Flanders
State of
the Art

00_144_1
FHR reports

Scientific Assistance towards a Probabilistic Formulation of Hydraulic Boundary Conditions

Extreme Value Analysis Reference Guide

DEPARTMENT
MOBILITY &
PUBLIC
WORKS

www.flandershydraulicsresearch.be

Scientific Assistance towards a Probabilistic Formulation of Hydraulic Boundary Conditions

Extreme Value Analysis Reference Guide

Leyssen, G.; Blanckaert, J.; Pereira, F.; Nossent, J.; Mostaert, F.

Flanders Hydraulics Research
Flanders State of the Art

Legal notice

Copyright and citation

D/2017/3241/26

This publication should be cited as follows:

Leyssen, G.; Blanckaert, J.; Pereira, F.; Nossent, J.; Mostaert, F. (2019). Scientific Assistance towards a Probabilistic Formulation of Hydraulic Boundary Conditions: Extreme Value Analysis Reference Guide. Version 3.0. FHR Reports, 00_144_1. Flanders Hydraulics Research & IMDC: Antwerp.

Document identification

Customer:	Flanders Hydraulics Research	Ref.:	WL2019R00_144_1
Keywords (3-5):	Extreme Value; Flooding; Floodrisk		
Text (p.):	16	Appendices (p.):	12
Confidentiality:	☒ No	☒ Available online	

Author(s):	Leyssen, G.; Blanckaert, J.

Control

	Name	Signature
Reviser(s):	Pereira, F.; Nossent, J.	Getekend door: Fernando Pereira (Signature) Getekend op: 2019-11-05 14:03:13 +01:00 Reden: Ik keur dit document goed; Fernando Pereira Getekend door: Jiri Nossent (Signature) Getekend op: 2019-11-05 13:57:46 +01:00 Reden: Ik keur dit document goed; Jiri Nossent
Project leader:	Pereira, F.	Getekend door: Fernando Pereira (Signature) Getekend op: 2019-11-05 14:03:34 +01:00 Reden: Ik keur dit document goed; Fernando Pereira

Approval

Head of Division:	Mostaert, F.	Getekend door: Frank Mostaert (Signature) Getekend op: 2019-11-05 15:14:29 +01:00 Reden: Ik keur dit document goed; Frank Mostaert

Abstract

This report gives an extensive overview of the univariate extreme value theory including all the formulae necessary to carry out an extreme value analysis. This report is the base for a software tool developed for extreme value analysis.

The difference between the marginal and conditional distribution are highlighted as an introduction. The parameter estimation techniques of the maximum likelihood theory are discussed and selected for further implementation. Some techniques that help in the selection of the appropriate distribution are explained. Finally the changes between the updated and the previous methodology are discussed. Formulae of selected distributions are summarized in the annexes.

Contents

Abstract III
Contents V
List of tables VII
List of figures VIII
1 introduction 1
1.1 Scope of contract 1
1.2 Overview 1
1.3 This report 2
2 Extreme value statistics: an overview 3
2.1 Introduction 3
2.2 Marginal distributions: theory 3
2.2.1 Generalized extreme value distribution (GEV) 3
2.3 Conditional distributions: theory 6
2.3.1 Generalized Pareto distribution (GPD) 6
3 Parameter estimation 8
3.1 Introduction 8
3.2 Empirical 8
3.3 Maximum likelihood estimation 8
3.3.1 GEV distribution 9
3.3.2 GPD distribution 9
4 Selection of appropriate conditional distribution 11
4.1 Excess functions 11
4.2 RMSE as a parameter of the goodness of fit 12
4.3 Confidence intervals of the parameters 13
4.4 Visual control of the return level plot 13
4.5 Check of the Poisson process 13
5 Changes in the methodology 14
5.1 Parameters estimation by the MLE method 14
5.2 Weight factor 14
6 References 16

Appendix A: Extreme value model formulae A1
GEV A1
Gumbel distribution A1
GPD (y=x-u) A2
Exponential distribution (conditional form y=x-u) A2
Pareto distribution (conditional form y=x/u) A3
Conditional Weibull distribution (y=x-u) A3
Appendix B: Validation of conditional distributions A4
GPD A4
Exponential distribution A5
Pareto distribution A5
Conditional Weibull distribution A5
Appendix C: Derivation of the return level A6
GEV A6
Gumbel distribution A6
GPD A7
Exponential distribution A8
Pareto distribution A9
Conditional Weibull distribution A10
Appendix D: Maximum likelihood estimations A11
Gumbel distribution A11
Exponential distribution A11
Pareto distribution A12
Conditional Weibull distribution A12

