

# **ECQ: Hydrological extreme value analysis tool**

P. Willems  
K.U.Leuven – Hydraulics Laboratory  
Kasteelpark Arenberg 40 – B-3001 Heverlee  
Patrick.Willems@bwk.kuleuven.ac.be

## ECQ: Methodology (Reference Manual)

### Reference:

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.

## Extreme value analysis

### 1. Introduction

In many water engineering applications, the accurate description of extreme surface water states (floodings, deteriorated water quality) and their recurrence rates is of primary importance. This description can be done either based on long-term time series of measurements (for discharges, water levels, pollutant concentrations, etc.), or long-term simulation results from mathematical models. For this purpose, an extreme value analysis is needed.

In extreme value analysis the tail of the distribution describing the probability of occurrence of extreme events is analysed and modelled by a separate distribution. The considered extremes might exist of extreme rainfall intensities, storm volumes, water levels, discharges, water quality parameters, etc.

### 2. Extreme value theory

Consider the following set of ordered and independent observations in a sample of the variable  $X$ , having probability distribution  $F_X$ :

$$x_1 \geq x_2 \geq \dots \geq x_m$$

where  $m$  denotes the total number of observations considered in the extreme value analysis.

As was shown first by Fisher & Tippett (1928), the probability distribution of the maximum observation  $x_t$ , eventually after locating and scaling, converges to a limited number of possible distributions as  $m$  tends to infinity. These distributions are called Generalized Extreme Value (GEV) distributions  $H(x)$ :

$$\begin{aligned}
 H(x) &= \exp\left(-\left(1 + \gamma \frac{x - x_t}{\beta}\right)^{-1/\gamma}\right) && \text{if } \gamma \neq 0 \\
 &= \exp\left(-\exp\left(-\frac{x - x_t}{\beta}\right)\right) && \text{if } \gamma = 0
 \end{aligned}$$

Pickands (1975) has shown moreover that, if only values of  $X$  above a sufficiently high threshold  $x_t$  are taken into consideration, the conditional distribution converges to the Generalized Pareto Distribution (GPD)  $G(x)$  as  $x_t$  becomes higher :

$$G(x) = 1 - \left(1 + \gamma \frac{x - x_t}{\beta}\right)^{-1/\gamma} \quad \text{if } \gamma \neq 0$$

$$= 1 - \exp\left(-\frac{x - x_t}{\beta}\right) \quad \text{if } \gamma = 0$$

In the case  $\gamma=0$ , this distribution matches the exponential distribution (with threshold  $x_t$ ).

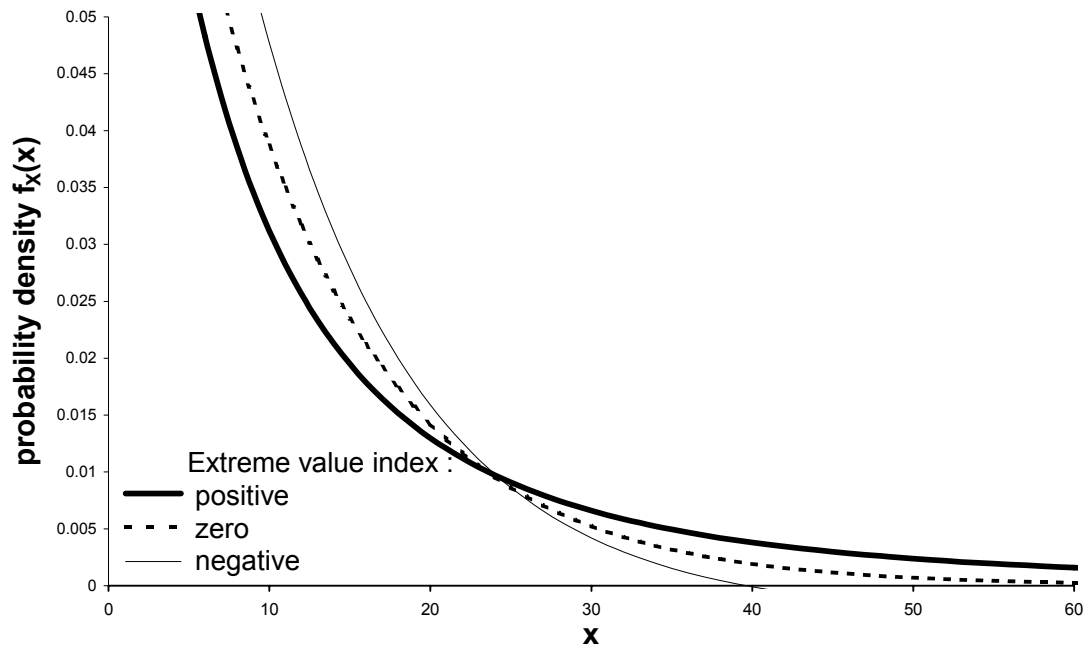
The parameter  $\gamma$  is called the extreme value index and shapes the tail of the distribution. For increasing  $\gamma$ , the tail becomes larger or ‘heavier’; the probability of occurrence becomes higher for extreme values of  $X$ . In the sub-cases  $\gamma < 0$ ,  $\gamma = 0$  and  $\gamma > 0$ , the shape of the tail is different (Figure 1). In the sub-case  $\gamma = 0$ , the tail decreases in an exponential way. In the sub-case  $\gamma > 0$ , it decreases in a polynomial way. This polynomial decrease can also be considered as an exponential decrease after logarithmic transformation of the variable  $X$ . It is clear that in this case the variable can take extremely high values. Just as the extremes can take unlimited high values in the cases  $\gamma > 0$  and  $\gamma = 0$ , they are limited in the case  $\gamma < 0$  and a maximum value exists for  $X$ . Some distributions having the same shape in excess of a threshold (at least asymptotically to infinity) are listed here for the three classes:

$\gamma < 0$  : Beta,

$\gamma = 0$  : Exponential, Normal, Lognormal, Gamma, Weibull, Gumbel,

$\gamma > 0$  : Pareto, Fréchet, Loggamma, Loghyperbolic.

The underlined distributions are the ones to which the distribution classes are often referred to. In hydrological applications, the classes  $\gamma = 0$  and  $\gamma > 0$  most frequently appear. In many of these applications, the extreme value index  $\gamma$  has a small positive value and the distinction between the two classes is of primary importance.



**Figure 1. Illustration example of the difference of the tail's shape for the 3 classes of extreme value distributions.**

### 3. Extreme value analysis based on quantile plots

#### 3.1. Different types of quantile plots

In a Q-Q plot (quantile-quantile plot), empirical quantiles are shown against theoretical quantiles. The empirical quantiles match the observed extremes  $x_i$ ,  $i=1, \dots, m$  ( $x_1 \leq \dots \leq x_m$ ), with  $p_i = i/(m+c)$  as their corresponding empirical probabilities of exceedance. In this study, the scores  $c$  ( $0 \leq c \leq 1$ ) are given value 1, corresponding to the Weibull plotting position of a quantile plot. For each empirical quantile  $x_i$ , the theoretical quantile is defined as  $F^{-1}(1-p_i)$ . The function  $F(x)$  is the cumulative distribution that is tested in the Q-Q plot and the Q-Q plot is named according to this distribution (exponential Q-Q plot, Pareto Q-Q plot, Weibull Q-Q plot, etc.). In some Q-Q plots, logarithmic transformed quantile values are plotted on both axes. If the observations agree with the considered distribution  $F(x)$ , the points in the QQ-plot approach the bisector.

In practice, however, one wants to test the validity of the distribution  $F(x)$  without knowledge of the parameter values. Adapted Q-Q plots are therefore used. In these adapted Q-Q plots, the so-called quantile function  $U(p)$  is plotted instead of the inverse distribution  $F^{-1}(1-p)$ . The quantile function  $U(p)$  is defined as the simplest function that is linearly dependent on  $F^{-1}(1-p)$  and independent on the parameter values of  $F(x)$ . These adapted Q-Q plots are denoted hereafter as 'quantile plots'.

The quantile function is also called reduced variate in the literature. Quantile functions do not exist, however, for all types of distributions. In the case of exponential, Pareto and Weibull quantile plots, the same quantile function is found:

$$U(p) = -\ln(p) = -\ln(1-G(x))$$

Expressions to draw exponential, Pareto and Weibull quantile plots are listed hereafter ( $U(p)$  or  $\ln(U(p))$  on the horizontal axis;  $x$  or  $\ln(x)$  on the vertical axis):

- Exponential quantile plot :

$$\left(-\ln\left(\frac{i}{m+1}\right); x_i\right)$$

- Pareto quantile plot :

$$\left(-\ln\left(\frac{i}{m+1}\right); \ln(x_i)\right)$$

- Weibull quantile plot :

$$\left(\ln\left(-\ln\left(\frac{i}{m+1}\right)\right); \ln(x_i)\right)$$

In section 7 examples are shown for the exponential and Pareto quantiles plots (Figures 4.d and 5.a for the exponential quantile plots, and Figure 4.b for the Pareto quantile plot).

