Weighted estimate of extreme quantile : an application to the estimation of high flood return periods

Alexandre Lekina*†, Fateh Chebana†, Taha B. M. J. Ouarda†‡

November 2, 2012

Abstract

Parametric models are commonly used in Frequency Analysis of extreme hydrological events. To estimate extreme quantiles associated to high return periods, these models are not always appropriate. Therefore, estimators based on Extreme value Theory (EVT) are proposed in the literature. The Weissman estimator is one of the popular EVT-based semi-parametric estimators of extreme quantiles. In the present paper we propose a new family of EVT-based semi-parametric estimators of extreme quantiles. To built this new family of estimators, the basic idea consists in assigning the weights to the k observations being used. Numerical experiments on simulated data are performed and a case study is presented. Results show that the proposed estimators are smooth, stable, less sentitive, and less biased than Weissman estimator.

Keywords : flood, extreme quantile, bias reduction, heavy tailed distribution, order statistics, Weissman estimator.

*Corresponding author, Alexandre.Lekina@ete.inrs.ca.

†Canada Research Chair to the Estimation of Hydro-meteorological Variables, INRS-ETE, 490 rue de la Couronne, Quebec, Canada G1K 9A9.

‡Masdar Institute of Science and technology, PO Box 54224, Abu Dhabi, UAE.

1 Introduction

Extreme events and natural disasters (*e.g.* earthquakes, floods, storms, droughts, nuclear accidents, stock market crashes) dominate the daily news by their unpredictable nature. Given their considerable economic and social impacts, it is of high importance to develop the appropriate models for the prediction of these events. Frequency analysis (FA) procedures are commonly used for the analysis of extreme hydrological events. The main goal of the FA of flood events is the assessment of the probability of exceedence of an event x_T, *i.e.* $\mathbb{P}(X > x_T)$. Alternatively, given a return period T, it is also of interest to estimate the quantity x_T such that $\mathbb{P}(X > x_T) = 1/T$. The event x_T corresponds to the quantile associated to a return period T (*e.g.* Salvadori et al., 2007, chapter 1).

In hydrology, the floods x_T of interest are typically such that T is larger than n, where n denotes the sample size (for instance, the number of years of record at the gauging site). The traditional estimation procedure of x_T or T consists in choosing a parametric probability model $f(x;\theta)$ that is fully indexed by a finite parameter set θ (*e.g.* shape, scale and location parameters). Once the parameters θ of the model are estimated, the exceedance probability $1/T$ (resp. quantile x_T) is evaluated directly through the Cumulative Distribution Function (CDF) $F(x;\theta)$ of the fitted distribution (resp. via an estimator of the generalized inverse of $F(x;\theta)$) (*e.g.* Young-Il et al., 1993; Haddad and Rahman, 2011).

Despite all efforts, the topic of the choice of the best fitting parametric probability model $f(x;\theta)$ and parameter estimation method for flood FA remains elusive (Bobée et al., 1993). In some countries, standard distributions are recommended to fit hydrometeorological variables, *e.g.* the Generalized Extreme Value (GEV) distribution in the United Kingdom for flood FA and in the United States for precipitation, the Log-Pearson type 3 distribution in the United States and China for streamflows, the Lognormal distribution in China for low flows and floods (*e.g.* Chen et al., 2004; Chebana et al., 2010). Nevertheless, in practice several problems remain to be solved.

The FA approach based on the selection of a parametric probability distribution has a number of drawbacks especially for large T. First, this approach relies heavily on the initial choice of the parametric family of probability distributions. If this choice of distribution is inappropriate then, especially for large values of T, significant errors in quantile estimates are obtained. Second, the sample sizes of hydrological records are often too short for the appropriate selection of the best fitting distribution. Stedinger (2000) recommended a minimum sample size ($n = 50$) for robust estimates of quantiles. However, this size is often not sufficient to make the judicious choice of the appropriate distribution by using goodness-of-fit tests (*e.g.* Adlouni et al., 2008). The latter are rather sensitive to the behavior of the tail of the distribution. Third, the classical parametric estimation procedures are heavily weighted towards fitting the main body (central region) of the assumed probability density. On the other hand, they attribute a relatively low weight to the estimation of the distribution tail. Moreover, Young-Il et al. (1993) argued that this estimation procedure is an onerous mismatch in objectives since such parametric fits are not robust to outliers in the tail of the sample distribution. Also, as natural disasters may come from different causes, this can lead to mixtures of distributions. The tail behavior of a mixture is often dictated by the tail behavior of the distribution with the heaviest tail and by the relative proportion of events that correspond to each component (*e.g.* Young-Il et al., 1993).

The above drawbacks indicate that the parametric approach can be relatively unreliable. Since non-parametric approaches capture better any distributional features homogeneous or heterogeneous exhibited by the data, Apipattanavis et al. (2010) proposed a non-parametric FA estimator based on local polynomial regression. Notice that Adamowski et al. (1998) showed the advantages of using non-parametric methods in flood FA for both annual maximum and partial duration flood series. The local polynomial regression does not require a "priori" assumption of the underlying CDF and the estimation is local and data driven. The local aspect of the estimation provides the ability to capture any arbitrary features that might be present in the data. Kernel-based estimators have been studied respectively by (Lall et al., 1993; Moon and Lall, 1994), and Quintela-del-Río and Francisco-Fernández (2011) for flood FA and air quality modeling. In Regional flood frequency estimation, Epanechnikov kernel has been used by Ouarda et al. (2001)

Moreover, several authors have investigated methods based on the extreme value theory (EVT) (Fisher and Tippet, 1928; Gnedenko, 1943). These methods are based on the properties of the k upper order statistics of the sample and on extrapolation methods. Currently, three main categories of methods can be identified : *(i)* extrapolation method based on (GEV) (*e.g.* Prescott and Walden, 1980; Smith, 1985; Hosking et al., 1985; Guida and Longo, 1988); *(ii)* extrapolation method based on the excesses method and Generalized Pareto Distributions (GPD) (*e.g.* Balkema and de Haan, 1974; Pickands, 1975; Hosking and Wallis, 1987; Lang et al., 1999) with its variants so-called exponential tail and quadratic tail (Breiman et al., 1990); *(iii)* the semi-parametric and non-parametric methods (*e.g.* Hill, 1975; Pickands, 1975; Weissman, 1978; Dekkers and de Haan, 1989; Beirlant et al., 2005). All three categories are based on the statistical model given by the maximum domain of attraction (MDA) condition that governs EVT. Some comparison studies (theory and simulation) between the different methods can be found in Rosen and Weissman (1996); de Haan and Peng (1998); Tsourti and Panaretos (2001).

In the semi-parametric approach, one seeks to develop estimators of the right tail quantiles according to the tail behavior of the distribution. Thus, one assumes a parametric form only for the tail part and not for the entire probability density. The methods based on this approach are more flexible than parametric ones. The well-known Weissman (1978) estimator is a semi-parametric estimator of extreme quantiles. However, most semi-parametric estimators of quantiles x_T share a number of common problems. Most importantly, they are biased and sensitive to the selection of the k upper order statistics of the sample (Gomes and Oliveira, 2001).

The main objective of the present paper is to show that the usual practice in hydrological FA to estimate quantiles by inverting the CDF is not appropriate for extreme quantiles. Therefore, we present a number of alternatives to estimate these quantiles including, for instance, the Weissman (1978) estimator. In addition, we propose a new family of EVT-based semi-parametric estimators of extreme quantiles that are smooth, stable, less sentitive to the number of observations being used, and less biased than Weissman (1978) estimator.

The paper is organized as follows. In section 2, we present the statistical framework of the study and the background of EVT. In section 3, we propose the estimators of quantiles from heavy-tailed distributions. The numerical experiments on simulated data are presented and discussed in section 4 and the case study is carried out in section 5. Conclusions and some directions for future work are presented in section 6.

2 Statistical framework and background of EVT

2.1 General statistical framework

Let us denote by F the CDF of a random variable X and x_p the associated quantile of order $1-p$ defined by :

$$\mathbb{P}(X \leq x_p) = 1 - \mathbb{P}(X > x_p) = F\left(x_p\right) = 1 - p, \text{ for } p \in (0,1). \tag{1}$$

We consider a sample $\{X_i,\ i = 1, \ldots, n\}$ of independent and identically distributed random variables with distribution function F. We denote by $X_{1,n} \leq \ldots \leq X_{n,n}$ their associated order statistics. From the observations of these variables, the aim is to built an estimator of the quantile x_p when $p = 1/T$ is very small, *i.e.* close to zero since the return period T is large. In this context, we talk about *high return period.* Given any $p \in (0,1)$, the quantile x_p is defined via the generalized inverse of the CDF, *i.e.* $x_p = F^{\leftarrow}(1-p)$. Thus a natural estimator of x_p is given by :

$$\hat{x}_p = \hat{F_n}^{\leftarrow}(1-p), \tag{2}$$

where $\hat{F}_n$ is an estimator of the CDF F. In Extreme value analysis, in order to preserve (in the asymptotic analysis) the fact that the number of observations np above the quantile x_p should be much smaller than any positive constant, one assumes that p depends on n, *i.e.* $p = p_n$, and that $p_n \to 0$ as n increases (*e.g.* Dekkers and de Haan, 1989; de Haan and Ferreira, 2006). The terms *extreme quantile*, *large quantile* or *high quantile* mean that p_n converges to zero, see *e.g.* Gardes et al. (2010) and Embrechts et al. (1997, chapter 6). In particular, for n large enough, the non-exceedance probability $\mathbb{P}(X_{n,n} < x_p)$, can be approximated as :

$$\mathbb{P}\left(X_{n,n} < x_p\right) \simeq \mathrm{e}^{-np_n} \text{ as } p_n \to 0, \tag{3}$$

which represents the probability that the quantity of interest x_p falls outside the range of the sample. From a mathematical point of view, two cases can be considered from (3). Depending on the rate of convergence of p_n to zero, the probability in (3) could be 0 or not :

First, if $p_n \to 0$ and $np_n \to \infty$ as $n \to \infty$, then $\mathbb{P}(X_{n,n} < x_p) \to 0$. In this situation, p_n goes to zero slower than $1/n$ and x_p is eventually *almost surely* smaller than the largest observation

$X_{n,n}$. Consequently, the estimation of the extreme quantile requires to interpolate inside the sample. In this context, the natural and basic estimator of x_p is given by (2). For instance, the $\lfloor np_n \rfloor$-th largest observation of the sample $\{X_i,\ i = 1, \ldots, n\}$, *i.e.* $X_{n-\lfloor np_n \rfloor+1,n}$, is an option (refer to Rényi, 1953; Dekkers and de Haan, 1989), where the symbol $\lfloor \bullet \rfloor$ denotes the floor function.

Second, if $p_n \to 0$ and $np_n \to c \neq \infty$ as $n \to \infty$, then $\mathbb{P}(X_{n,n} < x_p) \to \mathrm{e}^{-\mathrm{c}}$. In this context, the estimation of extreme quantiles may need extrapolation beyond the observations since x_p could be outside the sample, *i.e.* after the largest observation. According to the value of c, two situations arise :

When $c \in [1, \infty)$, it is possible to estimate x_p by (2), or basically by the $\lfloor c \rfloor$-th largest observation of the sample, since the estimation is based on the largest observations located near the border of the sample, but still within the data set. Nevertheless, recall that the $\lfloor c \rfloor$-th largest observation of a sample is asymptotically not Gaussian (Embrechts et al., 1997, corollaire 4.2.4).

When $c \in [0, 1)$, then p_n goes to zero at the same speed or faster than $1/n$ and x_p is eventually larger that the maximal observation $X_{n,n}$ with probability $\mathrm{e}^{-\mathrm{c}} \geq \mathrm{e}^{-1}$. In this case, the estimation of x_p is more difficult since it requires an estimation outside the sample. For instance, the quantile of order $(1-p_n)$ with $p_n < 1/n$ is extreme and is eventually larger than the maximum observation of the sample. Therefore, it is not appropriate to estimate it simply by inverting the CDF F. In predictions, the values of quantiles exceeding the length of the series are generally extrapolation values that exceed the largest observation of the sample.

We illustrate in Figure 1 the difference between large quantiles within and outside the sample. More precisely, Figures 1-(a) and 1-(b) describe the large quantile within the sample, while Figure 1-(c) describes the large quantile outside the sample. To illustrate the difference between the two quantiles, we generated a Fréchet distributed sample of size $n = 500$. In hydrology, this distribution is applied to extreme events such as river discharges and annual maximum 1-day rainfall (*e.g.* Coles, 2001).

In Figure 1-(a), $p = 1/25 = 0.04$ and the quantile $x_{1/25}$ is clearly smaller than the largest observation of the sample. Since we have $c = 20$ observations above $x_{1/25}$, then a non-parametric estimator of quantile $x_{1/25}$ obtained by interpolation is the 20-th largest observation, *i.e.* $X_{481,500}$. In Figure 1-(b), $p = 1/250 = 0.004$ and the estimation of the quantile becomes difficult since it is based on the $c = 2$ observations above $x_{1/250}$ and located near the border of the sample. In the case of Figure 1-(c) $p = 1/600 \simeq 0.0017$ and the quantile $x_{1/600}$ is larger than the largest observation of the sample. To estimate $x_{1/600}$ one needs to extrapolate beyond the largest observation of the sample.

When the number of observations above x_p is finite, *i.e.* $c \neq \infty$, one has to extend the empirical distribution function beyond the sample. EVT studies the behavior of the k largest observations of a sample and provides laws governing these values, and as such forms the natural framework for estimating the event x_p when $c \in [0, 1)$, where the quantile of interest is eventually larger than the maximal observation.

de Haan (1984) has established the first result in the case where $c = 0$. Dekkers and de Haan (1989) have studied the case $c = \infty$ and $c \in [0, 1)$. A summary of these results can be found in (Embrechts et al., 1997, Theorem 6.4.14 and Theorem 6.4.15). Gardes et al. (2010), Daouia et al. (2011) and Lekina (2010) provide an extension of situations $c = \infty$, $c \geq 1$ and $c \in [0, 1)$ in the conditional case, that is to say in the situation where the variable of interest X is recorded simultaneously with some covariate information. In the next section, we present a brief summary of EVT.