List of tables

Table 4-1: Mean excess functions (Beirlant 2004) 11

List of figures

Figure 1-1: Overview of reports, tools and manuals 1
Figure 2-1: GEV with parameters (μ, σ, ξ)= (3.87, 0.198, -0.05) 4
Figure 2-2: Tail behaviour of GEV with different values of ξ 5
Figure 2-3: The three marginal extreme value families 5
Figure 2-4: GPD with ξ=0.184, σ=7.44 and u=30. 7
Figure 4-1: Mean excess functions 12
Figure 5-1: Weight factor of the standard methodology 14

1 introduction

1.1 Scope of contract

Flanders Hydraulic Research (FHR) has commissioned IMDC NV to adapt its standardized methodology for rendering composite hydrographs, developed at KU Leuven (Willems 2001 & 2002), to recent evolutions (e.g. climate change), updated data series (e.g. recent measurements) and diversifying applications (e.g. coastal zone, flood risk calculations,…).

The project team consists of Sarah Doorme (advisor), Gert Leyssen (Jr. Eng.), Lorens Coorevits (Jr. Eng.) and Joris Blanckaert (Sr. Eng. and project manager for IMDC). On behalf of FHR, Eng. Fernando Pereira is in charge of the general supervision of the project. Eng. Toon Verwaest of FHR ministers scientific support towards coastal zone applications.

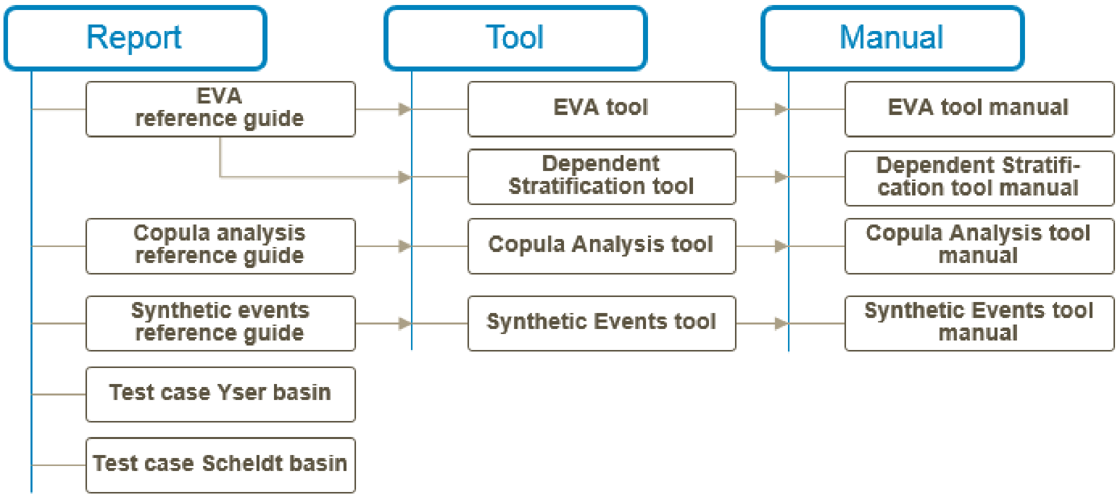
1.2 Overview

A new methodology is presented, which is based on extended literature review and expertise of the project team members. The methodology is described in a set of technical reports and is implemented in a suite of software tools for use in flood risk analysis and probabilistic design projects. The Graphical User Interfaces of the software tools are described in a set of manuals.

The new methodology is tested within two representative test cases, i.e. for the Yzer basin and the Scheldt basin (navigable waterways in Flanders). The test cases are described in two reports.

Figure 1-1 presents an overview of the reports, tools and manuals.

Figure 1-1: Overview of reports, tools and manuals



1.3 This report

This report provides an overview of the univariate extreme value theory which consists of a combination of the standardized methodology for Extreme Value Analysis, developed at KU Leuven (Willems 2001 & 2002), international standard literature (Coles, 2001, Kotz, 2000, Nelsen, 2004) and Beirlant's masterpiece (Beirlant, 2004). This overview is used to implement a software tool for extreme value analysis (see Figure 1-1). To keep the text readable a lot of the derivations are attached in the different annexes.