Another type of quantile plot was introduced by *Beirlant et al.* [1996] to estimate in a visual way the extreme value index and the corresponding distribution class. It is referred to as the generalized quantile plot and it is defined as:

$$\left(-\ln\left(\frac{i}{m}\right); \ln(UH_i)\right)$$

The values  $UH_i$  ( $i=2, \dots, m$ ) are empirical values of the so-called excess function  $UH(p)$  (also denoted as the UH function). This function can be found empirically as:

$$UH_i = x_{i+1} \left( \frac{\sum_{j=1}^i \ln(x_j)}{i} - \ln(x_{i+1}) \right)$$

and it is the product of the quantile function  $U(p)$  and an estimation of the so-called mean excess function  $e_{\ln(x)}(\ln(x))$ :

$$e_{\ln(x)}(\ln(x)) = E[\ln(X) - \ln(x) | X > x]$$

This function can be understood theoretically as the mean exceedance of the threshold  $x$ , after logarithmic transformation.

The generalized quantile plot is hereafter called UH-plot, according to the notation of the excess function  $UH(p)$  from which empirical values are plotted.

### 3.2. Three classes of distributions

In terms of the mean excess function, the three classes of the extreme value index can be marked off:

- $\gamma < 0$  : the mean excess decreases (towards the higher values) in the tail of the distribution; this means that the mean exceedance of a threshold decreases for higher thresholds; this corresponds to the limitation of  $X$  in this class.
- $\gamma = 0$  : the mean excess has a constant value (this is a well-known property of the exponential distribution).
- $\gamma > 0$  : the mean excess increases in the tail.

#### *UH-estimation extreme value index*

Beirlant et al. (1996) have shown that, whenever the tail of the studied distribution converges to a GEV or GPD distribution, the UH-plot has an asymptotic linear behaviour. The slope of the line is equal to the extreme value index  $\gamma$ :

$$\ln(UH(p)) \sim \gamma (-\ln(1-p)) \quad \text{when} \quad -\ln(1-p) \rightarrow \infty$$

Even if only a few extreme observations are available, the asymptotic linear behaviour can already be observed in the UH-plot in many water engineering applications. In some cases with many observations, even a perfect linear trend can be observed.

Because of the asymptotic nature of the linear path of the points in the UH-plot, more weight has to be given to the higher observations in the estimation of  $\gamma$ . For this reason, a weighted linear regression above a certain threshold  $x_t$  ( $t \geq m$ ) is performed on the UH-plot to estimate  $\gamma$ :

$$\hat{\gamma}_t = \frac{\sum_{j=1}^{t-1} w_j \ln\left(\frac{j}{t}\right) (\ln(UH_t) - \ln(UH_j))}{\sum_{j=1}^{t-1} w_j \ln^2\left(\frac{j}{t}\right)}$$

with weights  $w_j$  that can be chosen equal to (these are derived by Hill (1975), and are therefore also called 'Hill-weights':

$$w_j = -\frac{1}{\ln\left(\frac{j}{t}\right)}$$

An example of a plot of the UH-estimations versus the threshold rank  $t$  is given in Figure 4.a.

The optimal threshold  $x_t$ , above which the weighted regression has to be performed in an optimal estimation of  $\gamma$ , is the threshold that minimizes the mean squared error (MSE) of the regression:

$$MSE_t = \frac{1}{t-1} \left( \sum_{j=1}^{t-1} w_j \left( \ln\left(\frac{UH_j}{UH_t}\right) - \hat{\gamma}_t \ln\left(\frac{t}{j}\right) \right)^2 \right)$$

In Figure 4.a, also the results of this MSE calculation are added.

The MSE values increase at both low and high threshold values. At high threshold values (small  $t$ ), few observations are available and an accurate regression cannot be performed. The statistical uncertainty  $\sigma^2(\hat{\gamma}_t)$ , in terms of variance, increases for smaller  $t$ . For lower threshold values (higher  $t$ ) the points in the UH-plot start to deviate from the linear path and a bias appears in the estimation of  $\gamma$ . The mean squared error of the estimation is considered as the sum of the statistical variance and the squared bias:

$$MSE(\hat{\gamma}_t) = \sigma^2(\hat{\gamma}_t) + BIAS^2(\hat{\gamma}_t)$$

An asymptotic linear path can also be observed in the first derivative plot of  $\gamma$ . The slope  $\rho$  in this plot is related to the rate of convergence to a linear path in the UH-plot. The rate is higher for more negative values of  $\rho$ .

The separate estimation of  $\rho$  makes the determination of the optimal threshold possible; optimal weights for the linear regression in the UH-plot can be found as a function of  $\rho$ . These optimal weights are derived by Vynckier (1996). When they are used,  $\gamma$  can only be determined iteratively because the optimal weights, which are used to estimate  $\gamma$ , are themselves dependent on  $\gamma$  and  $\rho$ . In many applications, however, accurate approximations are found without iteration by applying the Hill-weights.

### 3.3. Positive extreme value index – Pareto quantile plot

To most typical distribution of the class  $\gamma > 0$  is the strict Pareto distribution  $F_X(x)$ :

$$F_X(x) = 1 - x^{-\alpha}$$

The other distributions of this class are therefore called Pareto-type distributions and have the same shape in the tail. The parameter  $\alpha = 1/\gamma$  is also called Pareto index.

Because of the convergence of the quantile function  $U(p)$  to a linear path in the Pareto quantile plot (with slope  $\gamma$ ):

$$\ln(UH(p)) \rightarrow \gamma (-\ln(1-p)) \quad \text{when } -\ln(1-p) \rightarrow \infty$$

the extreme value index  $\gamma$  can also be estimated in the Pareto quantile plot in the same way as it is estimated in the UH-plot. Asymptotic expressions of the variance  $\sigma^2(\hat{\gamma}_t)$  of the estimator even show a more accurate estimation.

When using the Hill-weights, the estimator of  $\gamma$  by means of the Pareto quantile plot equals the one that was suggested by Hill (1975):

$$\hat{\gamma}_t = \frac{1}{t-1} \left( \sum_{j=1}^{t-1} \ln(x_j) \right) - \ln(x_t) \quad \text{for } \gamma > 0$$

This ‘Hill-estimator’ is the first estimator introduced for the estimation of the extreme value index in the case  $\gamma > 0$  and it has been used successfully in many applications. It is easily shown that the Hill-estimator also equals the mean excess (the average increase in the Pareto quantile plot from the threshold  $x_t$  on), as was defined before. At this place, the notation  $UH$  can be explained. The letter  $H$  is standing for the Hill-estimator (from the threshold on) and the letter  $U$  is standing for the tail quantile function (in this case evaluated at the threshold).

In a more general way, using any function for the weights, the following equation can be used for  $\hat{\gamma}_t$ :

$$\hat{\gamma}_t = \frac{\sum_{j=1}^{t-1} w_j \ln\left(\frac{j}{t}\right) (\ln(x_t) - \ln(x_j))}{\sum_{j=1}^{t-1} w_j \ln^2\left(\frac{j}{t}\right)} \quad \text{for } \gamma > 0$$

When the formula for the Hill-weights is filled in this formula and using the approximation:

$$-\frac{1}{t-1} \sum_{j=1}^t \ln\left(\frac{j}{t}\right) \approx 1$$

the Hill-estimation formula is found.

The MSE of the weighed linear regression in the Pareto quantile plot is given by:

$$MSE_t = \frac{1}{t-1} \left( \sum_{j=1}^{t-1} w_j \left( \ln\left(\frac{x_j}{x_t}\right) - \hat{\gamma}_t \ln\left(\frac{t}{j}\right) \right)^2 \right)$$



The asymptotic variance of the Hill-estimator of  $\gamma$  (using the Hill-weights) is found to be:

$$\begin{aligned}\sigma^2(\hat{\gamma}_t) &= \frac{\gamma^2}{t-1} && \text{via Pareto quantile plot} \\ &= \frac{(\gamma+1)^2}{t-1} && \text{via UH - plot}\end{aligned}$$

If the observations are POT-values, a GPD distribution can be fitted above the optimal threshold  $x_t$ , which is also a parameter in the GPD. The remaining parameter  $\beta$  has not to be calibrated in that case, it equals:

$$\beta = \gamma x_t$$

This last expression can be determined considering the asymptotic linear path in the Pareto quantile plot :

$$-\ln(1-G(x)) + \ln(1-G(x_t)) = -\frac{1}{\gamma}(\ln(x) - \ln(x_t))$$

and by inserting the expression of the GPD distribution  $G(x)$ :

$$-\ln(1-G(x)) + \ln(1-G(x_t)) = -\frac{1}{\gamma} \ln\left(1 + \gamma \frac{x - x_t}{\beta}\right)$$

When higher thresholds  $x_s$  ( $s > t$ ) are considered, the GPD keeps being valid after adapting the parameter  $\beta$ :

$$\beta \rightarrow \beta + \gamma (x_s - x_t)$$

For lower thresholds, one has to look for another type of distribution in the class  $\gamma > 0$ , which can describe the deviation from the linear path below  $x_t$  in the Pareto quantile plot. For each distribution  $F_X(x)$  in this class, a relation can be found between the parameters of the distribution and the extreme value index (see also Beirlant et al. (1996)). The other parameters can be determined by maximizing the agreement between the distribution and the linear path above  $x_t$  in the Pareto quantile plot.