2.2 EVT background

In the literature, several estimation methods of the extreme quantile x_p where $p \simeq 0$ have been proposed, for instance in finance (Embrechts et al., 1997), in engineering structures (Ditlevsen, 1994) and in hydrology (Smith, 1987, 1986). These methods are based on the statistical model given by the MDA condition that governs EVT (Fisher and Tippet, 1928; Gnedenko, 1943). The main result of EVT shows that under some regularity conditions on the CDF F of X, there exist a parameter $\gamma \in \mathbb{R}$ and two sequences $(a_n)_{n\geq 1} > 0$ and $(b_n)_{n\geq 1} \in \mathbb{R}$ such that for all $x \in \mathbb{R}$,

$$\lim_{n\to\infty} \mathbb{P}\left[\frac{X_{n,n} - b_n}{a_n} \leq x\right] = \mathcal{H}_\gamma(x), \tag{4}$$

where $\mathcal{H}_\gamma(.)$ is a non-degenerate extreme value distribution defined by

$$\mathcal{H}_\gamma(x) = \begin{cases} \exp\left[-(1+\gamma x)^{-1/\gamma}\right] & \text{if } \gamma \neq 0 \\ \exp\left[-\exp(-x)\right] & \text{if } \gamma = 0 \end{cases} \quad \text{and for all } x \text{ such that } 1+\gamma x > 0. \tag{5}$$

The main result in (4) is true for most usual distributions F. If we make a parallel with the Central Limit Theorem (CLT), the sequence a_n plays the role of $n^{-1/2}\sigma(X)$ where $\sigma(X)$ denotes the standard deviation of X and the sequence b_n plays the role of the mathematical expectation of X. The sequences a_n and b_n are respectively interpreted as scale and location parameters. Note that these sequences are not unique. The reader is referred to Embrechts et al. (1997) for some examples of a_n and n_n in the fields of insurance and finance. A limited number of examples are presented in Table 1.

The parameter γ in (5) is called *extreme value index* and it has no equivalent in CLT. This index is known to be the crucial indicator for the decay behaviour of the distribution tail. It clearly governs the tail behavior, with larger values indicating heavier tails. If the cdf F satisfies the Fisher and Tippet (1928) theorem conditions, then F belongs to MDA of $\mathcal{H}_\gamma(.)$. According to the sign of γ, we distinguish the cases :

- Fréchet MDA ($\gamma > 0$) includes the distributions with polynomially decreasing Pareto-type tails, *e.g.* Cauchy, Pareto and Burr. This family has a rather heavy right tail;
- Weibull MDA ($\gamma < 0$) includes the distributions with finite right endpoint, *e.g.* uniform and beta;
- Gumbel MDA ($\gamma = 0$) includes distributions with exponentially decreasing tails, *e.g.* normal, exponential and Gamma. The distributions of this MDA are rather light tailed.

To check the assumption that F belongs to MDA of $\mathcal{H}_\gamma(.)$, several techniques are available. For a review on exploratory data analysis methods for extremes the reader is refereed *e.g.* to Embrechts et al. (1997, section 6.2). In extreme value-analysis, the Pareto quantile plot (PQ-plot) is based on :

$$\left\{\left(\log\frac{n+1}{j},\ X_{n-j+1}\right),\ j=1,\ \ldots,n\right\}, \tag{6}$$

and is widely used to graphically check if data are distributed according to a MDA(Fréchet) or not. If F is heavy-tailed, *i.e.* belongs to MDA(Fréchet), then the PQ-plot will be approximately linear with a positive slope for small values of j associated to the extremes points. Alternately, we can use the quantile-quantile plot (QQ-plot) or the generalized quantile plot (GQ-plot). The GQ-plot is based on (*e.g.* Willems et al., 2007) :

$$\left\{ \left(\log \frac{n+1}{j},\ \frac{X_{n-j}}{j} \sum_{i=1}^{j} \log \frac{X_{n-i+1,n}}{X_{n-j,n}} \right),\ j = 1,\ \ldots, n \right\}. \tag{7}$$

According to the curve of this graph, we can deduce the MDA associated to F. If for the extreme points, *i.e.* small value of j, the slope is positive, then F belongs to MDA(Fréchet) and if it is approximately constant, then F belongs to MDA(Gumbel). Finally, the case of a linear decrease means that F belongs to MDA(Weibull).

3 Proposed extreme quantile estimators

The aim of this section is to propose estimators of extreme quantiles when $c \neq \infty$. We deal with an estimation problem within the case where the CDF F is heavy-tailed or Pareto-type. The case where the distribution F is light-tailed or finite endpoint will be examined in future work. However, there exist abundant literature on light-tailed distributions (*e.g.* Diebolt et al., 2008; Beirlant et al., 1995, 1996a; Dierckx et al., 2009) and finite endpoint distributions (*e.g.* Falk, 1995; Hall and Park, 2002; Girard et al., 2012; Li and Peng, 2009). In the considered situation, for all $x > 0$ and for some unknown tail index $\gamma > 0$, the CDF F is of the form :

$$F(x) = 1 - x^{-1/\gamma} L(x), \tag{8}$$

where $L(.)$ is a slowly varying function at infinity, *i.e.* for all $\lambda > 0$,

$$L(\lambda x)/L(x) \to 1 \text{ as } x \to \infty. \tag{9}$$

Assumption (8) is also equivalent to stating that $\bar{F} = 1 - F$ is regularly varying at infinity with an index $-1/\gamma$. The reader is referred to Bingham et al. (1987) for a detailed reference on regular variation theory. The heavy-tailed model in (8) can also be stated in an equivalent way in terms of the quantile function as :

$$x_{p_n} = p_n^{-\gamma} \ell(p_n^{-1}), \tag{10}$$

where $p_n \in [0,1]$ and $\ell(.)$ is a slowly varying function at infinity (see Bingham et al., 1987, Theorem 1.5.12). Property (10) characterizes heavy-tailed distributions. Note that from condition (9) and property (10), the quantile x_{p_n} decreases towards 0 at a polynomial rate driven by γ. We remark that model (8) (resp. (10)) includes a parametric part $x^{-1/\gamma}$ (resp. $p_n^{-\gamma}$) depending only on a parameter γ and a non-parametric part $L(.)$ (resp. $\ell(.)$). Hence, (8) and (10) represent semi-parametric models.

Let $(k_n)_{n\geq 1}$ be an *intermediate sequence* corresponding to the fraction sample such that $1 \leq k_n < n$. Under (10), Weissman (1978) proposed to estimate, semi-parametrically, the extreme quantile x_{p_n} by :

$$\hat{x}_{p_n}^{\mathrm{W}} := \hat{x}_{p_n}^{\mathrm{W}}(k_n) = X_{n-k_n+1,n} \left(\frac{k_n}{np_n} \right)^{\hat{\gamma}_{k_n}^{\mathrm{H}}}, \tag{11}$$

where $\hat{\gamma}_{k_n}^{\mathrm{H}}$ is the Hill (1975) estimator of γ defined by :

$$\hat{\gamma}_{k_n}^{\mathrm{H}} = \frac{1}{k_n} \sum_{j=1}^{k_n} j \left\{ \log X_{n-j+1,n} - \log X_{n-j,n} \right\}. \tag{12}$$

Often used in hydrology (*e.g.* Young-Il et al., 1993), Weissman estimator (11) includes two terms. The first term, $X_{n-k_n+1,n}$ is the k_n-th largest observation of the sample, and the second term, $(k_n/(np_n))^{\hat{\gamma}_{k_n}^{\mathrm{H}}}$ is the extrapolation factor that allows to estimate extreme quantiles of an order $(1-p_n)$ arbitrarily large, *i.e.* p_n arbitrarily small.

The accuracy of estimators (11) and (12) depends on a precise choice of the sample fraction k_n, that corresponds to the number of order statistics, on which the estimation is based. The Weissman plot $\{(k_n, \hat{x}_{p_n}^{\mathrm{W}}),\ k_n = 1, \ldots, n-1\}$ described in section 4 shows a large volatility which represents a practical difficulty if no prior indication on k_n is available. Moreover, this estimator is biased. Indeed most semi-parametric estimators of extreme quantile x_{p_n} or tail index γ have similar problems : high variance for small values of k_n and high bias for large value of k_n (*e.g.* Gomes and Oliveira, 2001).

The limiting distributions for several semi-parametric estimators of γ and x_{p_n}, especially $\hat{\gamma}_{k_n}^{\mathrm{H}}$ and $\hat{x}_{p_n}^{\mathrm{W}}$, are established usually under a second order condition, not too restrictive, on the tail behavior. This second order condition assumes that there exists a constant $\rho < 0$ and the *bias function* $b(x) \to 0$ as $x \to \infty$, such that for all $\lambda > 1$,

$$\log \frac{\ell(\lambda x)}{\ell(x)} \sim b(x) \frac{\lambda^\rho - 1}{\rho} \text{ as } x \to \infty. \tag{13}$$

To improve the bias of the estimators $\hat{\gamma}^{\mathrm{H}}_{k_n}$ and $\hat{x}^{\mathrm{W}}_{p_n}$, the most common approach consists in assuming that the second order condition (13) holds with the bias function $b(x) = \gamma D x^\rho$ where $\rho < 0$ is a *second order shape parameter* and $D \neq 0$ is a *second order scale parameter* (de Wet et al., 2012; Goegebeur et al., 2010; Caeiro and Gomes, 2006; Caeiro et al., 2009). Thus, the problem of estimation of γ or x_{p_n} can be summarized in the estimation of the second order parameters ρ and D. This is the currently challenging estimation problem. Concisely, the second order parameter $\rho < 0$ tunes the convergence rate of $\ell(\lambda x)/\ell(x)$ to 1 in (9). The closer ρ is to 0, the slower the convergence will be, and the estimation of the tail parameter γ or quantile x_{p_n} will typically be difficult in practice.

In order to obtain an estimator of extreme quantile that is less sensitive to the selection of the sample fraction k_n, the basic idea of the present work involves doing the geometric mean of Weissman estimators. Intuitively, this idea is due to the fact that the bias of extreme quantiles increases for large values of k_n. Thus, instead of considering only the k_n-th largest observation of the sample as in Weissman (1978), one proposes to attribute equal importance to the k_n largest observations of the same sample. It consists in assigning the same weight to each observation of the subsample $\{X_{n-i+1,n},\ i = 1, \ldots, k_n\}$. Note that Drees (1995) applied a similar idea for the tail index estimator proposed by Pickands (1975). Here, unlike in bias correction methods, prior knowledge of new tuning parameters (especially the second-order parameters ρ and D) is not required and thus there is no need for an analysis related to these extra parameters. Therefore, the second-order refinements are not used in the remainder of the paper.

In order to estimate extreme quantiles of an order $(1 - p_n)$ arbitrarily large, we propose an estimator of high quantiles originally introduced in Lekina (2010, chapter 2) and defined by :

$$\hat{x}^{\mathrm{WG}}_{p_n} = \left[\prod_{i=1}^{k_n} X_{n-i+1,n} \left(\frac{i g_{k_n}}{n p_n} \right)^{\hat{\gamma}^{\mathrm{H}}_i} \right]^{1/k_n}, \tag{14}$$

where $g_{k_n} = \exp\left[\log(k_n + 1) - 1 - \log(k_n!)/k_n\right]$ and $\hat{\gamma}^{\mathrm{H}}_i$ is the Hill tail index estimator defined in (12). In order to obtain properties of the extreme quantile estimator in (14), $\hat{x}_{p_n}^{\mathrm{WG}}$ can be decomposed as follows (see Lekina, 2010, Proposition 2.2.1) :

$$\log \hat{x}_{p_n}^{\mathrm{WG}} \stackrel{\mathcal{D}}{=} \hat{\gamma}_{k_n}^{\mathrm{H}} - \gamma \log V_{k_n+1,n} + \log \ell\left(1/V_{k_n+1,n}\right) + \log\left(\frac{1}{\mathrm{e}}\frac{(k_n+1)}{np_n}\right)\hat{\gamma}_{k_n}^{\pi}, \tag{15}$$

where $\ell(.)$ is a slowly varying function at infinity, $V_{k_n+1,n}$ is the $(n-k_n)$-th upper order statistic of a sample of independent random variables $\{V_i,\ i = 1,\ldots,n\}$ uniformly distributed on $(0,1)$ and $\hat{\gamma}_{k_n}^{\pi}$ is a tail index estimator given by :

$$\hat{\gamma}_{k_n}^{\pi} = \sum_{j=1}^{k_n} j\left\{\log X_{n-j+1,n} - \log X_{n-j,n}\right\}\pi_j \Bigg/ \sum_{j=1}^{k_n}\pi_j\ , \tag{16}$$

with $\{\pi_j,\ j = 1,\ldots,k_n\}$ is a weighted function defined by

$$\pi_j = \sum_{i=j}^{k_n}\frac{1}{i}\log\left(\frac{ig_{k_n}}{np_n}\right). \tag{17}$$

Notice that the weights $\{\pi_j,\ j = 1,\ldots,k_n\}$ are a consequence of decomposition (15) and are not to be selected and one cannot attribute to them other quantities. Recall that the decomposition of the Weissman estimator is (*e.g.* Beirlant et al., 2004) :

$$\log \hat{x}_{p_n}^{\mathrm{W}} \stackrel{\mathcal{D}}{=} -\gamma \log V_{k_n,n} + \log \ell\left(1/V_{k_n,n}\right) + \log\left(\frac{k_n}{np_n}\right)\hat{\gamma}_{k_n}^{\mathrm{H}}, \tag{18}$$

where $V_{k_n,n}$ is the $(n-k_n+1)$-th upper order statistic of a sample of independent random variables $\{V_i,\ i = 1,\ldots,n\}$ uniformly distributed on $(0,1)$.