The 2st chapter gives an introduction in the univariate extreme value theory with its marginal and conditional domain. The 3nd chapter deals with parameter estimation by the maximum likelihood methodology. The 4th chapter contains some methods to select the correct distribution for a set of observation. Finally the 5th chapter gives an overview of the most important changes in the new methodology.

2 Extreme value statistics: an overview

2.1 Introduction

Extreme value statistics is unique as a statistical discipline in that it develops techniques and models for describing the unusual rather than the usual. Extreme values are by definition scarce which implies that estimates are required for values that are much greater (or smaller) than values that are already observed. Extreme value theory provides a number of models for the extrapolation to these extreme values combined with a number of tools to choose the correct model for the specific phenomena.

A distinction can be made between the classic extreme value theory with marginal distributions based on block maxima and the threshold extreme value theory with conditional distributions based on POT values. It should be noted that every marginal distribution has a conditional counterpart and vice versa.

2.2 Marginal distributions: theory

2.2.1 Generalized extreme value distribution (GEV)

The classic extreme value theory focuses on the statistical behavior of block maxima (M)

$$M_n = max\{X_1, \dots, X_n\}$$

A linear renormalization of the variable M_n (set of block maxima) by appropriate choices of the constants a_n and b_n yields M^*_n.

$$M_n^* = \frac{M_n - b_n}{a_n} \; with \; a_n > 0$$

It is stated that if there exist sequences of constants { a_n >0} and { b_n } such that

$$Pr\{M_n^* < z\} \rightarrow G(z) \; as \; n \rightarrow \infty$$

For a non-degenerate distribution function G, then G is a member of the GEV family.

$$G(z) = exp\left\{-\left[1 + \xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-1/\xi}\right\} \qquad 1$$

Defined by $\left\{z : 1 + \frac{\xi(z-\mu)}{\sigma} > 0\right\}$, where $-\infty<\mu<\infty$, $\sigma>0$ en $-\infty<\xi<\infty$. In subset of GEV with $\xi=0$ is interpreted as the limit of the above distribution as $\xi \rightarrow 0$:

$$G(z) = exp\left\{-\left[-\left(\frac{z-\mu}{\sigma}\right)\right]\right\}, \quad -\infty < z < \infty \qquad 2$$

The problem that the constants a_n and b_n are unknown, can be easily resolved by:

$$Pr\left\{\frac{M_n - b_n}{a_n} < z\right\} \approx G(z) \; as \; n \rightarrow \infty$$

$$Pr\{M_n < z\} \approx G\left(\frac{z - b_n}{a_n}\right) \; as \; n \rightarrow \infty$$

$$Pr\{M_n < z\} \approx G^*(z) \; as \; n \rightarrow \infty$$

Where G* is another member of the GEV family, with a different set of model parameters. The model parameters of the GEV distribution can be estimated by fitting the distribution through a set of block maxima obtained from a set of independent observations.

The GEV distribution is determined by three parameters:

- ξ = shape parameter
- μ = location parameter
- σ = scale parameter

By inverting equation 1 the return levels corresponding to the probabilities p are given by (see Annex C):

$$z_p = \begin{cases} \mu - \frac{\sigma}{\xi}\left[1 - \{-log(1-p)\}^{-\xi}\right] & if\ \xi \neq 0 \\ \mu - \sigma log\{-log(1-p)\} & if\ \xi = 0 \end{cases} \qquad 3$$

Where zp is the return level, p the exceedance frequency and 1/p is the return period which implies that zp is exceeded by a random block maximum with a frequency p (Coles 2001; Kotz 2002).

To obtain a better insight in the behavior of the GEV distribution an example is displayed in Figure 2-1. The tail behavior is illustrated in Figure 2-2. The four formulae used in Figure 2-1 are summarized in annex A.1.

Figure 2-1: GEV with parameters (μ, σ, ξ)= (3.87, 0.198, -0.05)