In some cases, a GPD keeps being valid if  $s$  is not much higher than  $t$  and if  $MSE_s$  is not much higher than  $MSE_t$ .

The limiting case  $\gamma = 0$  can be observed as a continuously decreasing slope in the Pareto quantile plot; the points keep turning down to a horizontal line (asymptotic slope of zero).

### 3.4. Zero extreme value index – Weibull quantile plot

The most typical distribution of the class  $\gamma = 0$  is the Weibull distribution:

$$F_X(x) = 1 - \exp\left(-\left(\frac{x}{\beta}\right)^\tau\right)$$

The other distributions of the class are called Weibull-type distributions.

The shape parameter  $\tau$  is called the Weibull index and, like the extreme value index, is a measure of the tail-heaviness of the distribution, although in another way :

- $\tau > 1$  : light or ‘super-exponential’ tail
- $\tau = 1$  : exponential distribution
- $0 < \tau < 1$  : heavy or ‘sub-exponential’ tail

The sub-case  $\tau = 0$  can be considered as a limiting case for the Weibull-type distributions; it corresponds to the Pareto class.

Analogous to the extreme value index, the Weibull index can be estimated by a weighted linear regression in the Weibull quantile plot; Weibull-type distributions converge to a linear path with slope  $1/\tau$  in the Weibull quantile plot:

$$\ln(UH(p)) \sim \frac{1}{\tau} \ln(-\ln(1-p)) \quad \text{when } -\ln(1-p) \rightarrow \infty$$

The following Hill-type estimator of  $1/\tau$ , using the Hill-weights, can be determined :

$$\frac{1}{\hat{\tau}_t} = \frac{\hat{\gamma}_t}{a_t}$$

$\hat{\gamma}_t$  is the Hill-estimator based on the Pareto quantile plot and  $a_t$  equals:

$$a_t = \frac{\sum_{j=1}^{t-1} w_j \ln\left(\frac{j}{t}\right) \left( \ln\left(-\ln\left(\frac{t}{m+1}\right)\right) - \ln\left(-\ln\left(\frac{j}{m+1}\right)\right) \right)}{\sum_{j=1}^{t-1} w_j \ln^2\left(\frac{j}{t}\right)}$$

The asymptotic variance of this estimator for  $\gamma$  (using the Hill-weights) is found to be:

$$\sigma^2\left(\frac{1}{\hat{\tau}_t}\right) = \frac{1}{(t-1)\tau^2}$$

The optimal threshold can also be determined by minimizing the MSE of the weighted regression in the Weibull quantile plot:

$$MSE_t = \frac{1}{t-1} \sum_{j=1}^{t-1} w_j \left( \ln\left(\frac{x_j}{x_t}\right) - \frac{1}{\hat{\tau}_t} \ln\left(\frac{\ln\left(\frac{j}{m+1}\right)}{\ln\left(\frac{t}{m+1}\right)}\right) \right)^2$$

### 3.5. Zero extreme value index – exponential quantile plot

When a Weibull index close to 1 is determined, the possibility  $\tau = 1$  can be tested by means of an exponential QQ-plot. The slope of an (eventual) asymptotic linear path of the points in this plot is equal to the only parameter  $\beta$  of the exponential distribution:

$$G(x) = 1 - \exp\left(-\frac{x - x_t}{\beta}\right)$$

Analogous to the extreme value index and the Weibull index, the index  $\beta$  can be estimated by a weighted linear regression in the exponential quantile plot:

$$\hat{\beta}_t = \frac{\sum_{j=1}^{t-1} w_j \ln\left(\frac{j}{t}\right) (x_t - x_j)}{\sum_{j=1}^{t-1} w_j \ln^2\left(\frac{j}{t}\right)}$$

Using Hill-weights, the following estimator is found (the Hill-type estimator for  $\beta$ ):

$$\hat{\beta}_t = \frac{1}{t-1} \left( \sum_{j=1}^{t-1} x_j \right) - x_t$$

The optimal threshold can also be determined by minimizing the MSE of the weighted regression in the exponential quantile plot:

$$MSE_t = \frac{1}{t-1} \left( \sum_{j=1}^{t-1} w_j (x_j - x_t - \hat{\beta}_t \ln\left(\frac{t}{j}\right)) \right)^2$$

### 3.6. Summary

#### *summary formulas*

In a summarized way, the following formulas have to be considered in the extreme value analysis for the different plots (depending on the distribution class):

UH-plot:

$$\left(-\ln\left(\frac{i}{m}\right); \ln(UH_i)\right)$$

$$UH_i = x_{i+1} \left( \frac{\sum_{j=1}^i \ln(x_j)}{i} - \ln(x_{i+1}) \right)$$

$$\hat{\gamma}_t = \frac{\sum_{j=1}^{t-1} w_j \ln\left(\frac{j}{t}\right) (\ln(UH_t) - \ln(UH_j))}{\sum_{j=1}^{t-1} w_j \ln^2\left(\frac{j}{t}\right)}$$

$$MSE_t = \frac{1}{t-1} \left( \sum_{j=1}^{t-1} w_j \left( \ln\left(\frac{UH_j}{UH_t}\right) - \hat{\gamma}_t \ln\left(\frac{t}{j}\right) \right)^2 \right)$$

Pareto quantile plot:

$$\left(-\ln\left(\frac{i}{m+1}\right); \ln(x_i)\right)$$

$$\hat{\gamma}_t = H_t = \frac{\sum_{j=1}^{t-1} w_j \ln\left(\frac{j}{t}\right) (\ln(x_t) - \ln(x_j))}{\sum_{j=1}^{t-1} w_j \ln^2\left(\frac{j}{t}\right)} \quad (\text{only for } \gamma > 0 !!)$$

$$MSE_t = \frac{1}{t-1} \left( \sum_{j=1}^{t-1} w_j \left( \ln\left(\frac{x_j}{x_t}\right) - \hat{\gamma}_t \ln\left(\frac{t}{j}\right) \right)^2 \right)$$

$$\hat{\beta}_t = \hat{\gamma}_t x_t$$

Weibull quantile plot:

$$\left(\ln\left(-\ln\left(\frac{i}{m+1}\right)\right); \ln(x_i)\right)$$

$$\frac{1}{\hat{\tau}_t} = \frac{H_t}{a_t}$$

$$a_t = \frac{\sum_{j=1}^{t-1} w_j \ln\left(\frac{j}{t}\right) \left(\ln\left(-\ln\left(\frac{t}{m+1}\right)\right) - \ln\left(-\ln\left(\frac{j}{m+1}\right)\right)\right)}{\sum_{j=1}^{t-1} w_j \ln^2\left(\frac{j}{t}\right)}$$

$$MSE_t = \frac{1}{t-1} \sum_{j=1}^{t-1} w_j \left(\ln\left(\frac{x_j}{x_t}\right) - \frac{1}{\hat{\tau}_t} \ln\left(\frac{\ln\left(\frac{j}{m+1}\right)}{\ln\left(\frac{t}{m+1}\right)}\right)\right)^2$$

Exponential quantile plot:

$$\left(-\ln\left(\frac{i}{m+1}\right); x_i\right)$$

$$\hat{\beta}_t = \frac{\sum_{j=1}^{t-1} w_j \ln\left(\frac{j}{t}\right) (x_t - x_j)}{\sum_{j=1}^{t-1} w_j \ln^2\left(\frac{j}{t}\right)}$$

$$MSE_t = \frac{1}{t-1} \left(\sum_{j=1}^{t-1} w_j (x_j - x_t - \hat{\beta}_t \ln\left(\frac{t}{j}\right))^2\right)$$

using Hill weights:

$$w_j = -\frac{1}{\ln\left(\frac{j}{t}\right)}$$

*summary distribution tail analysis*

Making use of these three types of quantile plots, an analysis can be made of the shape of the distribution's tail, and discrimination can be made between heavy tail, normal tail and light tail behaviour. This is summarized in the figures 2 and 3.

The distribution's tail can be considered normal when (Figure 2):

- in the exponential quantile plot: the upper tail points tend towards a straight line;
- in the Pareto quantile plot: the upper tail points continuously bend down;
- in the UH-plot: the slope in the upper tail approaches the zero value.