By comparing (15) and (18), notice that the representation of $\hat{x}_{p_n}^{\mathrm{WG}}$ involves an additional tail index estimator $\hat{\gamma}_{k_n}^{\pi}$. This estimator is a weighted sum of the log-spacings between the k_n largest order statistics $X_{n-k_n+1,n},\ldots,X_{n,n}$. According to Feuerverger and Hall (1999) and Beirlant et al. (2002), it is possible to establish the asymptotic distribution of $\hat{\gamma}_{k_n}^{\pi}$. In addition, under a restrictive condition $\log(k_n)/\log(np_n) \to 0$, Lekina (2010) has shown that the tail index estimator $\hat{\gamma}_{k_n}^{\pi}$ and the least-squares estimator of the tail index so-called Zipf (see Kratz and Resnick, 1996; Schultze and Steinebach, 1996) have the same limiting distribution. Thus, we can build confidence intervals for estimates of the extreme quantile $\hat{x}_{p_n}^{\mathrm{WG}}$. Indeed, decomposition (18) shows that the extreme quantile $\hat{x}^{\mathrm{W}}_{p_n}$ inherits its limiting distribution of the tail index estimator $\hat{\gamma}^{\mathrm{H}}_{k_n}$ or the largest upper order statistic $X_{n-k_n+1,n}$, in fact of $V_{k_n,n}$, (*e.g.* Gardes et al., 2010, for more details). Decomposition (15) shows that the limiting distribution of $\hat{x}^{\mathrm{WG}}_{p_n}$ may depend on the behavior of both $X_{n-k_n,n}$ (or $V_{k_n+1,n}$), $\hat{\gamma}^{\mathrm{H}}_{k_n}$ and $\hat{\gamma}^{\pi}_{k_n}$. In the EVT-literature, the limiting distribution of $\hat{\gamma}^{\mathrm{H}}_{k_n}$ and the upper order statistics have been established, for instance, respectively in Haeusler and Teugels (1985) and (Dekkers and de Haan, 1989; Rényi, 1953). Under the conditions $\log(k_n)/\log(np_n) \to 0$ and $k_n^{1/2} b(n/k_n) \to \lambda \in \mathbb{R}$ as $n \to \infty$, Lekina (2010, Theorem 2.2.1) showed that estimator $\hat{x}^{\mathrm{WG}}_{p_n}$ is asymptotically Gaussian and the asymptotic bias is given by $b(n/k_n)/(1-\rho)^2$. The latter is better, apart from the scale factor $1/(1-\rho)$, than the bias of estimator $\hat{x}^{\mathrm{W}}_{p_n}$.

The direct consequence of decomposition (15) is the introduction of an adaptation of the Weissman estimator given by :

$$\hat{x}^{\mathrm{L}}_{p_n} = X_{n-k_n+1,n} \left(\frac{k_n}{np_n} \right)^{\hat{\gamma}^{\pi}_{k_n}}, \tag{19}$$

which is valid for $p_n < 2/(n\mathrm{e})$ and $1 \leq k_n < n$. The condition $p_n < 2/(n\mathrm{e})$ is not restrictive since it ensures that the weight function $\{\pi_j,\ j = 1, \ldots, k_n\}$ is always positive and decreasing. If $p_n = 2/(n\mathrm{e})$ then, $\pi_j = 0$ for $j = k_n = 1$ and estimator (19) is valid for $2 \leq k_n < n$. Otherwise, if $p_n > 2/(n\mathrm{e})$ then for some integer $j \leq k_n < n$, the weight function is non-monotonous and can be even negative for small values of k_n. The decomposition in the distribution of $\hat{x}^{\mathrm{L}}_{p_n}$ is similar to that of $\hat{x}^{\mathrm{W}}_{p_n}$. It is sufficient to replace $\hat{\gamma}^{\mathrm{H}}_{k_n}$ in (18) by $\hat{\gamma}^{\pi}_{k_n}$. However, unlike $\hat{x}^{\mathrm{L}}_{p_n}$, $\hat{x}^{\mathrm{W}}_{p_n}$ can be used for $p_n \in (0,1)$ and $1 \leq k_n < n$.

It is also possible to redefine estimator (14) by replacing $\hat{\gamma}^{\mathrm{H}}_i$ by $\hat{\gamma}^{\pi}_i$. However, in this case, one needs to exactly reassess the renormalizing sequence g_{k_n}. In (14), g_{k_n} was computed by studying the asymptotic behaviour of estimator $\hat{x}^{\mathrm{W}}_{p_n}$. One can therefore use the same approach to evaluate the sequence f_{k_n} in definition (20) of the extreme quantile below. Nevertheless, since estimator (14) is interpreted as a geometric mean of (11), it follows that, for k_n large enough, $g_{k_n} \simeq 1$. Thus, it is still possible to fix $g_{k_n} = f_{k_n} = 1$ for the applications. Let f_{k_n} be a positive and non-decreasing sequence such that $f_{k_n} \simeq 1$ for k_n large enough. We introduce a second geometric estimator of extreme quantiles defined by :

$$\hat{x}_{p_n}^{\mathrm{LG}} = \left[\prod_{i=1}^{k_n} X_{n-i+1,n} \left(\frac{i f_{k_n}}{n p_n} \right)^{\hat{\gamma}_i^{\pi}} \right]^{1/k_n} \quad \text{with } p_n < 2/(n\mathrm{e}). \tag{20}$$

The following section provides an evaluation of the performance of this estimator.

4 Numerical experiments on simulated samples

In this section, we evaluate and compare the performance of the estimators $\hat{x}_{p_n}^{\mathrm{W}}$, $\hat{x}_{p_n}^{\mathrm{WG}}$, $\hat{x}_{p_n}^{\mathrm{L}}$ and $\hat{x}_{p_n}^{\mathrm{LG}}$ given in section 3 on a number of finite simulated samples. In order to evaluate the influence of the sequence f_{k_n}, we compute two versions of the estimator $\hat{x}_{p_n}^{\mathrm{LG}}$. Thus, we denote by $\hat{x}_{p_n}^{\mathrm{LG}(1)}$ (resp. $\hat{x}_{p_n}^{\mathrm{LG}(2)}$) the corresponding estimator associated to $f_{k_n} = 1$ (resp. $f_{k_n} = g_{k_n}$).

Let m, s and ρ be respectively a location, scale and second order parameter. We consider the following distributions which belong to the MDA(Fréchet) and are commonly used in hydrological frequency analysis (*e.g.* Brunet-Moret, 1969; Coles, 2001) :

- Fréchet with CDF $\mathcal{F}(x;\gamma,s,m) = \exp\left(-\left(\frac{x-m}{s}\right)^{-1/\gamma}\right)$ where $x > 0$, $m \in \mathbb{R}$ and $s > 0$,
- Burr with CDF $\mathcal{B}(x;\gamma,\rho) = 1 - \left(1 + x^{-\rho/\gamma}\right)^{1/\rho}$ where $x > 0$ and $\rho < 0$,
- Pareto with CDF $\mathcal{P}(x;\gamma,s) = 1 - \left(\frac{x}{s}\right)^{-1/\gamma}$ where $x \geq s > 0$,
- Student with CDF $\mathcal{ST}(x;\nu) = \dfrac{1}{2} + \dfrac{x\Gamma\left(\frac{1}{2}(\nu+1)\right) {}_2F_1\left(\frac{1}{2}, \frac{1}{2}(\nu+1); \frac{3}{2}; \frac{-x^2}{\nu}\right)}{(\nu\pi)^{1/2}\,\Gamma\left(\dfrac{1}{2}\nu\right)}$ where ν is the number of degrees of freedom, $x \in \mathbb{R}$, $\Gamma(z)$ is the gamma function and ${}_2F_1(a,b;c;z)$ is a hypergeometric function.

These four distributions satisfy models (8) and (10) but the Pareto distribution is the one for which the slowly varying functions $L(.)$ and $\ell(.)$ are constant.

For each of the distributions of Fréchet $\mathcal{F}(.;3/4,1,0)$, Burr $\mathcal{B}(.;3/4,-1)$, Pareto $\mathcal{P}(.;1,2)$ and Student $\mathcal{ST}(.;10)$, we generate $N = 1000$ samples of size $n \in \{30, 50, 100, 500\}$. Results for $N > 1000$ are not significantly different. The main goal is to estimate the extreme quantile of order $(1 - p_n)$ with $p_n = 1/(5n)$, *i.e.* for a return period $T = 5n$. For such a return period, an extrapolation is needed since $c = 1/5 \in [0,1)$ (the reader is referred to section 2). For each distribution and each sample size, we evaluate the mean for the bias and the modified mean square error (noted AMSE) of the considered estimators. The AMSE associated to estimator $\hat{x}^{\bullet}_{p_n}$ is defined by $\mathbb{E}\left(\log^2(\hat{x}^{\bullet}_{p_n}/x_{p_n})\right)$ which is estimated for a fixed sample fraction k_n by the quantity :

$$\mathrm{AMSE}\left(\hat{x}^{\bullet}_{p_n}\right) = \frac{1}{N}\sum_{j=1}^{N} \log^2(\hat{x}^{\bullet,\mathrm{j}}_{p_n}/x_{p_n}). \tag{21}$$

As those are the logarithms of extreme quantiles that are Gaussian, in EVA the logarithm employed in (21) is to insure the asymptotic normality (*e.g.* Beirlant et al., 2004, p. 120). We are also interested in the median estimator. This one is the estimator associated to median error.

For each sample size and for each of the four distributions, we superimposed in Figure 2 the mean estimators and the true theoretical quantile x_{p_n}, in Figure 3 the median estimators and x_{p_n} and in Figure 4 the AMSE corresponding to estimators $\hat{x}^{\mathrm{W}}_{p_n}$, $\hat{x}^{\mathrm{WG}}_{p_n}$, $\hat{x}^{\mathrm{L}}_{p_n}$ and $\hat{x}^{\mathrm{LG}}_{p_n}$. For visualization, we use a logarithmic scale in Figures 2 and 3. For each of the three Figures, we have sixteen pictures that we numbered for clarity (i)–(xvi).

In the remainder of the paper, for the sake of simplicity, the symbols ↑ and ↓ are employed to denote the expressions *increases* and *decreases* respectively. The discussion is done first and foremost by distribution, afterwards by sample size if there is no redundancy. Otherwise case are grouped.

Mean estimators

In Figure 2, except for the behavior of the mean estimators of $\hat{x}^{\mathrm{L}}_{p_n}$ when $k_n \simeq n$ with $n \geq 50$, the graphs of $\hat{x}^{\mathrm{W}}_{p_n}$, $\hat{x}^{\mathrm{WG}}_{p_n}$, $\hat{x}^{\mathrm{L}}_{p_n}$, $\hat{x}^{\mathrm{LG(1)}}_{p_n}$ and $\hat{x}^{\mathrm{LG(2)}}_{p_n}$ are convex. Except for the Pareto distribution for which the slowly varying $\ell(.)$ is constant, the simulations show that for the three other distributions (Fréchet, Burr and Student) the bias of the extreme quantile estimators ↑ as the sample size n ↑. This is due to the fact that the estimation of extreme quantiles of an order $(1 - 1/(5n))$ is more difficult when n ↑. In other words, this phenomenon is a consequence of $1/150 < 1/2500$ which means that estimating $x_{1/2500}$ in Figures 2-(d) is more difficult than estimating $x_{1/150}$ in Figures 2-(a).

For the distributions of Fréchet and Burr, the estimators $\hat{x}^{\mathrm{W}}_{p_n}$, $\hat{x}^{\mathrm{WG}}_{p_n}$ and $\hat{x}^{\mathrm{L}}_{p_n}$ have high bias for large values of the fraction sample k_n. For large values of k_n this bias ↑ as k_n ↑ while, for its small values this bias $\downarrow$ as $k_n \uparrow$. We note a different behavior of the estimators $\hat{x}_{p_n}^{\mathrm{LG}(1)}$ and $\hat{x}_{p_n}^{\mathrm{LG}(2)}$: **(1)** for sample size $n \in \{30, 50\}$, the bias of these estimators $\downarrow$ as $k_n \uparrow$; **(2)** for $n = 100$, this bias $\downarrow$ and becomes almost constant for large values of k_n; **(3)** when $n = 500$, for small values of k_n the bias $\downarrow$ as $k_n \uparrow$ and for large values of k_n the bias $\uparrow$ very slowly as $k_n \uparrow$.

Regarding the Student distribution, all estimators have high and $\uparrow$ bias for large values of k_n whatever the sample size. For very small values of k_n, this bias $\downarrow$ as $k_n \uparrow$.

In addition, whatever the sample size and for each of the three distributions viz Fréchet, Burr and Student, the bias of estimators $\hat{x}_{p_n}^{\mathrm{WG}}$, $\hat{x}_{p_n}^{\mathrm{L}}$, $\hat{x}_{p_n}^{\mathrm{LG}(1)}$ and $\hat{x}_{p_n}^{\mathrm{LG}(2)}$ becomes significantly less important than the one of $\hat{x}_{p_n}^{\mathrm{W}}$ as $k_n \uparrow$. Given a sample fraction k_n not too small, *e.g.* $k_n \simeq 2n/5$, the simulations in Figure 2 show that, for the small sample sizes $n \leq 100$, the bias of estimators $\hat{x}_{p_n}^{\mathrm{WG}}$, $\hat{x}_{p_n}^{\mathrm{L}}$, $\hat{x}_{p_n}^{\mathrm{LG}(1)}$ and $\hat{x}_{p_n}^{\mathrm{LG}(2)}$ is lower than the bias of Weissman estimator $\hat{x}_{p_n}^{\mathrm{W}}$. Thus, for these three distributions, the estimators $\hat{x}_{p_n}^{\mathrm{WG}}$, $\hat{x}_{p_n}^{\mathrm{L}}$, $\hat{x}_{p_n}^{\mathrm{LG}(1)}$ and $\hat{x}_{p_n}^{\mathrm{LG}(2)}$ improve the bias of $\hat{x}_{p_n}^{\mathrm{W}}$.

Regarding the Pareto distribution, since its slowly varying function $\ell(.)$ is constant and therefore its bias function $b(.) \equiv 0$ then, there is no asymptotic bias, *i.e.* the bias decreases and becomes negligible as the sample size n and the fraction sample $k_n \uparrow$. For small n, the Weissman estimator seems to be better than the other estimators. Nevertheless, when the sample size $n \uparrow$, all these estimators are approximately similar.

Median estimators

Generally, we observe from Figure 3 that the median estimators of $\hat{x}_{p_n}^{\mathrm{WG}}$, $\hat{x}_{p_n}^{\mathrm{L}}$, $\hat{x}_{p_n}^{\mathrm{LG}(1)}$ and $\hat{x}_{p_n}^{\mathrm{LG}(2)}$ are smooth and more stable than the Weissman estimator $\hat{x}_{p_n}^{\mathrm{W}}$ whatever the sample size. The previous findings in Figure 2 on the bias of the estimators $\hat{x}_{p_n}^{\mathrm{W}}$, $\hat{x}_{p_n}^{\mathrm{WG}}$, $\hat{x}_{p_n}^{\mathrm{L}}$, $\hat{x}_{p_n}^{\mathrm{LG}(1)}$ and $\hat{x}_{p_n}^{\mathrm{LG}(2)}$ are generally valid. Like the Weissman estimator $\hat{x}_{p_n}^{\mathrm{W}}$, the other estimators have high variance for small values of k_n and high bias for large values of k_n. Indeed for the Fréchet, Burr and Student distributions, if k_n is large then the approximation $\ell(.)$ is constant becomes worse and this implies a high bias. Nevertheless, the bias of $\hat{x}_{p_n}^{\mathrm{WG}}$, $\hat{x}_{p_n}^{\mathrm{L}}$, $\hat{x}_{p_n}^{\mathrm{LG}(1)}$ and $\hat{x}_{p_n}^{\mathrm{LG}(2)}$ is less significant than $\hat{x}_{p_n}^{\mathrm{W}}$. However for the Pareto distribution, the bias is negligible when k_n is large since $\ell(.)$ is constant. If k_n is small, one has too few observations, this implies then a high variance and a small bias since one remains in the tail of the distribution.