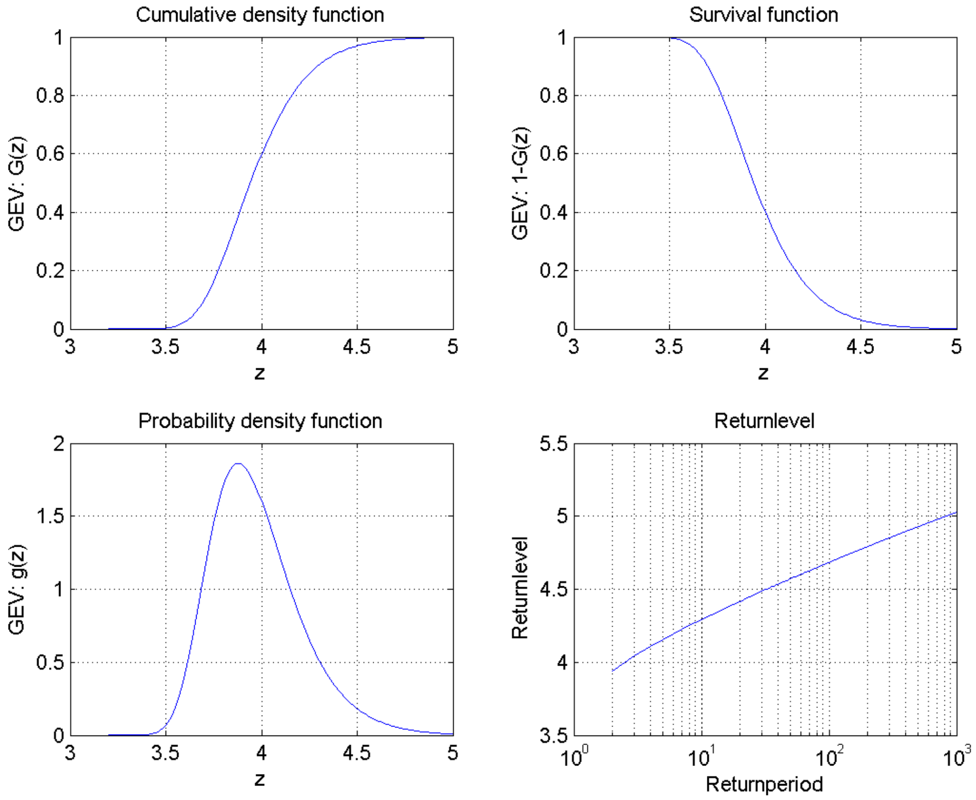
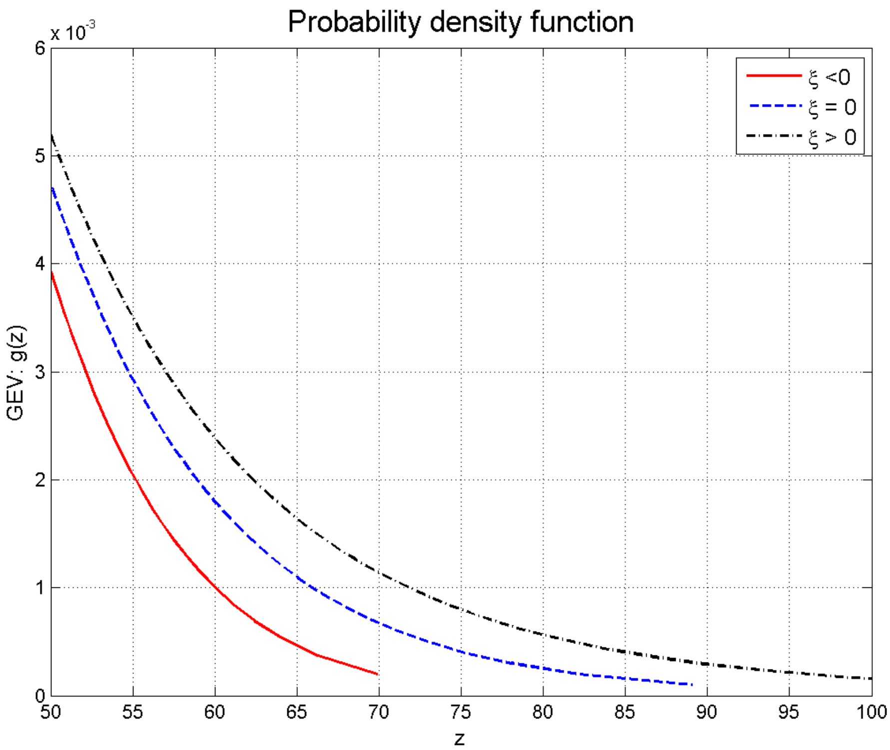
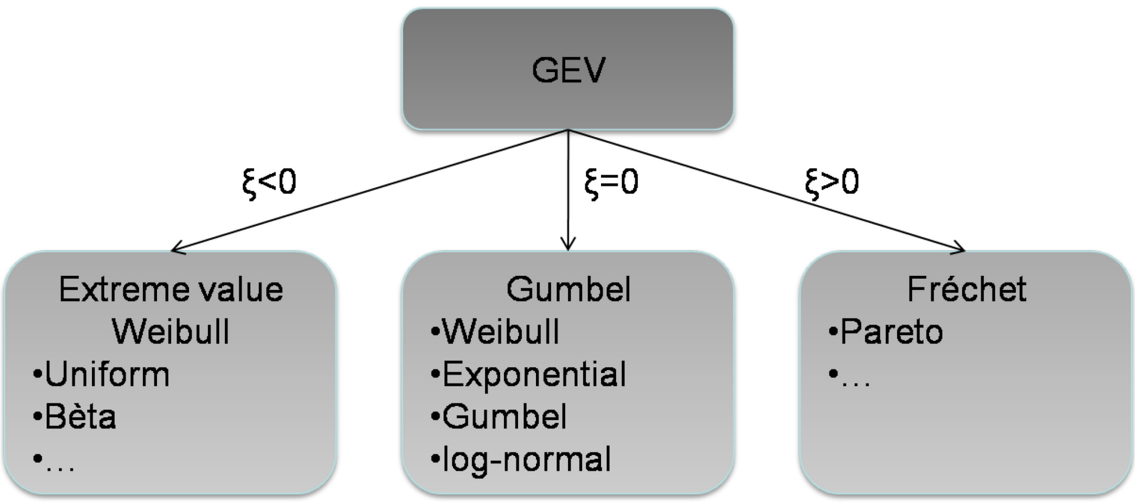