The distribution's tail is heavy when (Figure 3):

- in the exponential quantile plot: the upper tail points continuously bend up;
- in the Pareto quantile plot: the upper tail points tend towards a straight line;
- in the UH-plot: the slope in the upper tail is systematically positive.

The distribution's tail is light when:

- in the exponential quantile plot: the upper tail points continuously bend down;
- in the Pareto quantile plot: the upper tail points also continuously bend down;
- in the UH-plot: the slope in the upper tail is systematically negative.

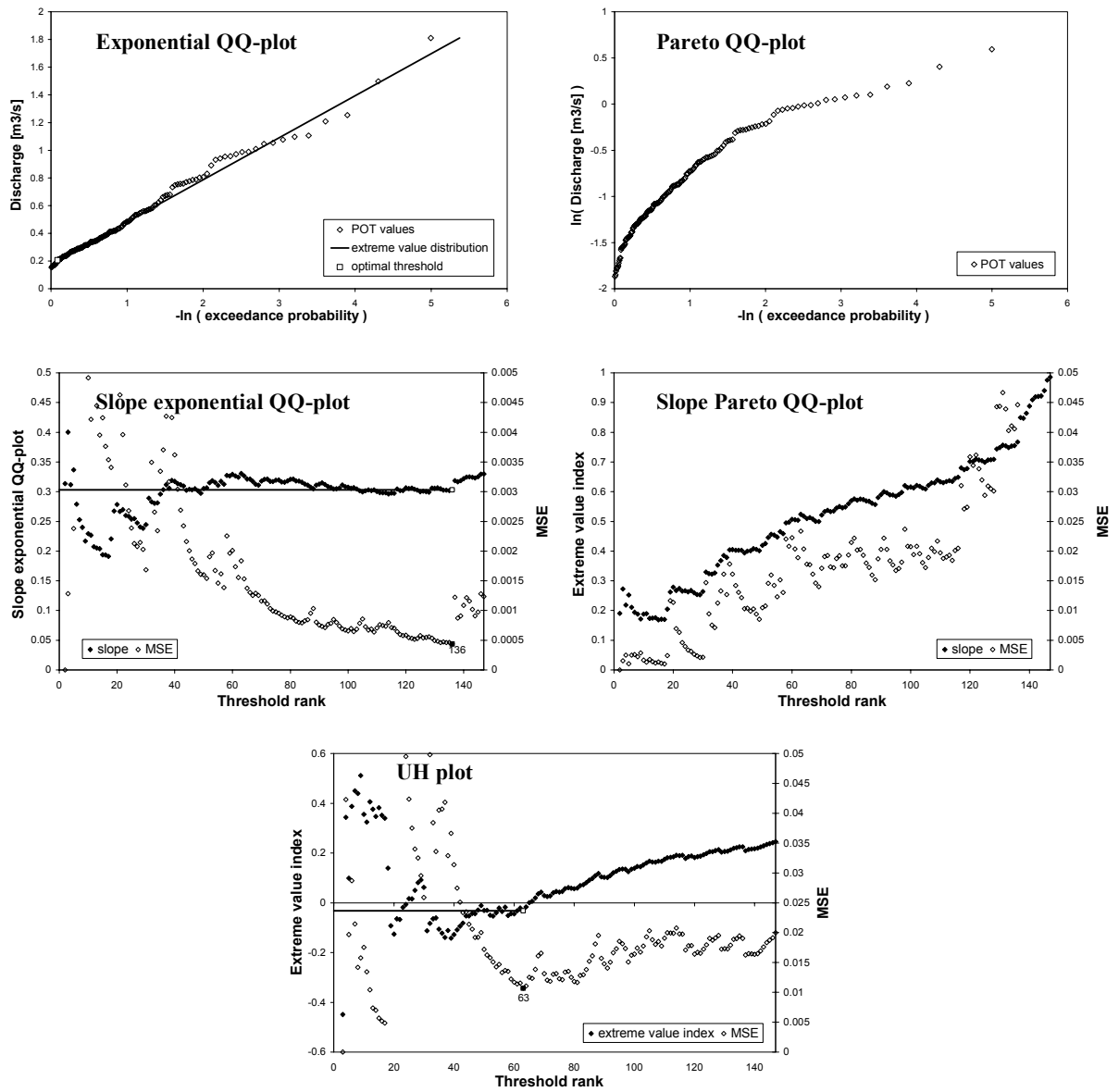


Figure 2. Example of QQ-plots for a normal tail case.

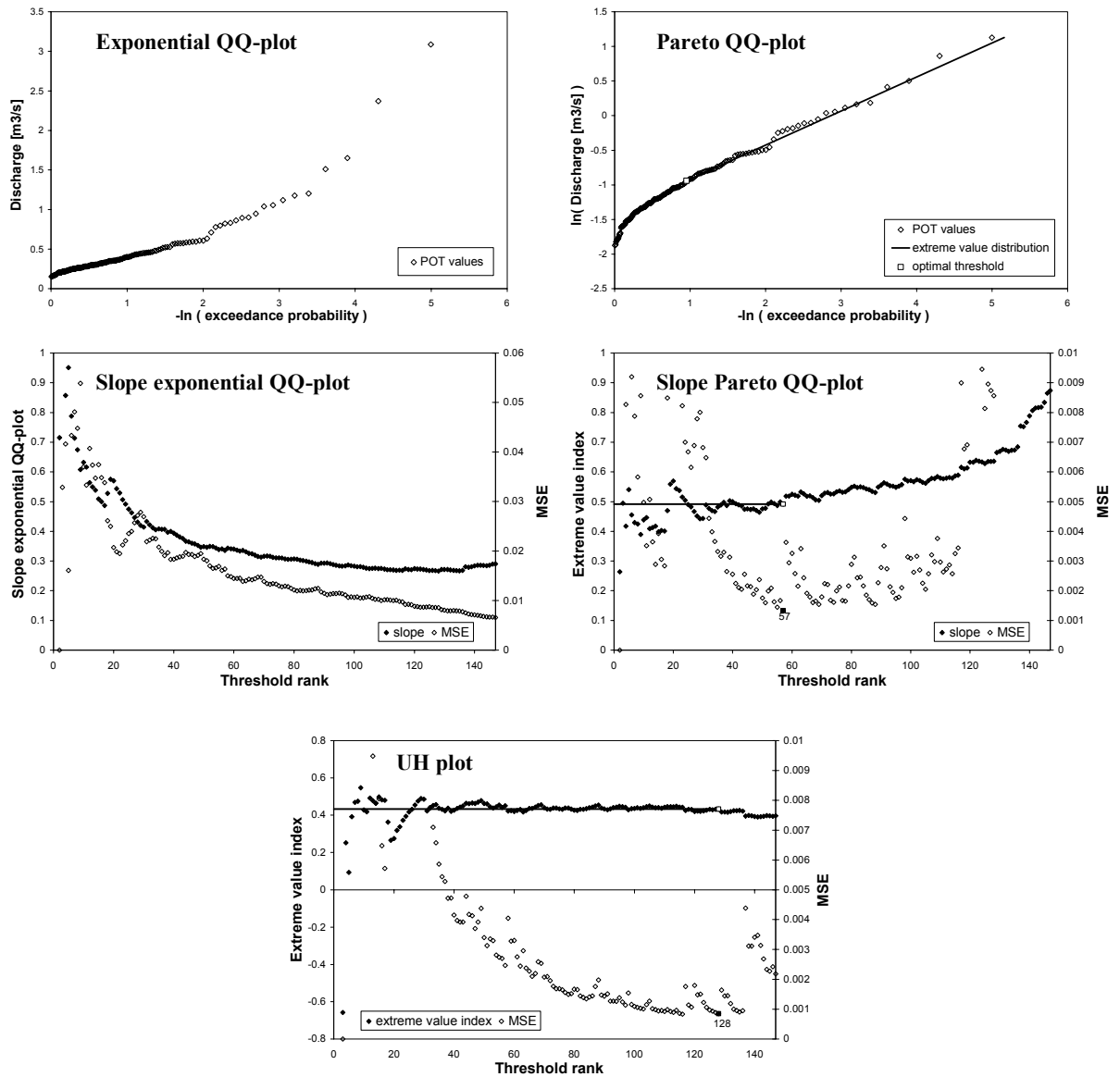


Figure 3. Example of QQ-plots for a heavy tail case.



#### 4. Extreme value analysis based on traditional calibration methods

Also the calibration techniques that are ‘traditionally’ used for calibration of distribution parameters can be used to calibrate the extreme value distributions. The different ‘typical’ extreme value distributions, as listed in section 2, are chosen and the parameters calibrated above threshold levels that are chosen arbitrarily. Method of moments (MOM), maximum likelihood method (ML), etc. can be used for that purpose. An overview of these statistical methods and their application to the GPD distribution is given by Hosking and Wallis (1987).

