSE E. et FILS-VILLETARD A. (2005), Estimation of the extreme value index and high quantiles under random censoring, *soumis*.

BEIRLANT J., VYNCKIER P. et TEUGELS J.L. (1996), Tail index estimation, Pareto quantile plots, and regression diagnostics, *J. Amer. Statist. Assoc.*, **91**, 1659-1667.

BERNIER J. (1967), Sur la théorie de renouvellement et son application en hydrologie, *Hyd.*, **67**, (10), Elec. France.

BOBÉE B. et RASMUSSEN P.F. (1995), Recent advances in flood frequency analysis, *Reviews of Geophysics, Supplement, American Geophysical Union*, 1111-1116.

BORGMAN L.E. (1963), Risk criteria, *J. Waterways and Harbors Division*, **80**, 1-35.

BUISHAND T.A. (1989), Statistics of extremes in climatology, *Statistica Neerlandica*, **43**, 1, 1-30.

CHAOUCHE A. et BACRO J.N. (2004), A statistical test procedure for the shape parameter of a generalized Pareto distribution, *Comput. Statist. Data Anal.*, **45**, 787-803.

CLAPS P. et LAIO F. (2003), Can continuous streamflow data support flood frequency analysis? An alternative to the partial duration series approach, *Water Resour. Res.*, **39**, 8, 1216, doi :10.1029/2002WR001868.

COLES S. (2001), *Introduction to statistical modelling of extremes values*, Springer Verlag.

CSÖRGŐ S., DEHEUVELS P. et MASON D. (1985), Kernel estimators of the tail index of a distribution, *Ann. Statist.*, **13**, 1050-1077.

DREES H. et KAUFMANN E. (1998), Selecting the optimal sample fraction in univariate extreme value estimation, *Stoch. Proc. Applications*, **75**, 149-172.

DURRANS S.R. et TOMIC S. (2001), Comparison of parametric tail estimators for low-flow frequency analysis, *J. American Water Resour. Assoc.*, **37**, 5, 1203-1214.

EMBRECHTS P., KLÜPPELBERG C. et MIKOSCH T. (1997), *Modelling extremal events*, Springer, Berlin.

FERRO C.A.T. et SEGERS J. (2003), Inference for clusters of extreme values, *J. Roy. Statist. Soc. Ser. B*, **65**, 545-556.

GNEDENKO B.V. (1943), Sur la distribution limite du terme maximum d'une série aléatoire, *Ann. Math.*, **44**, 423-453.

HALL P. (1982), On some simple estimates of an exponent of regular variation, *J. Roy. Statist. Soc. Ser. B*, **44**, 37-42.

HARREMOËS P. et MIKKELSEN P.S. (1995), Properties of extreme point rainfall I : Results from a rain gauge system in Denmark, *Atmos. Res.*, **37**, 277-286.

HILL B.M. (1975), A simple general approach to inference about the tail of a distribution, *Ann. Statist.*, **3**, 1163-1174.

HOSKING J.R.M. et WALLIS J.R. (1987), Parameter and quantile estimation for the generalized Pareto distribution, *Technometrics*, **29**, 339-349.

KAPLAN E.L. et MEIER P. (1958), Non-parametric estimation from incomplete observations, *J. Amer. Statist. Assoc.*, **53**, 457-481.

KLEMES V. (1993), Probability of extreme hydrometeorological events – a different approach, *In Extreme Hydrological Events : Precipitation, Floods and Droughts, IAHS Publ.*, **213**, 167-176.

KRATZ M. et RESNICK S. (1996), The qq-estimator and heavy tails, *Commun. Statist. Stochastic Models*, **12**, 699-724.

LANG M., OUARDA T.B.M.J. et BOBÉE B. (1999), Towards operational guidelines for over-threshold modeling, *J. Hydrol.*, **225**, 103-117.

NAVRATIL O., ALBERT M.B. et BREIL P. (2002), Water level time-series analysis for bank-full flow studies in rivers, in *River Flow 2002* (Eds. Bousmar & Zech), Swets & Zeitlinger, Lisse, 1167-1175.

PICKANDS III P. (1975), Statistical inference using extreme order statistics, *Ann. Statist.*, **3**, 119-131.

RASMUSSEN P.F. (2001), Generalized probability weighted moments : application to the generalized Pareto distribution, *Water Resour. Res.*, **37**, 1745-1751.

RASMUSSEN P.F., ASHKAR F., ROSBJERG D. et BOBÉE B. (1994), The POT method for flood estimation : a review, *Stochastic and Statistical Methods in Hydrology and Environmental Engineering*, K.W. Hipel (Ed.), Kluwer Academic Publishers, the Netherlands. Vol. 1 - Extreme values : floods and droughts : 15-26.

ROMBAUTS S. et WILLEMS P. (2003), Statistical analysis and composite hydrographs for the river Dender Basin (in Flemish), Technical Report, 165 pages, *Flemish Water Administration AWZ*, Antwerp, Belgium.

ROSBJERG D. (1987), On the annual maximum distribution in dependent partial duration series, *Stochastic Hydrol. Hydraul.*, **1**, 1, 3-16.

ROSBJERG D. et MADSEN H. (2004), Advanced approaches in PDS/POT modelling of extreme hydrological events, *Hydrology : Science & Practice for the 21st Century*, Vol. I.

ROSBJERG D., MADSEN H. et RASMUSSEN P.F. (1992), Prediction in partial duration series with generalized Pareto-distributed exceedances, *Water Resour. Res.*, **28**, 11, 3001-3010.

SCHULTZE J. et STEINEBACH J. (1996), On least squares estimates of an exponential tail coefficient, *Statist. Decisions*, **14**, 353-372.

SHANE R.M. et LYNN W.R. (1964), Mathematical model for flood risk evaluation, *J. Hydraulics Division*, **90**, 1-20.

SMITH R.L. (1987), Estimating tails of probability distributions, *Ann. Statist.*, **15**, 1174-1207.

TODOROVIC P. (1970), On some problems involving random number of random variables, *Ann. Math. Statist.*, **41**, 1059-1063.

USWRC (1976), Guidelines for determining flood flow frequency, *United States Water Resources Council*, Bull. 17, Hydrol. Comm. Washington, DC, 73 p.

WILLEMS P. (1998), Hydrological applications of extreme value analysis, *In : Hydrology in a changing environment*, H. Wheater and C. Kirby (ed.), John Wiley & Sons, Chichester, vol. III, 15-25.

WILLEMS P., VAES G., POPA D., TIMBE L. et BERLAMONT J. (2002), Quasi 2D river flood modelling, *In : River Flow 2002*, D. Bousmar and Y. Zech (eds.), Swets and Zeitlinger, lisse, Vol. 2, 1253-1259.

WILLEMS P. (2004), Extreme value analysis of rainfall-runoff and river discharges, under river flooding conditions, *soumis*.

WILLEMS P., GUILLOU A. et BEIRLANT J. (2005), Bias correction to the asymptotic properties of hydrological GPD based extreme value distributions, by means of a slowly varying function, *en révision à Water Resour. Res.*