Figure 2-2: Tail behaviour of GEV with different values of ξ



The Generalized Extreme Value distribution can be divided in three classes, as shown in Figure 2-3 and the formula below, based on the weight of the tail ξ. If the distribution has a light tail, ξ<0, the distribution is part of the extreme value Weibull domain. The distributions of these families have an upper boundary z_+. If ξ=0, the distribution belongs to the Gumbel domain and they decrease exponentially towards infinity. The third family is the Fréchet family. These distributions have a heavy tail which decrease polynomially towards infinity (Coles 2001).

Figure 2-3: The three marginal extreme value families



$$\begin{cases} III: EV\ Weibull\ (\xi < 0) & G(z) = \begin{cases} exp\left\{-\left[-\left(\frac{z-\mu}{\sigma}\right)^{\xi}\right]\right\} & ,z < \mu \\ 1 & ,z \geq \mu \end{cases} \\ II: Gumbel\ (\xi = 0) & G(z) = \exp\left[-exp\left\{-\left(\frac{z-\mu}{\sigma}\right)\right\}\right] \quad ,-\infty < z < \infty \\ I: Fréchet\ (\xi > 0) & G(z) = \begin{cases} 0 & ,z \leq \mu \\ exp\left\{-\left(\frac{z-\mu}{\sigma}\right)^{-\xi}\right\} & ,z > \mu \end{cases} \end{cases}$$

2.3 Conditional distributions: theory

The use of block maxima and the corresponding marginal distribution can be a wasteful way to use data. If continuous time series of hourly or daily observations are available the data can be used more efficient by abandoning the blocking procedure. Let X_1, X_2 , …X_n be a sequence of independent and identically distributed random variables, having a marginal distribution function F. We are interested in extreme events, i.e. the events with X_i that exceed some high threshold u. The description of the stochastic behavior of these extreme events is given by the conditional probability:

$$Pr\{X > u + y | X > u\} = \frac{1 - G(u+y)}{1 - G(u)} \quad ,y > 0 \quad ,y = x - u \qquad 4$$

This formula allows us to transform any marginal distribution to its conditional counterpart (Coles 2001).

2.3.1 Generalized Pareto distribution (GPD)

If block maxima have an approximate marginal distribution G, part of the GEV distribution, then the threshold excesses have a corresponding approximate distribution within the Generalized Pareto (GPD) family. The parameters of the GPD are uniquely determined by the associated GEV distribution of the block maxima. This statement can be validated by applying eq. 1 in eq. 4 (see derivation in Annex B.1). The survival function of the GPD is given by:

$$Pr\{X > u + y | X > u\} = 1 - G(y) = 1 - \left[1 + \frac{\xi y}{\tilde{\sigma}}\right]^{-1/\xi} \quad with\ y > 0\,, (1 + \frac{\xi y}{\tilde{\sigma}}) > 0 \qquad 5$$

Where $\tilde{\sigma}$ is uniquely determined by the parameters of the corresponding GEV distribution and the shape parameters are equal.

$$\tilde{\sigma} = \sigma + \xi(u - \mu)$$

An extra advantage is that where the parameters σ and μ of a GEV distribution will change by a change of the block size n, the parameter $\tilde{\sigma}$ of the corresponding GPD is unperturbed by the changes of σ and μ which are self-compensating (Coles 2001).