This ‘traditional’ calibration method has, however, some disadvantages. Only a very limited number of distributions are tried in practice, possibly all of the same class, and a distribution with a wrong index (extreme value or Weibull index) might be determined. In spite of a good fit in the range of  $X$  for which observations are available, an extrapolation outside this range, by means of the distribution, can be very erroneous in that case. Very extreme events can be strongly overestimated or underestimated. A visualization of the distribution in the quantile plot of the class to which the distribution belongs (exponential, Pareto or Weibull quantile plot; see previous section), makes this error undeniably clear.

The error in the index is caused by a wrong type of distribution (wrong sign for the index) and/or a non-optimal threshold (wrong value for the index). In many such erroneous cases, a set of parameter values that result in a good fit for the distribution is used, without recognizing the wrong shape in the tail. This wrong shape is indeed not recognizable in many traditional visualizations. In some cases, even many sets of parameter values that fit the distribution well exist. The more parameters in the distribution function, the more this problem manifests itself. The thought that more complex distributions can induce more accurate extreme value analyses is thus illusive. The extreme value analysis often becomes a curve fit application in that case and the distribution is no longer an accurate statistical basis for extrapolation, as it should be.

Not all parameters of the distributions can be estimated using the methodology presented in the previous section, especially not in the case of complex distributions. The UH- and Hill-estimators may also not be the most efficient. The important advantages of this methodology are, anyhow, the estimation of the sign and the order of magnitude of the indices, the determination of the distribution-class, the estimation of the optimal threshold level and its visual nature. A combination of this methodology and the traditional methods is most beneficial. The calibration can first be based on the methodology presented in the previous section to determine the distribution class (and in some cases also the more specific type of distribution), the confidence ranges of the indices and a valid range for the optimal threshold. Traditional methods can then be applied afterwards to estimate the parameter values of the distribution that meets the previously determined conditions.

## 5. Return period

If  $G(x)$  represents the probability distribution of the extremes above a threshold  $x_t$ , calibrated to  $t$  observations in  $n$  periods (e.g. years), the ‘return period’  $T$  (also called recurrence interval) of the exceedance level  $x$  then equals:

$$T[\text{number of years}] = \frac{n}{t} \frac{1}{1 - G(x)}$$

If a Poisson process can be assumed for the independent exceedances of a minimum threshold  $x_m$ , a relationship exist between the parameters of the GPD and the ones of the GEV (Buishand, 1989). The parameter  $\gamma$ , as the extreme value index, is identical in both cases and the same distribution classes exist.

In combination with the method of periodic maxima, the return period is calculated as the inverse of the population survival distribution of the annual maxima:

$$T[\text{number of years}] = \frac{1}{1 - H(x)}$$

After Langbein (1947) and some assumptions mentioned by Chow et al. (1988), it can be easily shown that this definition corresponds to the following formula whenever population distributions  $G(x)$  of POT-values are used ( $T_{AM}$  is the calculation using the formula for annual maxima, while  $T_{POT}$  for the POT-values):

$$\frac{1}{T_{JM}} = 1 - \exp\left(-\frac{1}{T_{POT}}\right)$$

On the basis of the linear regressions in the exponential, or Pareto, or Weibull quantile plots, the return period can also be calculated in an easier way:

- for the exponential case:

$$x = x_t + \hat{\beta} \left( \ln(T) - \ln\left(\frac{n}{t}\right) \right)$$

- for the Pareto case:

$$\ln(x) = \ln(x_t) + \hat{\gamma} \left( \ln(T) - \ln\left(\frac{n}{t}\right) \right)$$

- for the Weibull case:

$$\ln(x) = \ln(x_t) + \frac{1}{\hat{\tau}} \ln\left(\ln(T) - \ln\left(\frac{n}{t}\right)\right)$$

De ratio  $n/t$  indeed equals the ‘empirical’ return period at the threshold level  $x_t$ .

The return period has to be interpreted as the mean time between two independent exceedances (corresponding to the subjective choice of the independency criterion). It is an important parameter for studying flood and water quality problems as it jointly describes the probability of occurrence of extreme events and the temporal dimension. In this way, it is a

measure of safety (the ‘inverse risk’). Based on the calculation of the water levels with a certain return period along a river, flood protection can be achieved with a certain acceptable risk level (or with a certain safety level). In this way, flood protection programs can be set up, such as the Delta Plan and Sigma Plan in The Netherlands and Belgium. Related to pollution problems in rivers, the return period of exceedance of water quality standards can be studied to limit the risk of deterioration of the aquatic life and to define water restoration programs. Also for low flow or drought analyses, extreme value analysis is useful.

If only periodic maxima (e.g. yearly maxima) are available in the application, a more simplified extreme value analysis can be performed; the extraction of extremes is much simpler and no independency criterion is needed if the considered ‘periods’ are not short in time. Instead of the GPD distribution, the Generalized Extreme Value (GEV) distribution  $H(x)$  is used in that case.

## 6. Examples of flood frequency analysis

To illustrate the methodology of extreme value analysis, the application is shown for extreme rainfall intensities at the main meteorological station in Belgium at Uccle. The observed extremes are derived from the time series as peak-over-threshold values by the moving-average technique. The 324 highest extremes (the threshold corresponding to the rainfall intensity with an empirical return period of one month) are extracted; the threshold level being different for different aggregation-levels. For the independency criterion, only a condition is used for the critical interevent time. It is taken equal to the aggregation-level, with a minimum value of 12 hours. The minimum value of 12 hours is set out by the fact that two flood events within the same day or night are commonly interpreted to be one extreme event. As stated in section 3.9.4, this subjective choice of the independency criterion influences the interpretation given to the return period of a rainfall extreme. Based on the minimum interevent time of 12 hours in the independency criterion, the return period calculated in this study should be interpreted as the average time between two days or nights with extreme rainfall intensities.

For two aggregation-levels, 10 min (no aggregation, 10 min is the time step of the rainfall series) and 5 days, the extreme value analysis results are shown below.

### *Aggregation-level of 10 min*

First, an UH-estimation of the extreme value index  $\gamma$  is performed to determine the distribution class and the sign of  $\gamma$ . In Figure 4.a, this UH-estimation is shown as a function of the number  $t$  of extremes that are used in the estimation. An Hill-type estimator is used. In Figure 4.a, also the MSE of the weighted linear regression in the UH-plot is shown as a function of  $t$ . For  $t < 50$ , the UH-estimation of  $\gamma$  fluctuates around zero.

The same is concluded from the Pareto quantile plot (Figure 4.b); the points keep turning down. Also the Hill-estimation of  $\gamma$  keeps decreasing to lower values of  $t$  (Figure 4.c). The class  $\gamma = 0$  is therefore suggested and tested in the exponential quantile plot (Figure 4.d). An asymptotic linear path is observed for the points in this plot. The Hill-type estimation of the parameter  $\beta$  of the exponential distribution, together with the MSE, can be found in Figure 4.e. The optimal threshold  $t = 45$  is determined. At this threshold, the estimate of the parameter  $\beta$  equals:  $\hat{\beta}_t = 14.9$  mm/h. The same value is determined after use of the methods MOM and ML (both are identical in the exponential case).

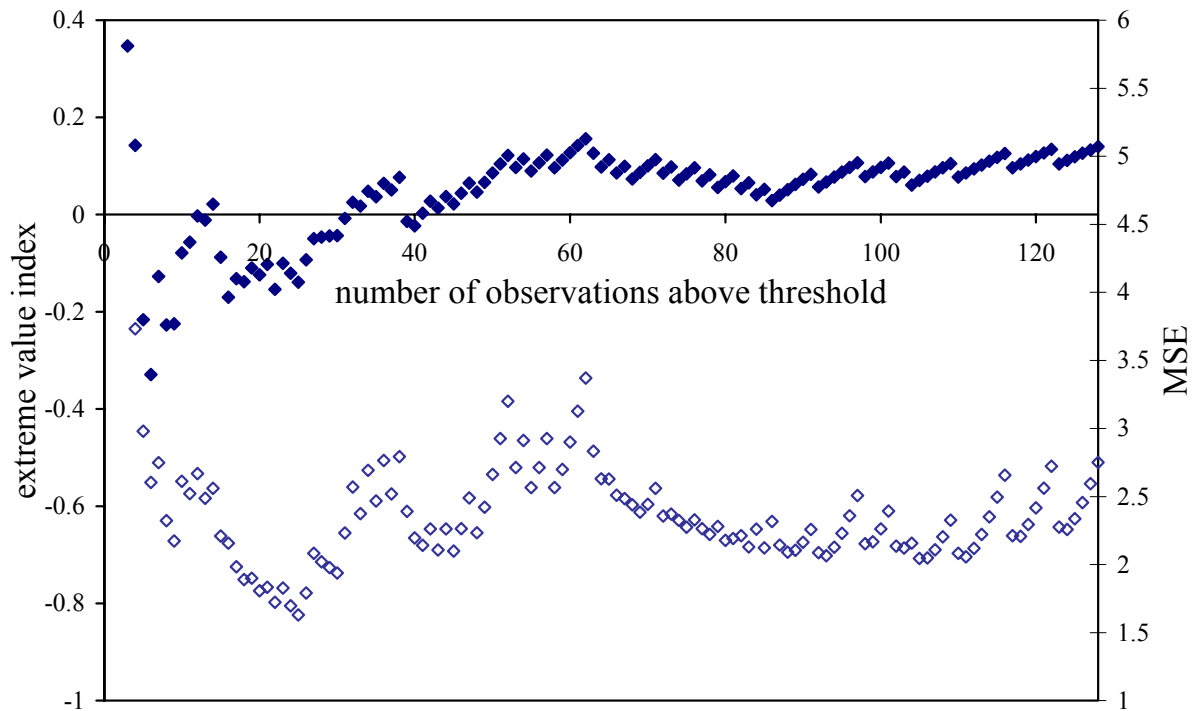


Figure 4.a. left vertical axis (◆): UH-estimation of the extreme value index; right vertical axis (◇): mean squared error of weighted linear regression in the UH-plot; 10 min averaged rainfall intensities at Uccle.