AMSE

In Figure 4, for the four distributions we observe that $\mathrm{AMSE}(\hat{x}_{p_n}^{\mathrm{W}})$ is slightly less smooth than those of its competing estimators. Except for $\mathrm{AMSE}(\hat{x}_{p_n}^{\mathrm{L}})$ when $k_n \simeq n$ with $n \geq 50$, the graphs of $\mathrm{AMSE}(\hat{x}_{p_n}^{\mathrm{W}})$, $\mathrm{AMSE}(\hat{x}_{p_n}^{\mathrm{WG}})$, $\mathrm{AMSE}(\hat{x}_{p_n}^{\mathrm{L}})$, $\mathrm{AMSE}(\hat{x}_{p_n}^{\mathrm{LG(1)}})$ and $\mathrm{AMSE}(\hat{x}_{p_n}^{\mathrm{LG(2)}})$ are convex. The geometric shape of these graphs is similar to the ones in Figure 2. The AMSE of all the estimators $\uparrow$ as the sample size $n \uparrow$ since the estimation of extreme quantiles of an order $(1 - 1/(5n))$ is more difficult when $n \uparrow$.

For the Pareto distribution, AMSE of all the estimators $\downarrow$ as $k_n \uparrow$ and, when the sample size $n \uparrow$ these AMSE are approximately similar for large values of k_n. This can be explained by the fact that there is no asymptotic bias. For this distribution, $\mathrm{AMSE}(\hat{x}_{p_n}^{\mathrm{WG}})$ and $\mathrm{AMSE}(\hat{x}_{p_n}^{\mathrm{W}})$ are approximately equal whatever k_n and n. Moreover, $\mathrm{AMSE}(\hat{x}_{p_n}^{\mathrm{LG(1)}})$ seems to be higher than the one of its competing estimators for the small sample sizes $n \leq 100$.

Unlike the Pareto distribution, for the Student distribution AMSE of all the estimators $\uparrow$ as $k_n \uparrow$. Moreover from a fraction sample k_n not too small, $\mathrm{AMSE}(\hat{x}_{p_n}^{\mathrm{W}})$ are clearly higher than $\mathrm{AMSE}(\hat{x}_{p_n}^{\mathrm{WG}})$ which is in turn higher than $\mathrm{AMSE}(\hat{x}_{p_n}^{\mathrm{L}})$ which is finally itself higher than $\mathrm{AMSE}(\hat{x}_{p_n}^{\mathrm{LG(1)}})$ and $\mathrm{AMSE}(\hat{x}_{p_n}^{\mathrm{LG(2)}})$. The two latter AMSE are approximately equal whatever k_n and n.

Regarding the Fréchet and Burr distributions, in general $\mathrm{AMSE}(\hat{x}_{p_n}^{\mathrm{W}})$ is higher than $\mathrm{AMSE}(\hat{x}_{p_n}^{\mathrm{WG}})$, $\mathrm{AMSE}(\hat{x}_{p_n}^{\mathrm{L}})$ and $\mathrm{AMSE}(\hat{x}_{p_n}^{\mathrm{LG(2)}})$ whatever the sample size. For small values of the fraction sample, $\mathrm{AMSE}(\hat{x}_{p_n}^{\mathrm{W}})$ is smaller than $\mathrm{AMSE}(\hat{x}_{p_n}^{\mathrm{LG(1)}})$ and for large values of k_n the opposite occurs, *i.e.* $\mathrm{AMSE}(\hat{x}_{p_n}^{\mathrm{LG(1)}}) < \mathrm{AMSE}(\hat{x}_{p_n}^{\mathrm{W}})$. Once the function AMSE reaches its minimum, we observe that : **(1)** $\mathrm{AMSE}(\hat{x}_{p_n}^{\mathrm{LG(1)}})$ and $\mathrm{AMSE}(\hat{x}_{p_n}^{\mathrm{LG(2)}}) \uparrow$ slowly as $k_n \uparrow$; **(2)** $\mathrm{AMSE}(\hat{x}_{p_n}^{\mathrm{WG}})$ and $\mathrm{AMSE}(\hat{x}_{p_n}^{\mathrm{L}})$ $\uparrow$ slightly faster as $k_n \uparrow$; **(3)** $\mathrm{AMSE}(\hat{x}_{p_n}^{\mathrm{W}}) \uparrow$ very faster as $k_n \uparrow$. When the sample size $n \uparrow$, the difference between $\mathrm{AMSE}(\hat{x}_{p_n}^{\mathrm{LG(1)}})$ and $\mathrm{AMSE}(\hat{x}_{p_n}^{\mathrm{LG(2)}}) \downarrow$ as $k_n \uparrow$.

As by definition, AMSE is equal to the sum of the variance and squared bias of the estimator, *i.e.*

$$\mathrm{AMSE}(\hat{x}_{p_n}^{\bullet}) = \mathrm{Avar}(\hat{x}_{p_n}^{\bullet}) + \mathrm{ABias}^2(\hat{x}_{p_n}^{\bullet}), \tag{22}$$

where letter “A” at the beginning of the notation refers to “asymptotic”, Figure 4 suggests the following interpretations :

- The variance of estimators $\hat{x}_{p_n}^{\mathrm{WG}}$, $\hat{x}_{p_n}^{\mathrm{L}}$, $\hat{x}_{p_n}^{\mathrm{LG(1)}}$ and $\hat{x}_{p_n}^{\mathrm{LG(2)}}$ seems smaller than the variance of $\hat{x}_{p_n}^{\mathrm{W}}$. The behaviour of the median estimators of $\hat{x}_{p_n}^{\mathrm{WG}}$, $\hat{x}_{p_n}^{\mathrm{L}}$, $\hat{x}_{p_n}^{\mathrm{LG(1)}}$ and $\hat{x}_{p_n}^{\mathrm{LG(2)}}$ in Figures 3 tend to confirm these statements. They are more stable than $\hat{x}^{\mathrm{W}}_{p_n}$. Notice that the variance of $\hat{x}^{\mathrm{W}}_{p_n}$ can be approximated by $\frac{\gamma^2}{k_n}\left(1+\log^2\left(\frac{k_n}{np_n}\right)\right)$ (see *e.g.* Beirlant et al., 2004, p. 120).

- The standard deviation of the proposed estimators may be negligible compared to their bias, *i.e.* $\mathrm{Avar}^{1/2}(\hat{x}^{\bullet}_{p_n}) \ll \mathrm{ABias}(\hat{x}^{\bullet}_{p_n})$. Thus, since the bias of estimators $\hat{x}^{\mathrm{WG}}_{p_n}$, $\hat{x}^{\mathrm{L}}_{p_n}$, $\hat{x}^{\mathrm{LG}(1)}_{p_n}$ and $\hat{x}^{\mathrm{LG}(2)}_{p_n}$ are smaller than the bias of Weissman estimator $\hat{x}^{\mathrm{W}}_{p_n}$ at a scale factor to be determined, then $\mathrm{AMSE}(\hat{x}^{\mathrm{W}}_{p_n})$ is larger than $\mathrm{AMSE}(\hat{x}^{\mathrm{WG}}_{p_n})$, $\mathrm{AMSE}(\hat{x}^{\mathrm{L}}_{p_n})$, $\mathrm{AMSE}(\hat{x}^{\mathrm{LG}(2)}_{p_n})$ and, from a sample fraction k_n not too small $\mathrm{AMSE}(\hat{x}^{\mathrm{WG}}_{p_n}) > \mathrm{AMSE}(\hat{x}^{\mathrm{LG}(1)}_{p_n})$.

Choice of the optimal sample fraction

The proposed estimators depend on the fraction sample k_n. Basically, the direct minimization of the AMSE errors can be used as a criterion to select k_n. However, this method can not be considered in practice since the AMSE is unknown. A number of methods for the selection of sample fraction k_n can be found in Beirlant et al. (1996b); Drees and Kaufmann (1998); Guillou and Hall (2001); Gomes and Oliveira (2001). Another option consists in choosing k_n corresponding to the range of stability of the estimators with respect to the fraction sample. In this study, one proposes to choose the largest integer k_n which minimizes a dissimilarity measure between the four estimators $\hat{x}^{\mathrm{W}}_{p_n}$, $\hat{x}^{\mathrm{WG}}_{p_n}$, $\hat{x}^{\mathrm{L}}_{p_n}$ and $\hat{x}^{\mathrm{LG}(2)}_{p_n}$, *i.e.*

$$\begin{aligned}\hat{k}_n = \underset{k_n=1,\dots,n-1}{\arg\min} & \left\{\left|\hat{x}^{\mathrm{W}}_{p_n}-\hat{x}^{\mathrm{WG}}_{p_n}\right|+\left|\hat{x}^{\mathrm{W}}_{p_n}-\hat{x}^{\mathrm{L}}_{p_n}\right|+\left|\hat{x}^{\mathrm{W}}_{p_n}-\hat{x}^{\mathrm{LG}(2)}_{p_n}\right|\right.\\ & \left.+\left|\hat{x}^{\mathrm{WG}}_{p_n}-\hat{x}^{\mathrm{L}}_{p_n}\right|+\left|\hat{x}^{\mathrm{WG}}_{p_n}-\hat{x}^{\mathrm{LG}(2)}_{p_n}\right|+\left|\hat{x}^{\mathrm{L}}_{p_n}-\hat{x}^{\mathrm{LG}(2)}_{p_n}\right|\right\}. \qquad (23)\end{aligned}$$

This heuristic is used in non-parametric estimation. It relies on the idea that, if $\hat{k}_n$ is properly chosen, all estimates should approximately give the same value. We refer to Gardes et al. (2010) for an illustration of this procedure on simulated data. In addition, we illustrated, in Figures 5 and 6, the dissimilarity procedure on the median estimators for $N = 1000$ simulated samples from the Fréchet and Burr distributions respectively. In both Figures, the selected $\hat{k}_n$ produce good results. Nevertheless, when selecting k_n independently for each estimator, better results may be produced as it is the case for instance$\hat{x}^{\mathrm{L}}_{p_n}$ in Figure 5-a and $\hat{x}^{\mathrm{W}}_{p_n}$ in Figure 5-d. In the other Figures, the dissimilarity procedure performs as well as selecting k_n independently for each estimator by minimization of the error.

A brief summary

To summarize, these numerical experiments confirm that, for a large enough fraction sample k_n and large simple size ($n > 100$), $\hat{x}_{p_n}^{\mathrm{LG}(1)} \simeq \hat{x}_{p_n}^{\mathrm{LG}(2)}$ which means that it is reasonable to fix $f_{k_n} = 1$. However, they show that the choice $f_{k_n} = 1$ is not optimal since $\hat{x}_{p_n}^{\mathrm{LG}(2)}$ is better than $\hat{x}_{p_n}^{\mathrm{LG}(1)}$ in almost all cases, especially when $n \leq 100$. Finally, despite the fact that we know there is no optimal estimator for all cases, the simulations confirm that estimators $\hat{x}_{p_n}^{\mathrm{WG}}$, $\hat{x}_{p_n}^{\mathrm{L}}$ and $\hat{x}_{p_n}^{\mathrm{LG}(2)}$ are better than the Weissman estimator $\hat{x}_{p_n}^{\mathrm{W}}$ especially for the bias and the AMSE for the distributions where the function $\ell(.)$ is not constant. The performance of all estimators are approximately equal when $\ell(.)$ is the constant.

5 Case study : estimation of high flood return period

In this section, we adapt and apply the proposed estimators to flood events. As illustrated in 7, a flood event is mainly described with three variables obtained from a typical flood hydrograph. These variables are the flood peak (Q), flood volume (V) and flood duration (D).

The data set used in this case study is taken from Yue et al. (1999) and consists in daily natural streamflow measurements from the Ashuapmushuan basin (reference number 061901). The gauging station, located in the province of Quebec (Canada) is near the outlet of the basin, at latitude $48.69°$N and longitude $72.49°$W. In this region, floods are generally caused by high spring snowmelt. Data are available from 1963 to 1995. The flood annual observations of flood peaks, durations and volumes were extracted from a daily streamflow data set.

The proposed estimators of extreme quantiles are built by assuming that the CDF is heavy tailed. An exploratory study is performed using the PQ-plot in (6) and the GQ-plot in (7). Figures 8-a and 8-b illustrates respectively the PQ-plots and GQ-plots corresponding to three variables characterising the flood event. These plots show that the flood peak and the flood volume belong to the MDA(Fréchet). Indeed, for extreme points, the PQ-plots in Figure 8-(iii, v) seem to be approximatively linear and the GQ-plots in Figure 8-(iv, v) reveal a positive slope. On the other hand, the duration is not heavy-tailed since the curves of its PQ-plot in Figure 8-(i) and GP-plot in Figure 8-(ii) are approximately constant for extremes points. Thus, we are only interested in estimating of peak and volume. We considered the return period $T \in \{66, 99, 132, 165\}$ years according to the sample size $n = 33$. Mathematically, the problem is to estimate the quantile of order

$$(1-p) \in \{0.9848485, 0.989899, 0.9924242, 0.9939394\}.$$

For each T, the extreme quantile is estimated with $\hat{x}_p^{\mathrm{W}}$, $\hat{x}_p^{\mathrm{L}}$, $\hat{x}_p^{\mathrm{WG}}$ and $\hat{x}_p^{\mathrm{LG}(2)}$. The fraction sample on which the estimation is based was chosen by using criterion (23). For each value of T, for each of the two selected variables (V and Q), we compute the mean and the standard deviation (stdev) of the estimators. The estimated peaks and volumes are presented, with their computed mean and standard deviation, in Table 2 and Table 3 respectively.

Unlike the stdev of the estimated volumes Table 3, we notice that the stdev of the estimated peaks in Table 2 do not $\uparrow$ too fast as the return period $T\uparrow$. Also, stdev is large for the estimated volumes. Thus, for this case study, the estimate of volume V deteriorates faster than the estimate of the peak as $T\uparrow$. The estimation remains more stable when the extreme quantile is not too far from the boundary of the sample, *i.e.* for a reasonable value of the return period T. Indeed, estimation errors increase with the return period.