Similar to the GEV distribution, the shape parameter ξ is dominant in determining the behavior of the GPD. If ξ<0 the distribution has an upper bound of $u - \tilde{\sigma}/\xi$. If ξ>0 as well as in case ξ=0 the distribution has no upper limit. In the latter case eq. 5 can be written:

$$G(y) = 1 - exp\left(\frac{-y}{\tilde{\sigma}}\right) \qquad 6$$

Further on in this report the parameter $\tilde{\sigma}$ of the GDP will be written as σ. Similar to the derivation of the GPD from the GEV, the conditional forms of the different families can be derived from the marginal form (Coles 2001) as shown in Annex B.

The return levels are given by (see Annex C):

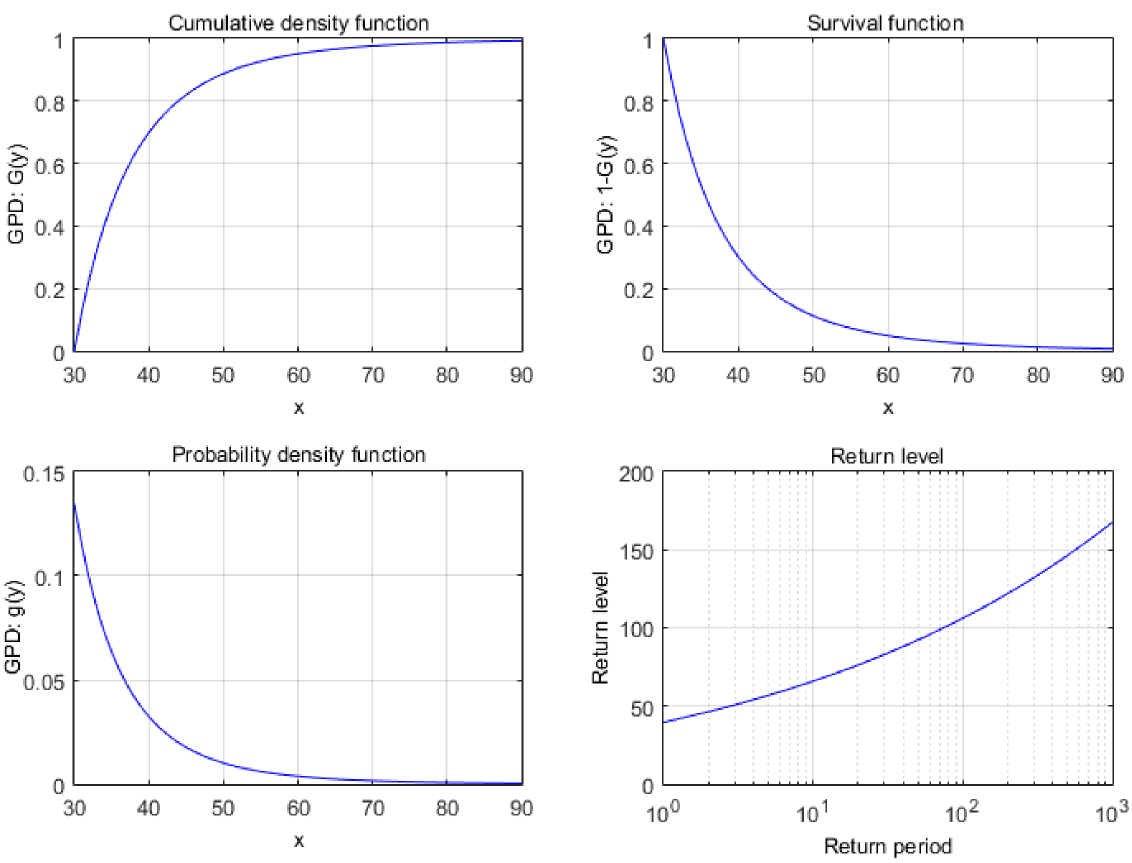
$$x_m = \begin{cases} u + \frac{\sigma}{\xi}\left(\left(\frac{T * n_y * k}{n}\right)^{\xi} - 1\right) & if\ \xi \neq 0 \\ u + \sigma log\left(\frac{T * n_y * k}{n}\right) & if\ \xi = 0 \end{cases} \quad 7$$