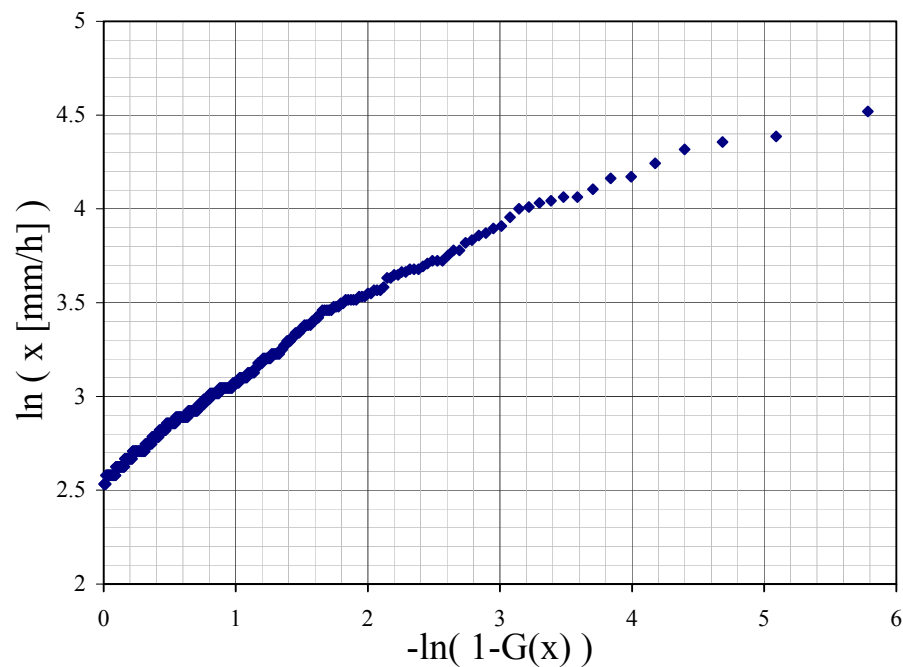


Figure 4.b. Pareto quantile plot; 10 min averaged rainfall intensities  $x$  at Uccle.

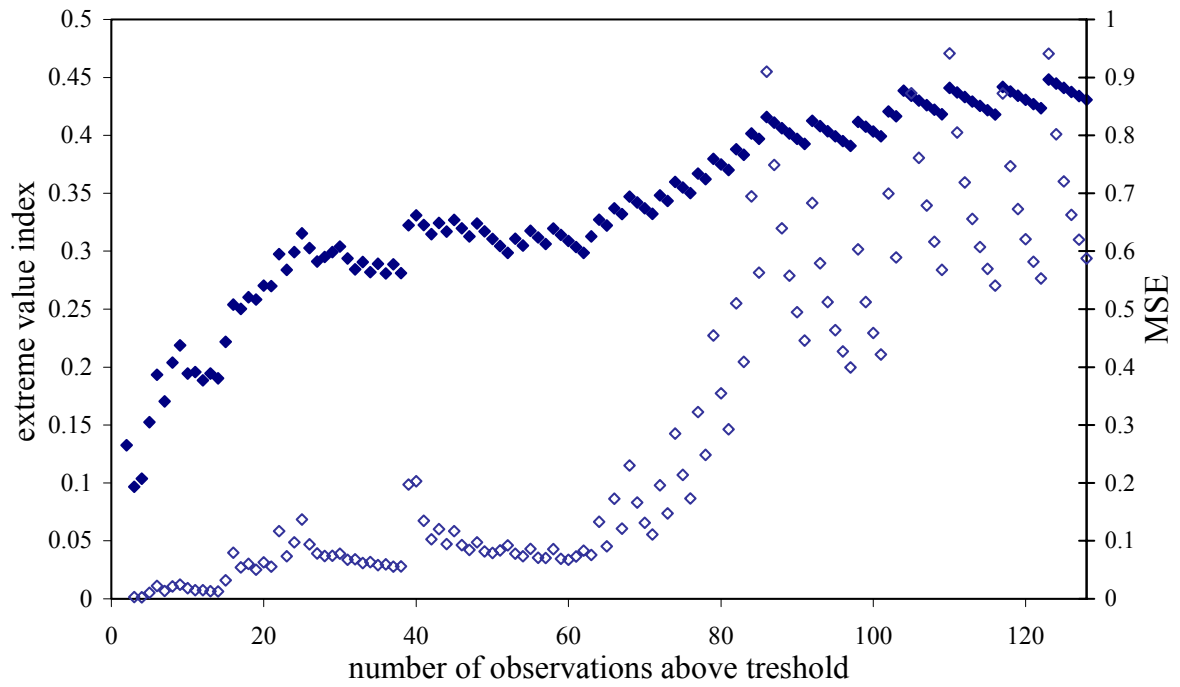


Figure 4.c. left vertical axis (♦): Hill-estimation of the extreme value index; right vertical axis (◊): mean squared error of weighted linear regression in the Pareto quantile plot; 10 min averaged rainfall intensities at Uccle.

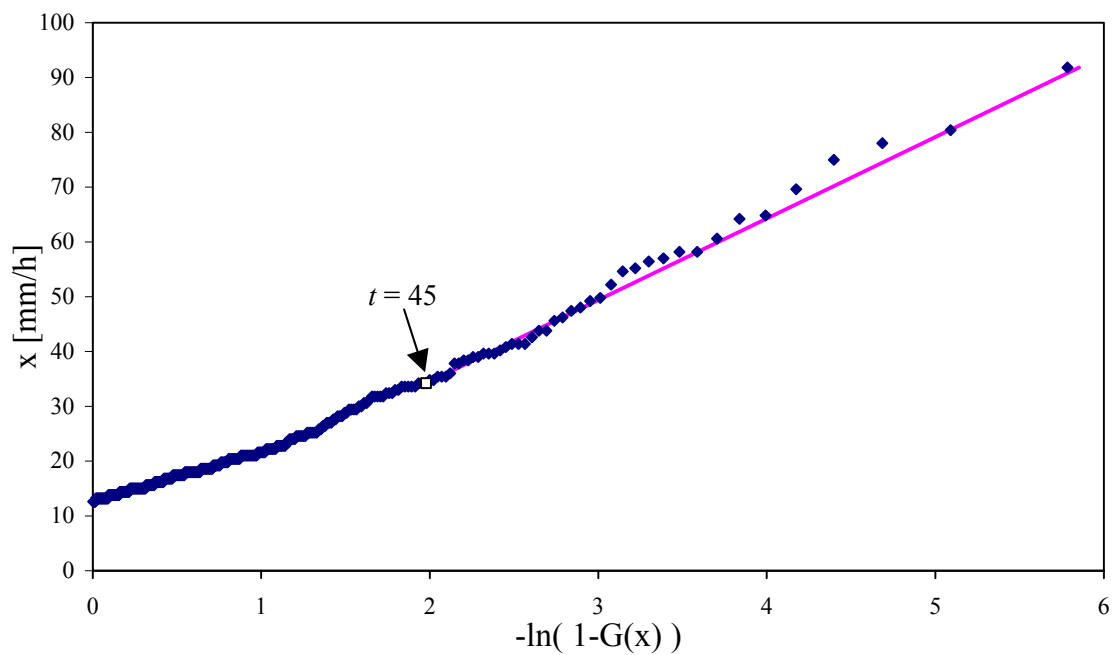
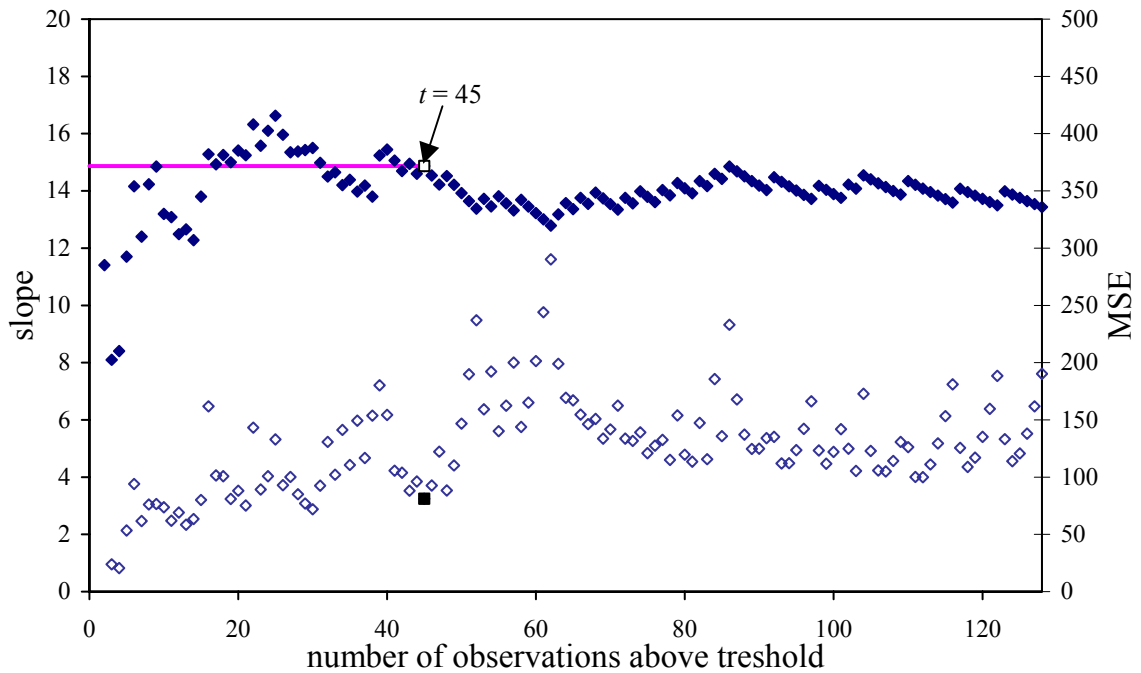