Figure 9 illustrates the selected fraction sample k_n and the estimators associated to each one of the considered variables Q and V for the return periods $T = 66$ and $T = 165$ years. For both variables of interest, we observe that the estimators $\hat{x}_p^{\mathrm{L}}$, $\hat{x}_p^{\mathrm{WG}}$ and $\hat{x}_p^{\mathrm{LG}(2)}$ are smooth and more stable compared to $\hat{x}_p^{\mathrm{W}}$. In addition, the difference between $\hat{x}_p^{\mathrm{W}}$ and the three other estimators $\uparrow$ as the fraction sample $k_n\uparrow$. This indicates a high bias for large values of k_n.

For Q series, criterion (23) suggests $\hat{k}_n = 16$ respectively for $T = 66$ and $T = 165$ years. Nevertheless, Figures 9-(a, b) show that we can choose $\hat{k}_n$ in the set $\{6, \ldots, 16\}$ where the four estimators seem to have similar values. Moreover, for the estimator $\hat{x}_p^{\mathrm{L}}$, Figures 9-(a, b) indicate that $\hat{k}_n$ can also be larger than 16 since this estimator is less sensitive to the selected k_n. $\hat{x}_p^{\mathrm{WG}}$ have a large volatility and for $k_n > 16$ the difference between this estimator and the other ones becomes important. Taking $k_n > 16$ could lead to an overestimation of the extreme quantiles.

Regarding the series of V, criterion (23) indicates that $\hat{k}_n = 8$ is a good choice for $T = 66$ and $T = 165$ years. In Figures 9-(c, d), the observation of the range of stability of the four estimators with respect to the fraction sample shows that $\hat{k}_n$ could be reasonably estimated in $\{5, \ldots, 10\}$. Figures 9-(c, d) confirm that $\hat{x}_p^{\mathrm{L}}$, $\hat{x}_p^{\mathrm{WG}}$, $\hat{x}_p^{\mathrm{LG}(2)}$ are smooth and less sensitive than $\hat{x}_p^{\mathrm{W}}$. Figure 9-(d) shows that one can build the estimator $\hat{x}_p^{\mathrm{L}}$ not only with the k_n largest observations but also with the entire sample, *i.e.* $k_n = n$.

Even through the estimator values in Tables 2 and 3 are relatively similar, Figure 9 indicates that $\hat{x}_p^{\mathrm{W}}$ is very sensitive to k_n. Therefore, a bad choice of k_n could lead to very different estimator values whereas the other proposed estimators have a very small volatility with respect to k_n. Despite the fact that all the estimators are similar for a reasonable choice of k_n, the results of the case study suggest that it is advantageous to estimate extreme quantiles with $\hat{x}_p^{\mathrm{WG}}$, $\hat{x}_p^{\mathrm{LG(2)}}$ and $\hat{x}_p^{\mathrm{L}}$ instead of $\hat{x}_p^{\mathrm{W}}$. The case study results confirm the findings of the simulation study, in particular the stability of the proposed estimators with respect to k_n.

6 Conclusions

The present paper introduced (1) the geometric estimators of extreme quantiles and (2) a "weighted" estimator of quantiles for high return periods $T \geq 2/(n\mathrm{e})$ where n is the sample size. Simulation results show that the proposed estimators given in (14), (19) and (20) are smooth and more stable than the Weissman estimator (11). In addition, they improve the bias. Since the accuracy of estimators depends on the precise choice of the number of order statistics k_n, a method of selection of k_n is proposed and illustrated in the case study. The case study shows that $\hat{x}_p^{\mathrm{W}}$ is very sensitive to the selected k_n which is not the case of the proposed estimators. Given the good performance of estimators (14), (19) and (20), we propose to explicit in future work, their asymptotic distributions. More precisely, we propose to study asymptotic properties of the proposed estimators under less restrictive conditions than those in Lekina (2010). This statistical result will allow, for instance, to build more accurate estimation confidence intervals. In other respects, this result would allow to validate the behaviour of the observed AMSE in the simulations and to identify the most efficient estimator. Finally, despite the fact that in EVA, it is often recommended to consider at the same time several estimators of extreme quantiles since there is no optimal estimator for all cases, according to the simulation results on simulated data in the present paper, we suggest to use estimateur $\hat{x}_p^{\mathrm{LG(2)}}$. Numerical experiments indicate that its AMSE is smaller than the one of its competitors especially for the small samples *i.e.* $n \leq 100$.

Acknowledgements

Financial support for this study was graciously provided by the Natural Sciences and Engineering Research Council (NSERC) of Canada and the Canada Research Chair Program. The authors wish to express their appreciation to the reviewers and the Editor-in-Chief for their invaluable comments and suggestions.

References

Adamowski, K., Liang, G.-C., and Patry, G. G. (1998). Annual maxima and partial duration flood series analysis by parametric and non-parametric methods. *Hydrological Processes*, 12(10-11):1685–1699.

Adlouni, S. E., Bobée, B., and Ouarda, T. (2008). On the tails of extreme event distributions in hydrology. *Journal of Hydrology*, 355(1-4):16–33.

Apipattanavis, S., Rajagopalan, B., and Lall, U. (2010). Local polynomial-based flood frequency estimator for mixed population. *Journal of Hydrologic Engineering*, 15(9):680–691.

Balkema, A. and de Haan, L. (1974). Residual life time at a great age. *Annals of Probability*, 2(5):792–804.

Beirlant, J., Broniatowski, M., Teugels, J. L., and Vynckier, P. (1995). The mean residual life function at great age: Applications to tail estimation. *Journal of Statistical Planning and Inference*, 45(1-2):21–48.

Beirlant, J., Dierckx, G., Guilllou, A., and Stărică, C. (2002). On exponential representations of log-spacings of extreme order statistics. *Extremes*, 5(2):157–180.

Beirlant, J., Dierckx, G., and Guillou, A. (2005). Estimation of the extreme value index and regression on generalized quantile plots. *Annals of Statistics*, 11(6):949–970.

Beirlant, J., Goegebeur, Y., Teugels, J., and Segers, J. (2004). *Statistics of extremes: theory and applications*. Wiley Series in Probability and Statistics. John Wiley and Sons.

Beirlant, J., Teugels, J., and Vynckier, P. (1996a). *Practical analysis of extreme values*. Leuven University Press.

Beirlant, J., Vynckier, P., and Teugels, J. (1996b). Excess functions and estimation of the extreme value index. *Bernoulli*, 2(4):293–318.

Bingham, N. H., Goldie, C. M., and Teugels, J. L. (1987). *Regular Variation*, volume 27. Encyclopedia of Mathematics and its Applications, Cambridge University Press.

Bobée, B., Cavadias, G., Ashkar, F., Bernier, J., and Rasmussen, P. (1993). Towards a systematic approach to comparing distributions used in flood frequency analysis. *Journal of Hydrology*, 142:121–136.

Breiman, L., Stone, C. J., and Kooperberg, C. (1990). Robust confidence bounds for extreme upper quantiles. *Journal of Statistical Computation and Simulation*, 37(3–4):127–149.

Brunet-Moret, Y. (1969). Étude de quelques lois statistiques utilisées en hydrologie. *Cahiers d'hydrologie*, 6(3).

Caeiro, F. and Gomes, M. (2006). A new class of estimators of a "scale" second order parameter. *Extremes*, 9(3-4):193–211.

Caeiro, F., Gomes, M., and Rodrigues, L. (2009). Reduced-bias tail index estimators under a third-order framework. *Communications in Statistics - Theory and Methods*, 38(7):1019–1040.

Chebana, F., Adlouni, S., and Bobée, B. (2010). Mixed estimation methods for halphen distributions with applications in extreme hydrologic events. *Stochastic Environmental Research and Risk Assessment*, 24(3):359–376.

Chen, Y., Xu, S., Sha, Z., Pieter, V. G., and Sheng-Hua, G. (2004). Study on L-moment estimations for log-normal distribution with historical flood data. *International Association of Hydrological Sciences*, (289):107–113.

Coles, S. (2001). *An introduction to statistical modeling of extreme values*. Springer series in statistics. Springer, 1 edition.

Daouia, A., Gardes, L., Girard, S., and Lekina, A. (2011). Kernel estimators of extreme level curves. *Test*, 20(2):311–333.

de Haan, L. (1984). Slow variation and characterization of domains of attraction. In J. Tiago de Oliveira, editor, *Statistical Extremes and Applications*, pages 31–48. Reidel, Dorchrecht.

de Haan, L. and Ferreira, A. (2006). *Extreme Value Theory: An Introduction*. Springer Series in Operations Research and Financial Engineering.

de Haan, L. and Peng, L. (1998). Comparison of tail index estimators. *Statistica Neerlandica*, 52(1):60–70.

de Wet, T., Goegebeur, Y., and Guillou, A. (2012). Weighted moment estimators for the second order scale parameter. *Methodology and Computing in Applied Probability*, 14:753–783.

Dekkers, A. and de Haan, L. (1989). On the estimation of the extreme value index and large quantile estimation. *Annals of Statistics*, 17(4):1795–1832.

Diebolt, J., Gardes, L., Girard, S., and Guillou, A. (2008). Bias-reduced extreme quantiles estimators of Weibull distributions. *Journal of Statistical Planning and Inference*, 138(5):1389–1401.

Dierckx, G., Beirlant, J., Waal, D. D., and Guillou, A. (2009). A new estimation method for Weibull-type tails based on the mean excess function. *Journal of Statistical Planning and Inference*, 139(6):1905–1920.

Ditlevsen, O. (1994). Distribution arbitrariness in structural reliability. In G. Schueller, M. Shinozuka, J. Y., editor, *6th International Conference on Structural Safety and Reliability*, pages 1241–1247. Balkema, Rotterdam.

Drees, H. (1995). Refined Pickands estimator of the extreme value index. *Annals of Statistics*, 23(6):2059–2080.

Drees, H. and Kaufmann, E. (1998). Selecting the optimal sample fraction in univariate extreme value estimation. *Stochastic Processes and their Applications*, 75(2):149–172.

Embrechts, P., Klüppelberg, C., and Mikosch, T. (1997). *Modelling Extremal Events for Insurance and Finance*. Springer Verlag.

Falk, M. (1995). On testing the extreme value index via the Pot-method. *Annals of Statistics*, 23(6):2013–2035.

Feuerverger, A. and Hall, P. (1999). Estimating a tail exponent by modelling departure from a Pareto distribution. *Annals of Statistics*, 27(2):760–781.

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

Gardes, L., Girard, S., and Lekina, A. (2010). Functional nonparametric estimation of conditional extreme quantiles. *Journal of Multivariate Analysis*, 101(2):419–433.

Girard, A., Guillou, A., and Stupfler, G. (2012). Estimating an endpoint with high order moments. *Test*. To appear.

Gnedenko, B. (1943). Sur la distribution limite du terme maximum d'une série aléatoire. *Annals of Mathematics*, 44(3):423–453.

Goegebeur, Y., Beirlant, J., and de Wet, T. (2010). Kernel estimators for the second order parameter in extreme value statistics. *Journal of Statistical Planning and Inference*, 140(9):2632–2652.

Gomes, M. I. and Oliveira, O. (2001). The bootstrap methodology in statistics of extremes: theory and applications - choice of the optimal sample fraction. *Extremes*, 4(4):331–358.

Guida, M. and Longo, M. (1988). Estimation of probability tails based on generalized extreme value distributions. *Reliability Engineering and System Safety*, 20(3):219–242.

Guillou, A. and Hall, P. (2001). A diagnostic for selecting the threshold in extreme value analysis. *Journal of the Royal Statistical Society, Series B*, 63(2):293–305.

Haddad, K. and Rahman, A. (2011). Selection of the best fit flood frequency distribution and parameter estimation procedure: a case study for Tasmania in Australia. *Stochastic Environmental Research and Risk Assessment*, 25(3):415–428.

Haeusler, E. and Teugels, J. (1985). On asymptotic normality of Hill's estimator for the exponent of regular variation. *Annals of Statistics*, 13(2):743–756.

Hall, P. and Park, B. U. (2002). New methods for bias correction at endpoints and boundaries. *Annals of Statistics*, 30(5):1460–1479.

Hill, B. (1975). A simple general approach to inference about the tail of a distribution. *Annals of Statistics*, 3(5):1163–1174.

Hosking, J. R. M. and Wallis, J. R. (1987). Parameter and quantile estimation for the generalized Pareto distribution. *Technometrics*, 29(3):339–1349.

Hosking, J. R. M., Wallis, J. R., and Wood, E. F. (1985). Estimation of the generalized extreme-value distribution by the method of probability-weighted comments. *Technometrics*, 27(3):251–261.

Kratz, M. and Resnick, S. (1996). The QQ-estimator and heavy tails. *Stochastic Models*, 12(4):699–724.

Lall, U., il Moon, Y., and Bosworth, K. (1993). Kernel flood frequency estimators: Bandwidth selection and kernel choice. *Water Resources Research*, 29(4).

Lang, M., Ouarda, T., and Bobée, B. (1999). Towards operational guidelines for over-threshold modeling. *Journal of Hydrology*, 225(3–4):103–117.

Lekina, A. (2010). *Estimation non-paramétrique des quantiles extrêmes conditionnels*. PhD thesis, Université de Grenoble.

Li, D. and Peng, L. (2009). Does bias reduction with external estimator of second order parameter work for endpoint? *Journal of Statistical Planning and Inference*, 139(6):1937–1952.

Moon, Y.-I. and Lall, U. (1994). Kernel quantite function estimator for flood frequency analysis. *Water Resources Research*, 30(11).

Ouarda, T. B., Girard, C., Cavadias, G. S., and Bobée, B. (2001). Regional flood frequency estimation with canonical correlation analysis. *Journal of Hydrology*, 254(1–4):157–173.

Pickands, J. (1975). Statistical inference using extreme order statistics. *Annals of Statistics*, 3(1):119–131.

Prescott, P. and Walden, A. T. (1980). Maximum likelihood estimation of the parameters of generalized extreme-value distribution. *Biometrika*, 67(3):723–724.

Quintela-del-Río, A. and Francisco-Fernández, M. (2011). Analysis of high level ozone concentrations using nonparametric methods. *Science of The Total Environment*, 409(6):1123–1133.

Rényi, A. (1953). On the theory of order statistics. *Acta Mathematica Hungarica*, 4(3–4):191–231.

Rosen, O. and Weissman, I. (1996). Comparison of estimation methods in extreme value theory. *Communication in Statistics-Theory and Methods*, 24(4):759–773.

Salvadori, G., De Michele, C., Kottegoda, N. T., and Rosso, R. (2007). *Extremes in Nature: An Approach Using Copulas*. Springer.