Where ny is the number of observations in a year, n the total amount of observations, k the number of observations that exceed the threshold u and T=1/p the return level. The number of years of data is in most cases used as an estimator for $\frac{n}{n_y}$. This yields:

$$x_m = \begin{cases} u + \frac{\sigma}{\xi}\left(\left(\frac{T * k}{A}\right)^{\xi} - 1\right) & if\ \xi \neq 0 \\ u + \sigma log\left(\frac{T * k}{A}\right) & if\ \xi = 0 \end{cases} \quad 8$$

To get a better understanding in the behavior of the GPD an example if displayed in Figure 2-4.

Figure 2-4: GPD with ξ=0.184, $\tilde{\sigma}$=7.44 and u=30.



3 Parameter estimation

3.1 Introduction

The parameter estimation of the extreme value distributions is done by the maximum likelihood estimation. The results of the obtained distribution can be evaluated by comparison with the empirical distribution.

3.2 Empirical

The empirical probability of exceedance of a set of observations, in this case block maxima or POT values, can be found by sorting the observation from large to small

$$z_1 > z_2 > \cdots > z_n$$

The exceedance probability (empirical cumulative density function) is given by:

$$\left(z_i, \frac{i}{n+1}\right)$$

3.3 Maximum likelihood estimation

The parameter values of the distributions can be determined by Maximum Likelihood Estimation (MLE). This method is a well-known statistical method used for fitting a statistical model to data and providing estimates for the model's parameters. Under the assumption that z1, …, zn are n independent random variables with a probability density function, g(z), with parameters θ, the likelihood function is given by

$$L(\theta) = \prod_{i=1}^{n} g(z_i; \theta) \qquad 9$$

In practice it is more convenient to work with the logarithm of the likelihood function, the so called log-likelihood (Beirlant, 2004).

$$L(\theta) = \sum_{i=1}^{n} log(g(z_i; \theta)) \qquad 10$$

Maximization of the log-likelihood in function of the parameters θ yields in the maximum likelihood estimator θMLE. This estimator has an approximate normal distribution which can be used to estimate a confidence interval by means of the variance-covariance matrix.

These confidence intervals can be used in the estimation of confidence intervals for the return level (zm). The variance of this return level can be calculated by the delta method (with T the transpose of the matrix).

$$Var(z_m) \approx \nabla {z_m}^T * V * \nabla z_m, \quad with\ \nabla {z_m}^T = \frac{\partial z_m}{\partial \theta}$$

Where V is the variance-covariance matrix of θMLE (Coles 2001).

The maximum likelihood functions of the GEV and the GPD distributions are evaluated in the next subsections. The equations for the conditional Weibull distribution, the exponential distribution, the Pareto distribution and the Gumbel distribution are shown in Annex D.

3.3.1 GEV distribution

In case of a GEV distribution the log-likelihood if ξ≠0 is given by (Coles 2001):

$$l(\mu,\sigma,\xi) = -n * log(\sigma) - \left(1+\frac{1}{\xi}\right)\sum_{i=1}^{n} log\left(1+\xi\left(\frac{z_i-\mu}{\sigma}\right)\right) - \sum_{i=1}^{n}\left(1+\xi\left(\frac{z_i-\mu}{\sigma}\right)\right)^{-1/\xi} \qquad 11$$

Provided that $\left\{z : 1+\frac{\xi(z-\mu)}{\sigma} > 0\right\}$. If ξ=0 this equation becomes

$$l(\mu,\sigma) = -n * log(\sigma) - \sum_{i=1}^{n}\left(\frac{z_i-\mu}{\sigma}\right) - \sum_{i=1}^{n} exp\left(-\left(\frac{z_i-\mu}{\sigma}\right)\right) \qquad 12$$

The obtained parameters can be used to calculate the maximum likelihood return level.

$$z_m = \begin{cases} \mu - \frac{\sigma}{\xi}\left[1-\{-log(1-p)\}^{-\xi}\right] & ,\xi \neq 0 \\ \mu - \sigma log\{-log(1-p)\} & ,\xi = 0 \end{cases}$$