Figure 4.d. Exponential quantile plot; 10 min averaged rainfall intensities  $x$  at Uccle.



**Figure 4.e.** left vertical axis (◆): Hill-type estimation slope in the exponential quantile plot; right vertical axis (◇): mean squared error of weighted linear regression in the exponential quantile plot; 10 min averaged rainfall intensities at Uccle.

#### *Aggregation-level of 5 days*

Also for the aggregation-level of 5 days,  $\gamma$  is not significantly different from zero. Based on a Hill-type regression in the exponential quantile plot (Figure 5.a), the MSE is minimized at the optimal threshold  $t=112$  (Figure 5.b). At the high threshold-levels, there is a large fluctuation of the parameter values, because of the statistical uncertainty; only a few observations can be used for estimation of the parameter. At lower threshold-levels ( $t > 170$ ), the parameter values start to deviate from the earlier values. They are continuously increasing with increasing  $t$ . This behaviour is caused by the deviation of the points from the linear path in the quantile plot of Figure 5.a. In the intermediate range ( $30 < t < 170$ ), the parameter values remain nearly constant. This is the range in which the optimal threshold is situated. In that range, the MSE reaches local minima at three threshold orders:  $t=53$ ,  $t=112$  and  $t=150$ , and an absolute minimum for  $t=112$ .

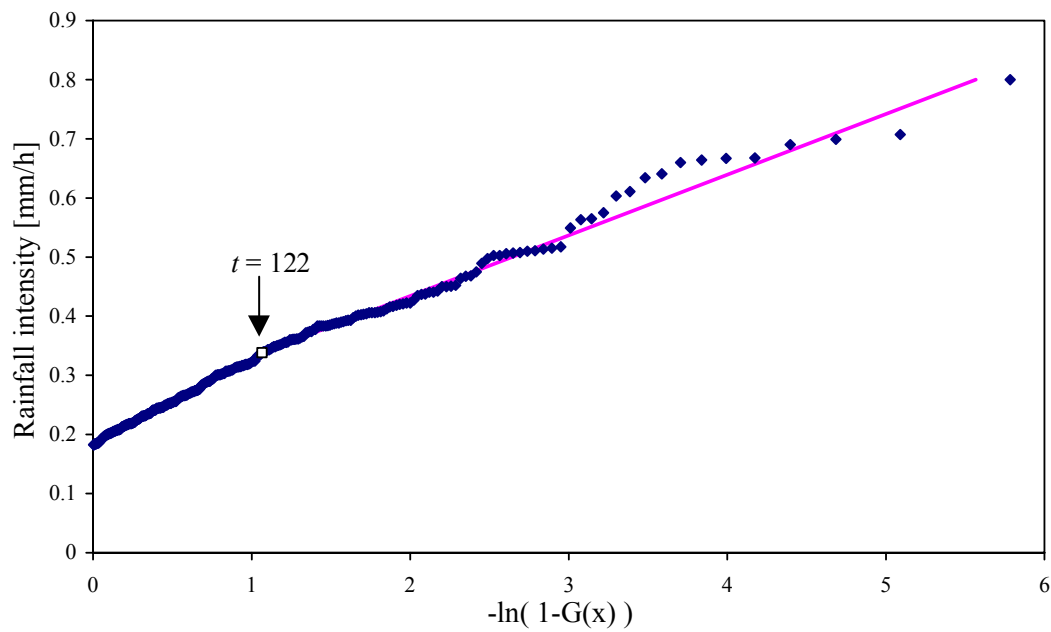


Figure 5.a. Exponential quantile plot; 5 days averaged rainfall intensities at Uccle.

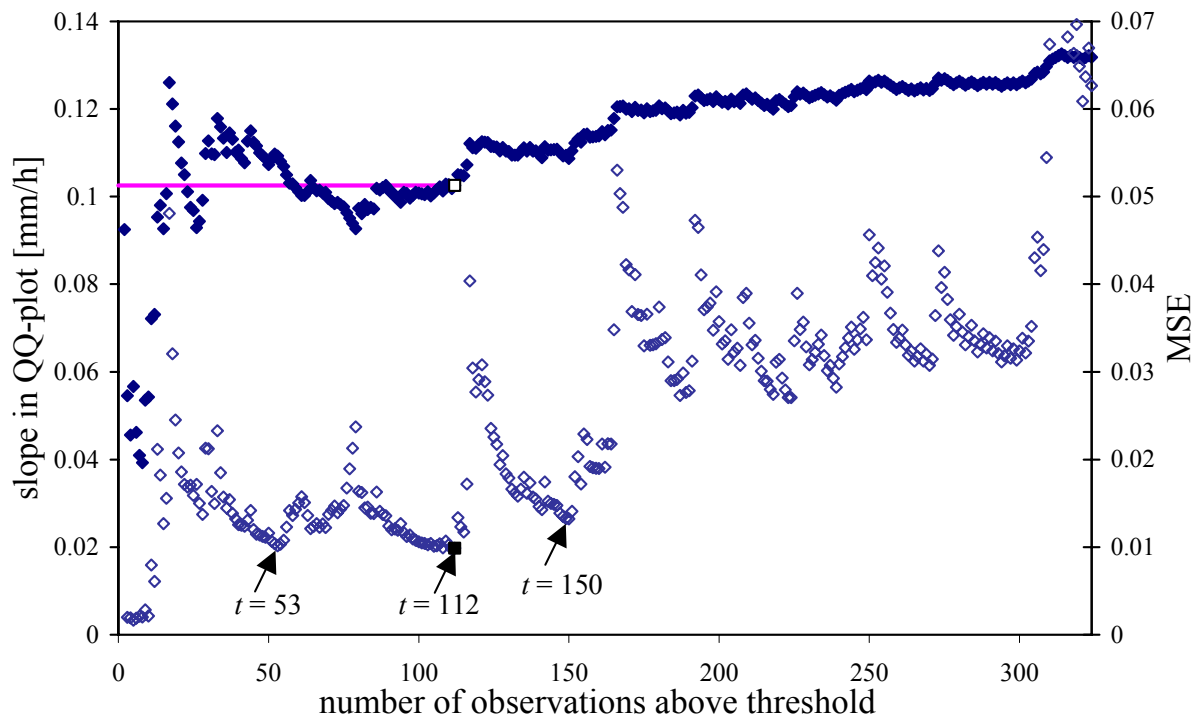


Figure 5.b. left vertical axis ( $\blacklozenge$ ): Hill-type estimation slope in the exponential quantile plot; right vertical axis ( $\diamond$ ): mean squared error of weighted linear regression in the exponential quantile plot; 5 days averaged rainfall intensities at Uccle.