Schultze, J. and Steinebach, J. (1996). On least squares estimates of an exponential tail coefficient. *Statistics and Decisions*, 14(3):353–372.

Smith, J. (1987). Estimating the upper tail of flood frequency distributions. *Water Resources Research*, 23(8):1657–1666.

Smith, R. L. (1985). Maximum likelihood estimation in a class of nonregular cases. *Biometrika*, 72(1):67–92.

Smith, R. L. (1986). Extreme value theory based on the r largest annual events. *Journal of Hydrology*, 86(1-2):27 – 43.

Stedinger, J. R. (2000). Flood frequency analysis and statistical estimation of flood risk. In *Inland Flood Hazards : Human, Riparian and Aquatic Communities*, chapter 12, pages 334–358.

Tsourti, Z. and Panaretos, J. (2001). A simulation study on the performance of extreme-value index estimators and proposed robustifying modifications. *5th Hellenic European Conference on Computer Mathematics and its Applications, Athens, Greece*, 2:847–852.

Weissman, I. (1978). Estimation of parameters and large quantiles based on the k-largest observations. *Journal of the American Statistical Association*, 73(364):812–815.

Willems, P., Guillou, A., and Beirlant, J. (2007). Bias correction in hydrologic GPD based extreme value analysis by means of a slowly varying function. *Journal of Hydrology*, 338(3–4):221–236.

Young-Il, M., Lall, U., and Bosworth, K. (1993). A comparison of tail probability estimators for flood frequency analysis. *Journal of Hydrology*, 151(2-4):343 – 363.

Yue, S., Ouarda, T., Bobée, B., Legendre, P., and Bruneau, P. (1999). The Gumbel mixed model for flood frequency analysis. *Journal of Hydrology*, 226(1-2):88–100.

Distribution	Density	Sequences
Normal	$f(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right)$	$a_n = (2\log n)^{-1/2}$
	$x \in \mathbb{R}$	$b_n = (2\log n)^{1/2} - \frac{\log\log n + \log 4\pi}{2\,(2\log n)^{1/2}}$
Exponential	$f(x) = \lambda \exp(-\lambda x)$	$a_n = 1/\lambda$
	$x \geq 0$	$b_n = \log(n)/\lambda$
Cauchy	$f(x) = \frac{1}{\pi} \frac{1}{1+x^2}$	$a_n = 0$
	$x \in \mathbb{R}$	$b_n = n/\pi$
Beta	$f(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}$	$a_n = \left(n\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b+1)}\right)^{-1/b}$
	$0 < x < 1,\ a, b > 0$	$b_n = 1$

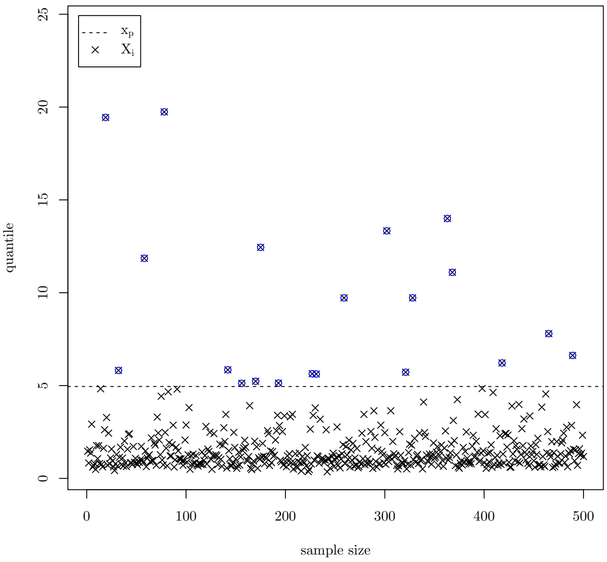
Table 1: Limited number of examples of the theoretical normalized sequences a_n et b_n.

Estimator \ Return period T	66	99	132	165
$\hat{x}^{\mathrm{W}}_{1/T}$	2435.00	2583.10	2693.62	2782.58
$\hat{x}^{\mathrm{L}}_{1/T}$	2456.14	2607.50	2720.55	2811.61
$\hat{x}^{\mathrm{WG}}_{1/T}$	2433.15	2583.25	2695.34	2785.62
$\hat{x}^{\mathrm{LG(2)}}_{1/T}$	2432.61	2584.59	2698.14	2789.64
mean	2439.45	2590.01	2702.45	2793.02
stdev	11.22	11.68	12.13	12.55

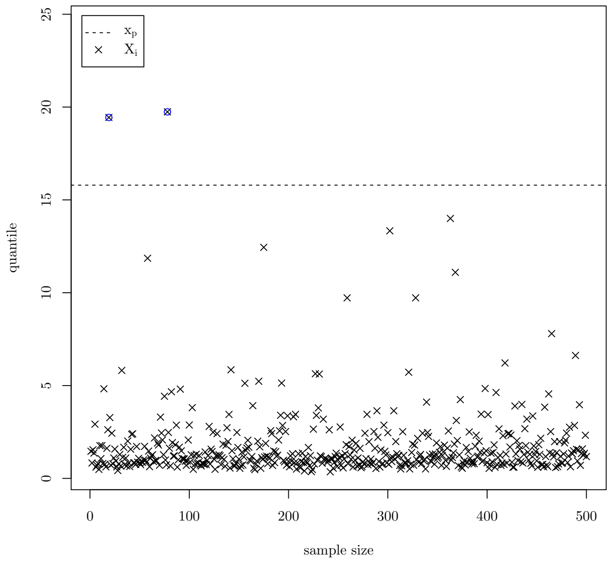
Table 2: Estimated flood peak Q.

Estimator \ Return period T	66	99	132	165
$\hat{x}^{\mathrm{W}}_{1/T}$	84979.31	89238.64	92389.53	94909.97
$\hat{x}^{\mathrm{W}}_{1/T}$	84267.36	88418.86	91485.84	93937.06
$\hat{x}^{\mathrm{WG}}_{1/T}$	84970.60	89146.48	92233.20	94700.84
$\hat{x}^{\mathrm{LG(2)}}_{1/T}$	84761.40	88953.95	92053.78	94532.40
mean	84957.27	89158.70	92264.57	94747.80
stdev	652.10	708.26	751.46	786.86

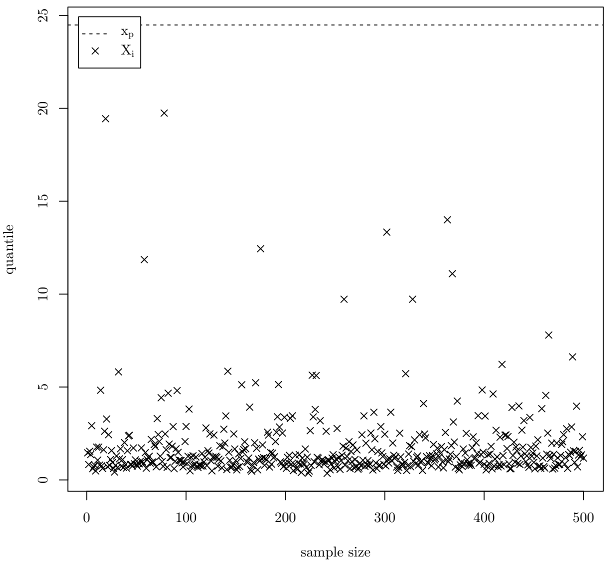
Table 3: Estimated flood volume V.



(a) $T = 25$



(b) $T = 250$



(c) $T = 600$

Figure 1: Difference between large quantiles within and outside the sample. Scatter plot of the Fréchet distributed sample $\{X_i,\ i = 1, \ldots, 500\}$ $(\times \times \times)$ with tail index $\gamma = 0.5$, location parameter $m = 0$ and scale parameter $s = 1$, the extreme quantile x_p $(- \; - \; -)$ and observations higher than x_p $(\otimes \otimes \otimes)$ with $p = 1/T$, for (a) $T = 25$, (b) $T = 250$ and (c) $T = 600$.

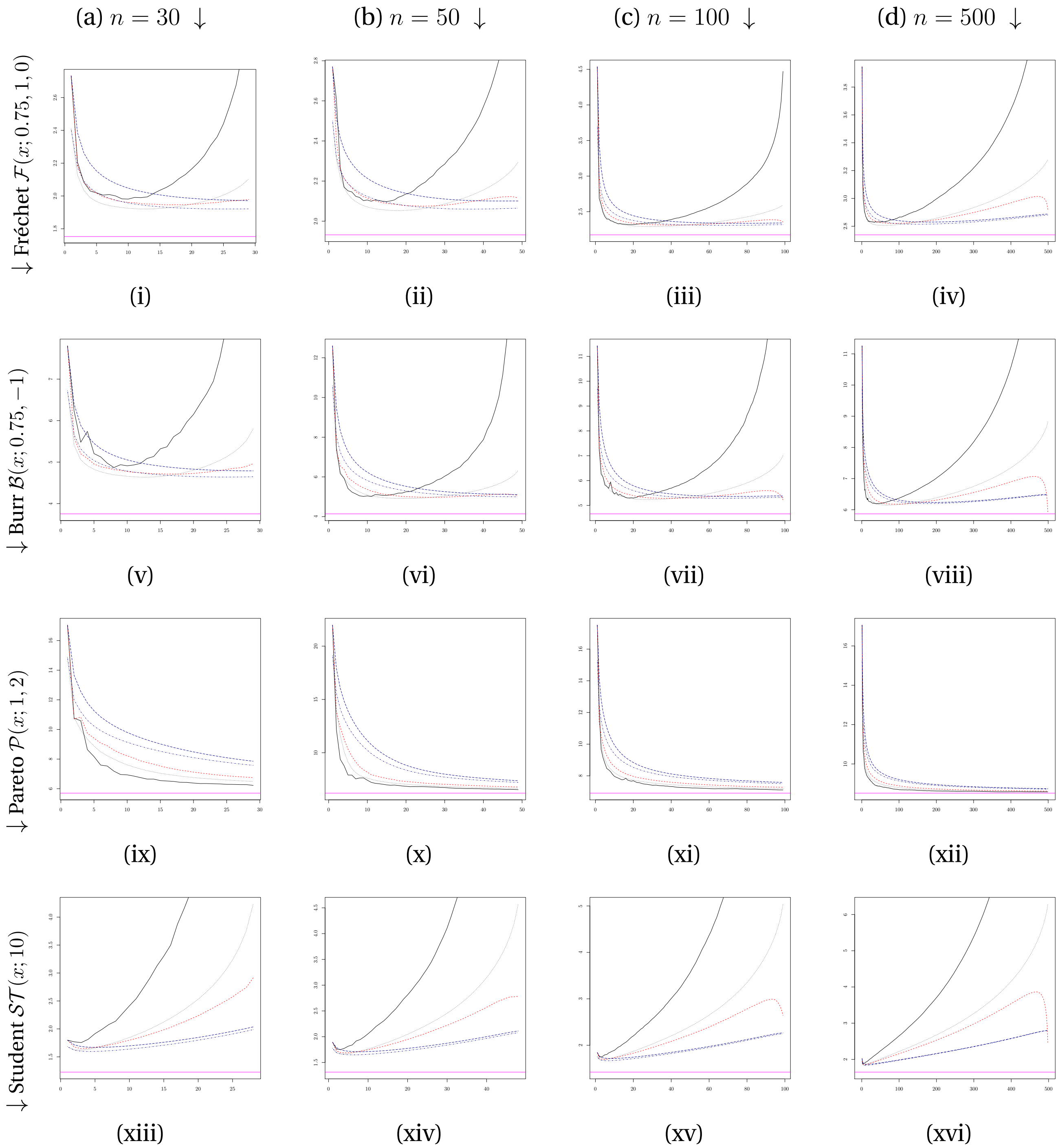


Figure 2: Mean estimators of $\log \hat{x}^{\mathrm{W}}_{p_n}$ (——), $\log \hat{x}^{\mathrm{WG}}_{p_n}$ $(\cdots\cdots)$, $\log \hat{x}^{\mathrm{L}}_{p_n}$ $(- - -)$, $\log \hat{x}^{\mathrm{LG}(1)}_{p_n}$ (— — —) and $\log \hat{x}^{\mathrm{LG}(2)}_{p_n}$ $(- \cdot -)$ for $N = 1000$ simulated samples of size $n \in \{30, 50, 100, 500\}$ from the distributions of Fréchet (i)–(iv), Burr (v)–(viii), Pareto (ix)–(xii) and Student (xiii)–(xvi). The horizontal line indicates the true value of log-quantile, *i.e.* $\log x_{p_n}$. The horizontal axis corresponds to the fraction sample $k_n = 1, \ldots, n-1$.