## 7. Low flow frequency analysis

For low flow frequency analysis, the same concepts as explained above are applicable. A transformation is however needed to transfer low values (minima) into high values (maxima). The transformation  $Y = 1/X$  is used by preference. Another useful transformation is  $Y = -X$ .

The discharge minima are selected during periods where the discharge approximately reaches the baseflow value. Two low flow periods then can be selected as independent when they are separated by a time length of at least the recession constant of groundwater runoff, and when there is at least one high flow period in between. In a practical way, the POT high flows are first selected (using the recession constant of baseflow as independency period) whereafter independent low flow values are calculated as the minimum discharges in between two successive high flow POT peaks.

The following remarks, however, have to be taken into account for the low flow analysis:

- For the low flow discharges, there is a lower limit for the discharges  $X$ . Flow discharges indeed cannot be negative. After transformation  $Y = 1/X$ , the lower limit  $X=0$  for  $X$  becomes an upper limit for  $Y$  at  $+\infty$ . Using the transformation  $Y = -X$ ,  $Y$  has an upper limit at 0. This means that, in this last case, the probability distribution of  $Y$  cannot have a heavy or normal tail. The tail has to be light (due to the upper limit).
- For low flow discharges, Weibull or Fréchet distributions most often appear as low flow extreme value distributions. The Fréchet distribution has the following equation:

$$P[X \leq x] = \exp\left(-\frac{x^{-\tau}}{\beta}\right)$$

while the Weibull-verdeling has the following shape (lower limit at  $x_l=0$ ):

$$P[X \leq x] = 1 - \exp\left(-\frac{x^\tau}{\beta}\right)$$

It is the left tail of this Weibull distribution (the tail towards the lower values; approaching zero) which is used for the low flow extreme value distribution. For flood frequency applications, the upper tail of the distribution is used. As explained above, this upper tail is of the normal type (extreme value index). The left tail is, however, limited and thus belongs to the light tail class (negative extreme value index).

The validity of the Fréchet distribution can be checked in the Weibull Q-Q plot after transformation  $Y = 1/X$ . The slope of a linear asymptotic tail in this Weibull Q-Q plot equals the parameter  $\tau$  of the Fréchet distribution. For the special case when  $\tau=1$ , also a linear tail behaviour can be observed in the exponential Q-Q plot for  $Y = 1/X$ . This will be proved hereafter mathematically.

The validity of the Weibull distribution has to be checked in the Weibull Q-Q plot after transformation  $Y=-X$ .

*Normal tail for  $Y = 1/X$* 

When an exponential tail is observed for  $Y$ , then an exponential extreme value distribution can be calibrated above a specific threshold  $y_t$  :

$$G(y) = P[Y \leq y | Y \geq y_t] = 1 - \exp\left(-\frac{y - y_t}{\beta}\right)$$

This equation can be transferred as follows towards a distribution for  $X$ : (using  $x_t = 1 / y_t$ ):

$$P[X \leq x | X \leq x_t] = P[Y \geq y | Y \geq y_t] = 1 - P[Y \leq y | Y \geq y_t] = \exp\left(-\frac{y - y_t}{\beta}\right) = \exp\left(-\frac{x^{-1} - x_t^{-1}}{\beta}\right)$$

or:

$$G(x) = P[X \leq x | X \leq x_t] = \exp\left(-\frac{x^{-1} - x_t^{-1}}{\beta}\right) = \frac{\exp\left(-\frac{x^{-1}}{\beta}\right)}{\exp\left(-\frac{x_t^{-1}}{\beta}\right)}$$

This last equation matches the Fréchet distribution for  $\tau=1$  (conditional distribution for values lower than  $x_t$ ).

For low flow minima, the return period  $T$  has to be calculated on the basis of the non-exceedance probability  $G(x)$  (and not based on the exceedance probability  $1 - G(x)$  as done for the maxima):

$$T[\text{years}] = \frac{n}{t} \frac{1}{G(x)} = \frac{n}{t} \frac{\exp\left(-\frac{x_t^{-1}}{\beta}\right)}{\exp\left(-\frac{x^{-1}}{\beta}\right)}$$

This equation for  $T$  is valid when  $t$  minima are considered below the threshold  $x_t$  during  $n$  years.

*Weibull tail for  $Y = 1/X$* 

In case of a Weibull tail for  $Y = 1/X$  the more general Fréchet distribution is found for  $X$ :

$$F(y) = P[Y \leq y] = 1 - \exp\left(-\frac{y^\tau}{\beta}\right)$$

$$F(x) = P[X \leq x] = \exp\left(-\frac{x^{-\tau}}{\beta}\right)$$

where:  $\beta = \frac{x_t^{-\tau}}{c}$

and  $c = -\ln\left(\frac{t}{m}\right)$  in case the analysis is based on  $m$  independent low flow minima.

The return period is given by:

$$T[\text{years}] = \frac{n}{t} \frac{1}{G(x)} = \frac{n}{t \exp\left(\frac{x^{-\tau}}{\beta}\right)} \frac{1}{\exp\left(-\frac{x^{-\tau}}{\beta}\right)}$$

#### *Consideration of zeros in low frequency analysis*

In the case of zeros present in the selection of independent low flows from the available series, the transformation  $1/Q$  cannot be applied. A bi-model probability model is needed, where a split is made between the zero low flow values and the non-zero low flow values:

$$F(y) = P[Y \leq y] = P[Y = 0] + P[Y \leq y | Y > 0] P[Y > 0]$$

The number of zero low flow values over the total number of selected independent low flow values gives an empirical estimate of  $P[Y = 0]$ . The number of non-zeros over the total number of low flow values can be used as empirical estimate for  $P[Y > 0]$ . The following relation applies between  $P[Y = 0]$  and  $P[Y > 0]$ :

$$P[Y = 0] + P[Y > 0] = 1$$

Based on the non-zero low flow values, the low flow extreme value analysis as presented above can be applied and the low flow frequency distribution calibrated. This distribution equals the conditional low flow frequency distribution for  $P[Y \leq y | Y > 0]$  (the distribution under the condition of positive low flow values:  $Y > 0$ ).

#### *Analysis of dry spells*

Dry spells can be analysed along the same lines as described above for the calibration of the distribution of maxima (see flood frequency analysis).

## ECQ: Short User's Manual

### Steps to follow:

- Input: Selected independent extremes to be pasted in ranked order (highest value first) in the column B.
- Optional in Input: Specify the name of the dataset in Cell E3.
- Main: Evaluate the results based on the UH plot, the exponential Q-Q plot, the Pareto Q-Q plot (and optional: the Weibull Q-Q plot):
  - Specify the optional parameters:
    - No bias (0) or bias (1) correction to the slope calculation in the selected Q-Q plot (select 0 when no use is made of this option); for 1 select also the  $\rho$  value
    - Weighting factors: 0 for weighting factors 1 and 0 for the Hill weighting factors (by default, the Hill factors are used)
    - The type of regression in the Q-Q plot to be considered: 0 for an unconstrained regression (non-fixed threshold), 1 for a constrained regression (fixed threshold)
    - When censoring of the highest values has to be considered, specify the number of values to be censored (the censoring threshold rank) (select 0 when no censoring is considered)
  - Select the type of Q-Q plot to be evaluated and press the “Calculate” button to calculate the Q-Q plot, to conduct the regression and to calculate the MSE of the regression line, for all possible threshold ranks
  - Look at the Q-Q plot (button “Q-Q plot”); and at the path of the slope and the MSE of the regression line for increasing threshold ranks (button “Slope Q-Q plots”); and evaluate the linearity of the points in the tail of the distribution.
  - Repeat this procedure for the different Q-Q plots
- Make a decision on the sign of the extreme value index (the distribution type) as calibration result (1 when selecting a normal tail (exponential or Gumbel distribution), 2 when selecting a heavy tail (GPD or GEV distribution with  $\gamma > 0$ ), 3 when selecting the Weibull class.
- Select the optimal threshold based on the Chart “Slope Q-Q plot”; and specify the rank number of this optimal threshold in ‘Selection optimal threshold’. For this selected threshold:

- the estimated slope in the Q-Q plot is shown in ‘Slope Q-Q plot at selected threshold’; and also graphically shown by the line in the Chart “Slope Q-Q plot”
  - the calibrated extreme value distribution is shown as a regression line in Chart “Q-Q plot”
  - the calibrated parameter values of this extreme value distribution are shown as ‘Calibration results’
- 
- Optional: the UH-estimator of the extreme value distribution can be derived by specifying the optimal threshold for this estimator on the basis of the UH-plot. Specify the rank number of this optimal threshold in ‘Selection optimal threshold’. For this selected threshold, the estimated extreme value index is shown in ‘Slope Q-Q plot at selected threshold’; and also graphically shown by the line in the Chart “Slope UH-plot”.