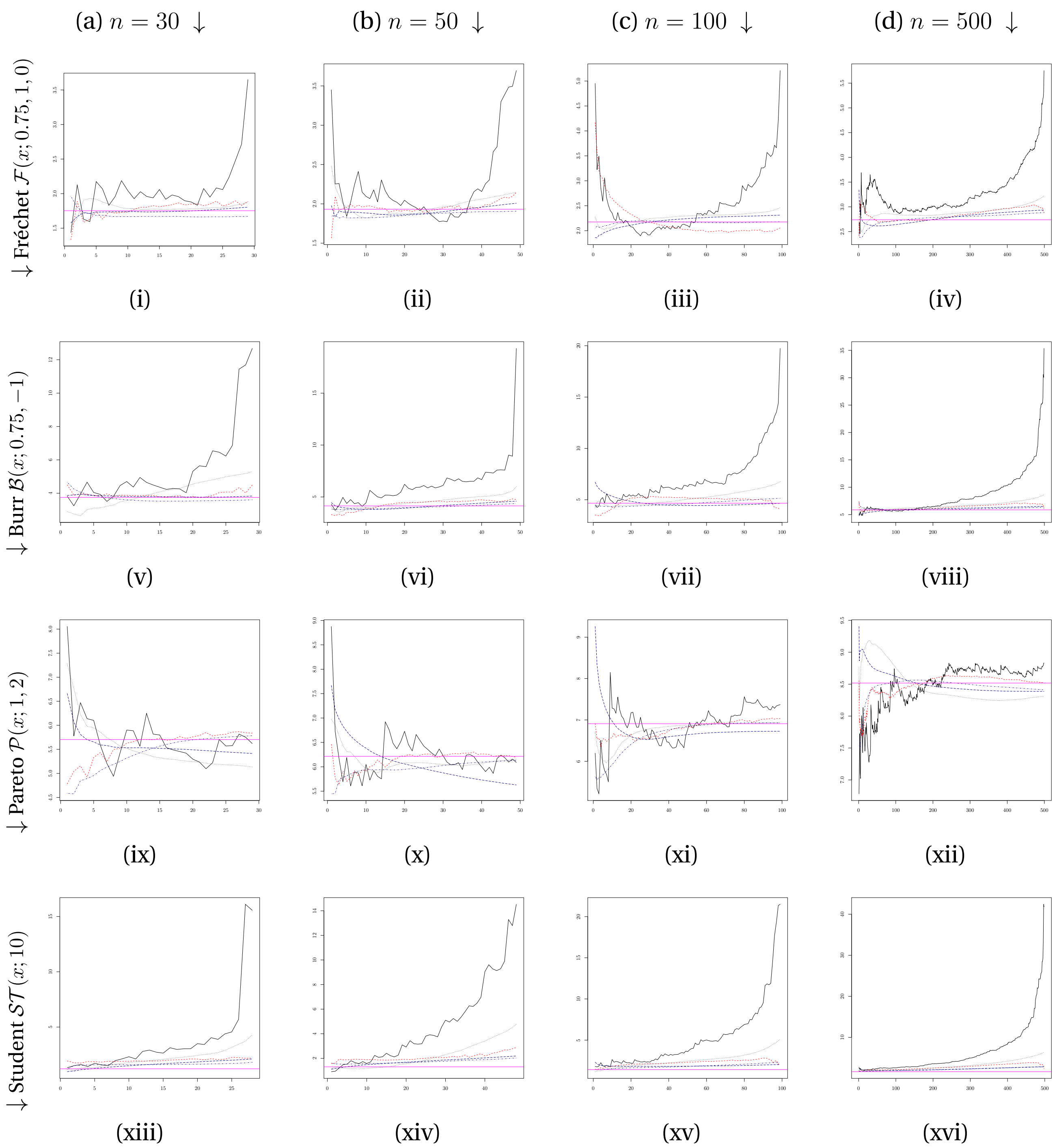


Figure 3: Median estimators of $\log \hat{x}^{\mathrm{W}}_{p_n}$ (——), $\log \hat{x}^{\mathrm{WG}}_{p_n}$ $(\cdots\cdots)$, $\log \hat{x}^{\mathrm{L}}_{p_n}$ $(- - -)$, $\log \hat{x}^{\mathrm{LG(1)}}_{p_n}$ (— — —) and $\log \hat{x}^{\mathrm{LG(2)}}_{p_n}$ $(- \cdot -)$ for $N = 1000$ simulated samples of size $n \in \{30, 50, 100, 500\}$ from the distributions of Fréchet (i)–(iv), Burr (v)–(viii), Pareto (ix)–(xii) and Student (xiii)–(xvi). The horizontal line indicates the true value of log-quantile, *i.e.* $\log x_{p_n}$. The horizontal axis corresponds to the fraction sample $k_n = 1, \ldots, n-1$.

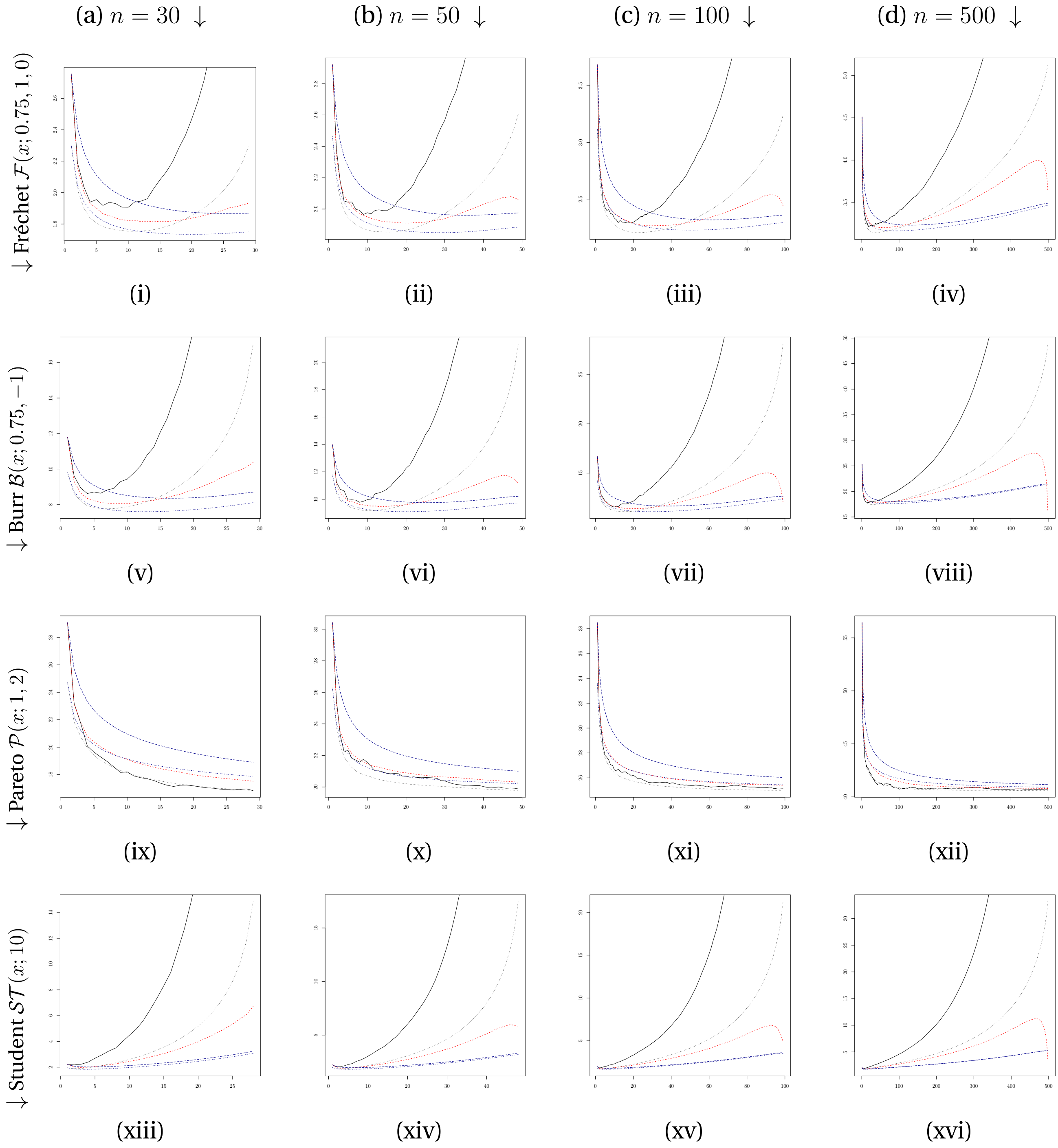
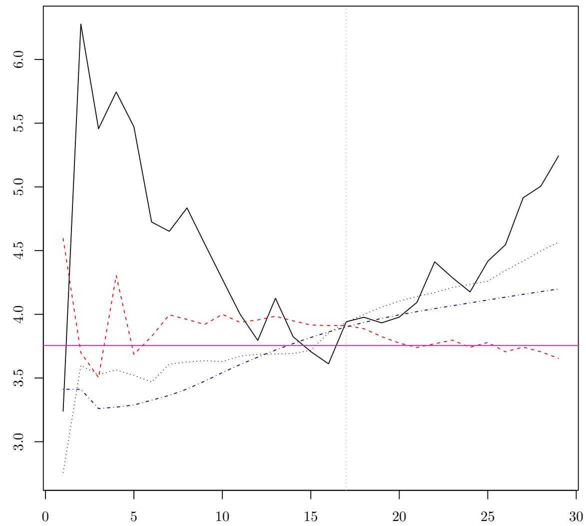
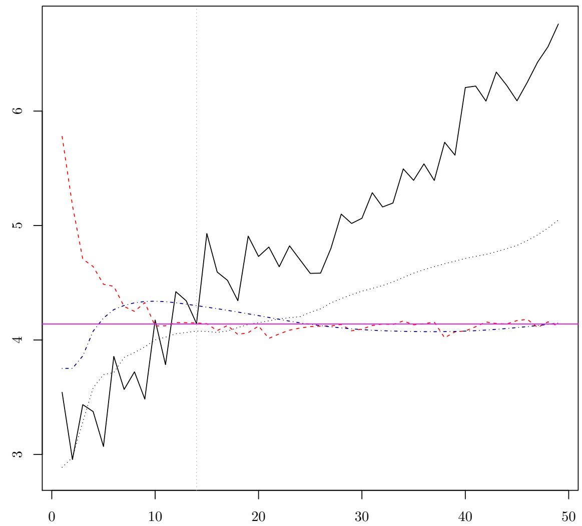


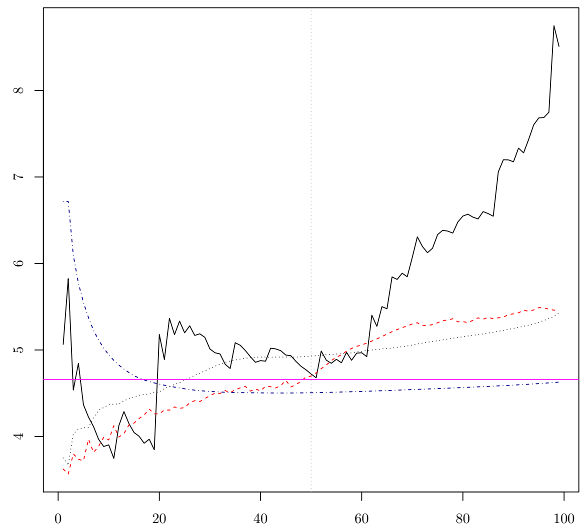
Figure 4: AMSE of the estimators $\hat{x}^{\mathrm{W}}_{p_n}$ (——), $\hat{x}^{\mathrm{WG}}_{p_n}$ ($\cdots\cdots$), $\hat{x}^{\mathrm{L}}_{p_n}$ ($-\ -\ -$), $\hat{x}^{\mathrm{LG(1)}}_{p_n}$ (—— —— ——) and $\hat{x}^{\mathrm{LG(2)}}_{p_n}$ ($- \cdot -$) for $N = 1000$ simulated samples of size $n \in \{30, 50, 100, 500\}$ from the distributions of Fréchet (i)–(iv), Burr (v)–(viii), Pareto (ix)–(xii) and Student (xiii)–(xvi). The horizontal axis corresponds to the fraction sample $k_n = 1, \ldots, n-1$.



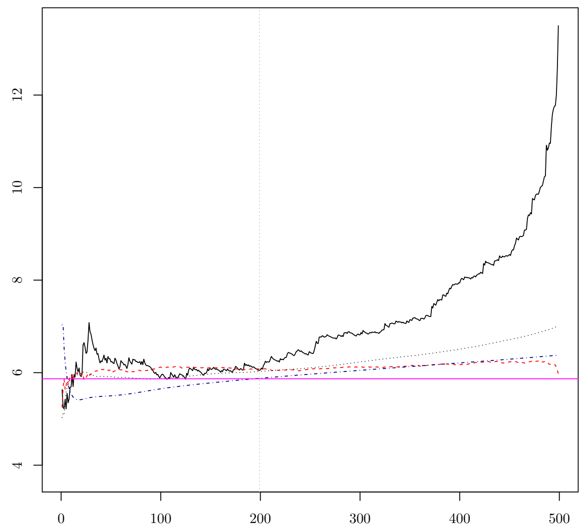
(a) $n = 30$



(b) $n = 50$

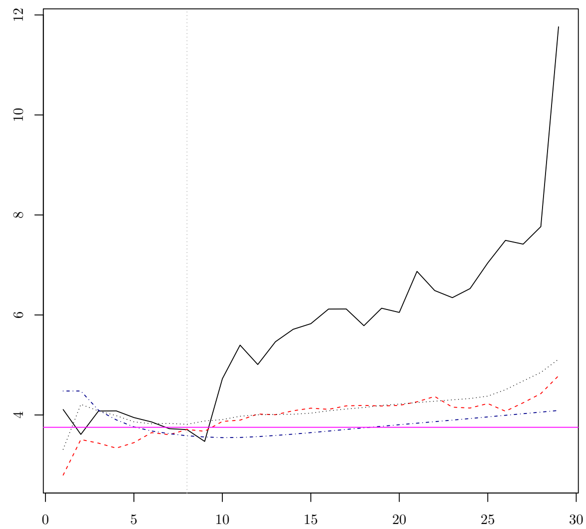


(c) $n = 100$

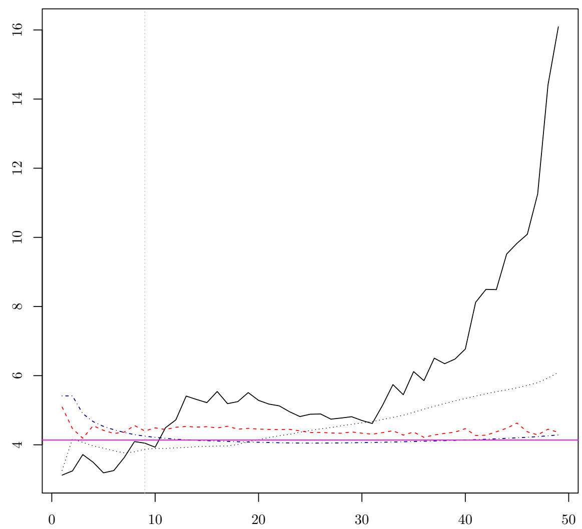


(d) $n = 500$

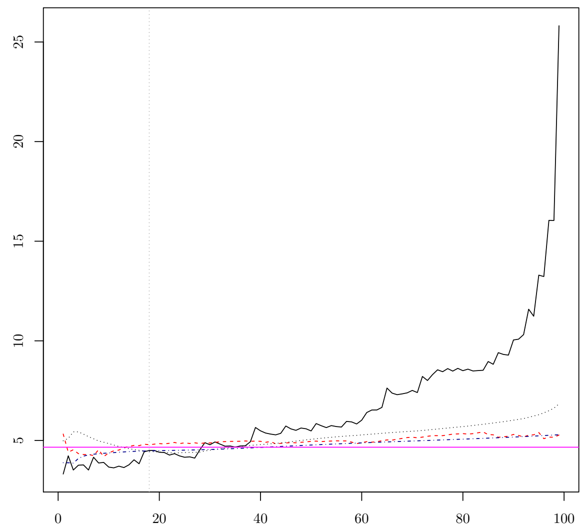
Figure 5: Choice of the sample fraction $\hat{k}_n$ *(vertical dotted line)* obtained by minimizing a dissimilarity measure between the estimators $\log \hat{x}_{p_n}^{\mathrm{W}}$ (——), $\log \hat{x}_{p_n}^{\mathrm{WG}}$ $(\cdots\cdots)$, $\log \hat{x}_{p_n}^{\mathrm{L}}$ $(- - -)$ and $\log \hat{x}_{p_n}^{\mathrm{LG(2)}}$ $(- \cdot -)$ for $N = 1000$ simulated samples of size $n \in \{30, 50, 100, 500\}$ from the Féchet distribution $\mathcal{F}(x; 0.75, 1, 0)$. The horizontal axis corresponds to the fraction sample $k_n = 1, \ldots, n-1$.



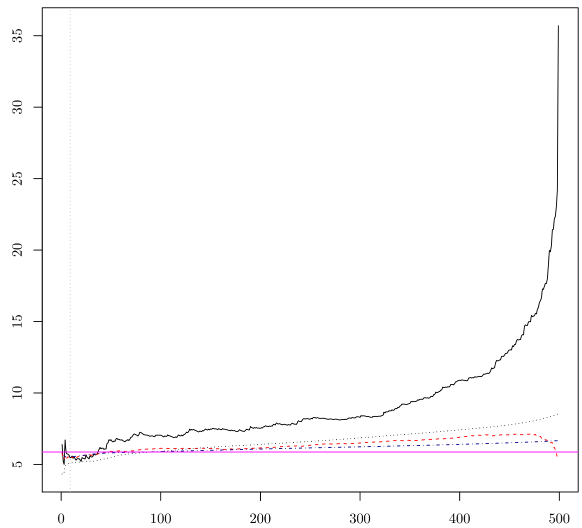
(a) $n = 30$



(b) $n = 50$



(c) $n = 100$



(d) $n = 500$

Figure 6: Choice of the sample fraction $\hat{k}_n$ *(vertical dotted line)* obtained by minimizing a dissimilarity measure between the estimators $\log \hat{x}^{\mathrm{W}}_{p_n}$ (——), $\log \hat{x}^{\mathrm{WG}}_{p_n}$ $(\cdots\cdots)$, $\log \hat{x}^{\mathrm{L}}_{p_n}$ $(- - -)$ and $\log \hat{x}^{\mathrm{LG(2)}}_{p_n}$ $(- \cdot -)$ for $N = 1000$ simulated samples of size $n \in \{30, 50, 100, 500\}$ from the Burr distribution $\mathcal{B}(x; 0.75, -1)$. The horizontal axis corresponds to the fraction sample $k_n = 1, \ldots, n-1$.

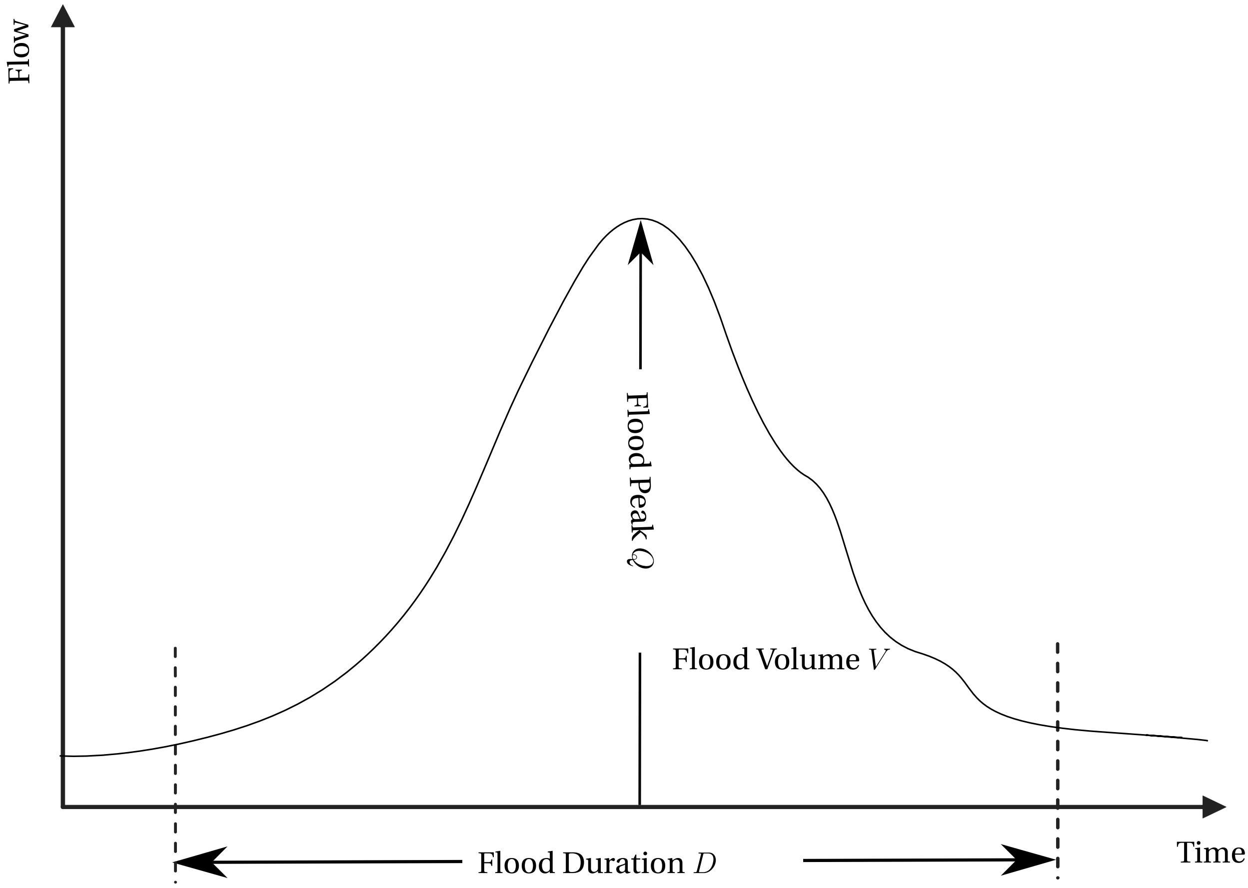


Figure 7: Typical flood hydrograph.

(a) PQ-plots ↓ (b) GQ-plots ↓

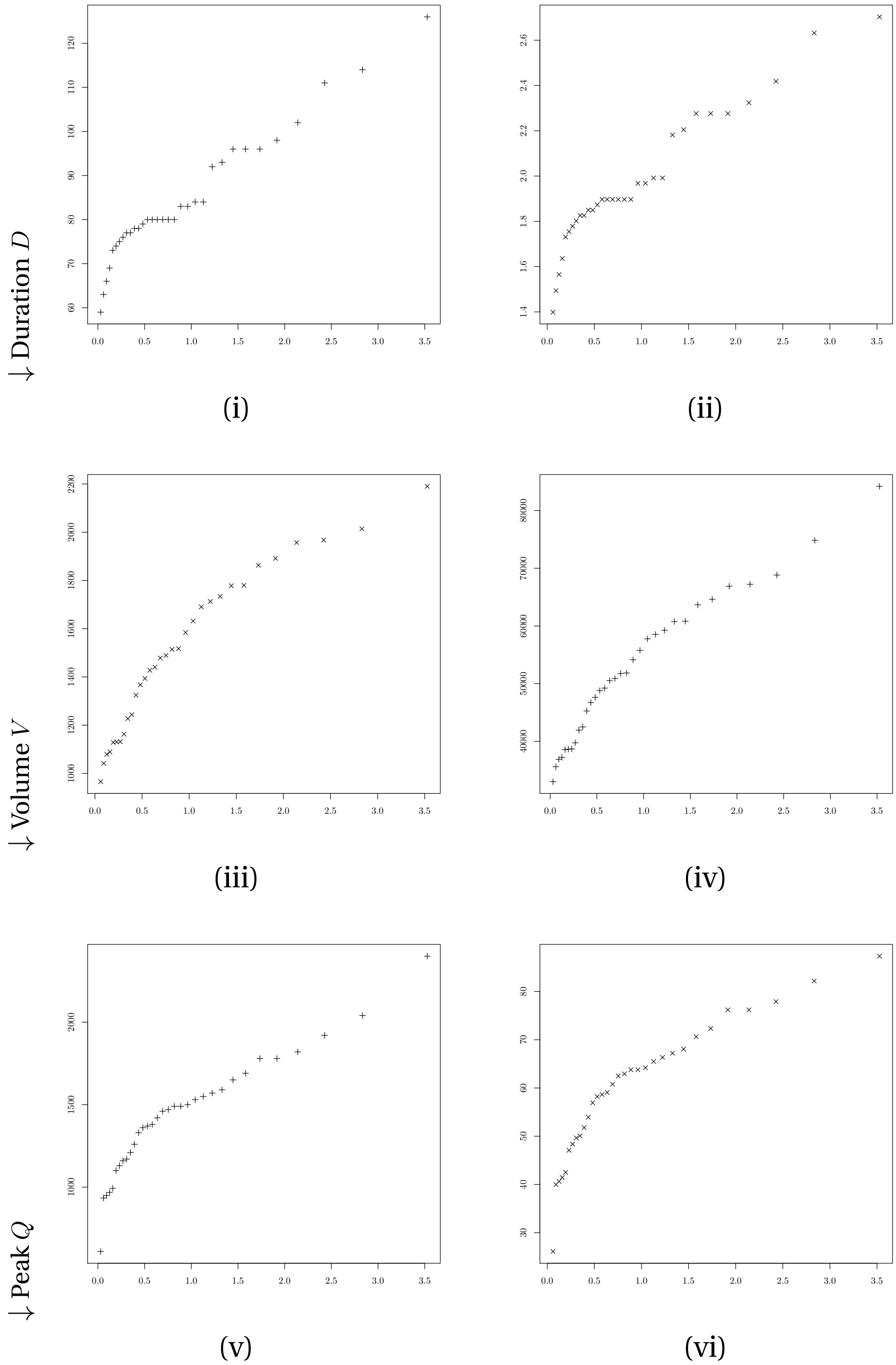
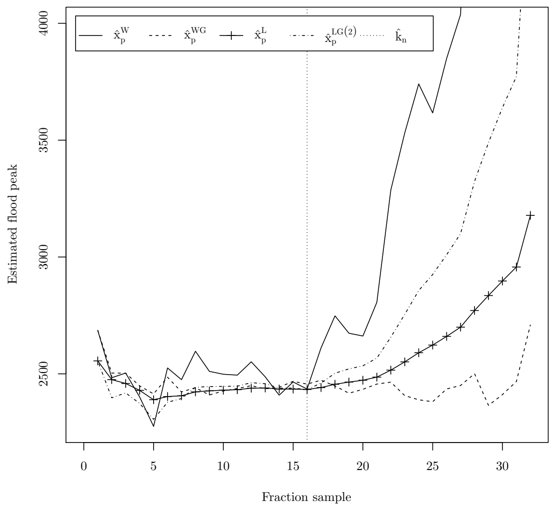
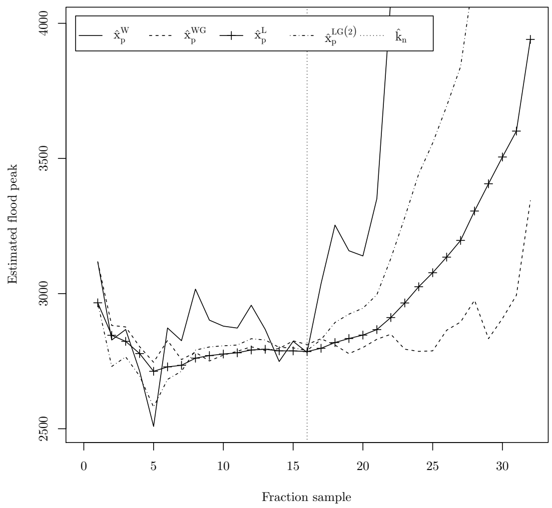


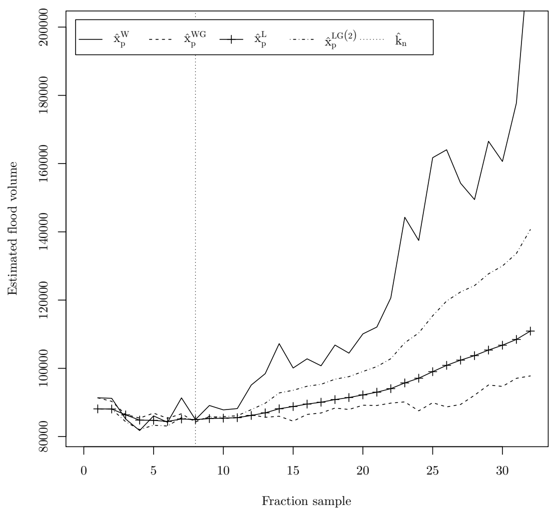
Figure 8: PQ-plots and GQ-plots obtained for duration (i)-(ii), flood volume (iii)-(iv) and flood peak (v)-(vi).



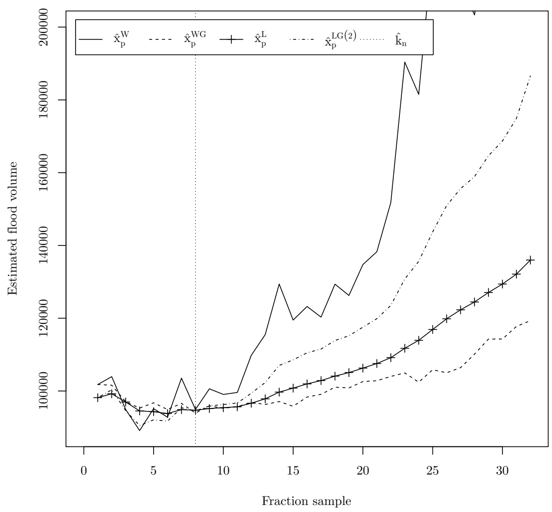
(a) $T = 66$ and $\hat{k}_n = 16$ for Q



(b) $T = 165$ and $\hat{k}_n = 16$ for Q



(c) $T = 66$ and $\hat{k}_n = 8$ for V



(d) $T = 165$ and $\hat{k}_n = 8$ for V

Figure 9: Estimated flood peaks (a)-(b) and estimated flood volumes (c)-(d) with $\hat{x}_p^{\mathrm{W}}$ (——), $\hat{x}_p^{\mathrm{WG}}$ $(- - -)$, $\hat{x}_p^{\mathrm{L}}$ $(- + -)$ and $\hat{x}_{p_n}^{\mathrm{LG}(2)}$ $(- \cdot -)$ for the indicated return period T, the selected fraction sample $\hat{k}_n$ $(\cdots\cdots)$.