The variance estimated by the delta method is given by:

$$Var(z_m) \approx \nabla {z_m}^{\mathrm{T}} * \mathrm{V} * \nabla z_m$$

Where V is the variance-covariance matrix of (μ,σ,ξ) and

$$\nabla {z_m}^{\mathrm{T}} = \left[\frac{\partial \mathrm{z_m}}{\partial \mu}, \frac{\partial \mathrm{z_m}}{\partial \sigma}, \frac{\partial \mathrm{z_m}}{\partial \xi}\right] = \left[1, \quad -\frac{\left(1-{\mathrm{y_m}}^{-\xi}\right)}{\xi}, \quad \frac{\sigma\left(1-{\mathrm{y_m}}^{-\xi}\right)}{\xi^2} - \frac{\sigma {\mathrm{y_m}}^{-\xi}\log(\mathrm{y_m})}{\xi}\right]$$

With

$$\mathrm{y_m} = -\log(1-\mathrm{p})$$

3.3.2 GPD distribution

The log likelihood function of the Generalized Pareto Distribution is given by (Coles 2001):

$$l(\sigma,\xi) = -k * log(\sigma) - \left(1+\frac{1}{\xi}\right)\sum_{i=1}^{k} log\left(1+\xi\frac{y_i}{\sigma}\right), \; y = z - u \qquad 13$$

Provided that $\left\{1+\frac{\xi y_i}{\sigma} > 0 \; voor \; i = 1,\dots,k\right\}$. If ξ=0 this equation becomes

$$l(\sigma) = -k * log(\sigma) - \frac{1}{\sigma}\sum_{i=1}^{k} y_i \qquad 14$$

The obtained parameters can be used to calculate the maximum likelihood return level.

$$z_m = \begin{cases} u + \frac{\sigma}{\xi}\left(\left(\frac{T * k}{A}\right)^{\xi} - 1\right) & voor\ \xi \neq 0 \\ u + \sigma log\left(\frac{T * k}{A}\right) & voor\ \xi = 0 \end{cases}$$

The variance estimated by the delta method is given by:

$$Var(z_m) \approx \nabla {z_m}^T * V * \nabla z_m$$

where V is the variance-covariance matrix of (σ,ξ) and

$$\nabla z^{\mathrm{T}} = \left[\frac{\partial \mathrm{z_m}}{\partial \sigma}, \frac{\partial \mathrm{z_m}}{\partial \xi}\right] = \left[\frac{\left(\left(\frac{\mathrm{T} * \mathrm{k}}{\mathrm{A}}\right)^{-\xi} - 1\right)}{\xi}, \quad -\frac{\sigma\left(\left(\frac{\mathrm{T} * \mathrm{k}}{\mathrm{A}}\right)^{-\xi} - 1\right)}{\xi^2} + \frac{\sigma\left(\frac{\mathrm{T} * \mathrm{k}}{\mathrm{A}}\right)^{\xi} \log\left(\frac{\mathrm{T} * \mathrm{k}}{\mathrm{A}}\right)}{\xi} \right]$$

4 Selection of appropriate conditional distribution

There are some tools to select the right distribution for a dataset. A maximum likelihood estimation will give the best fit of the selected distribution though a dataset but doesn't guarantee that this distribution is appropriate. The following tools are available to select the best conditional distribution.

4.1 Excess functions

By means of the mean excess function it is possible to get an estimate of the right distribution. The mean excess is the mean of excess values of all POT values exceeding the threshold. Every conditional distribution has a mean excess function which give the mean excess in function of a threshold (u). The shape of the empirical mean excess function can be compared with theoretical mean excess function of the different distributions. If the empirical mean excess function has an increasing trend the corresponding distribution will belong to the GPD ($\xi > 0$), the Pareto distribution or the Conditional Weibull distribution ($\tau < 1$). In case of a horizontal mean excess function the data will follow a GPD ($\xi = 0$), or an exponential distribution. If the mean excess function is decreasing in function of the threshold the observation set will belong to the GPD ($\xi < 0$), or the CWD ($\tau > 1$). This comparison can be made before fitting the parameters of a distribution (Figure 4-1; Table 4-1).

A second advantage of the mean excess function is its use in the determination of the optimal threshold. A too low threshold will likely violate the asymptotic basis of the model, leading to bias and a too high threshold will generate few excesses with which the model can be estimated, leading to high variance. The empirical mean excess function has to keep the same trend above this optimal threshold (Coles 2001; Beirlant 2004).

Table 4-1: Mean excess functions (Beirlant 2004)

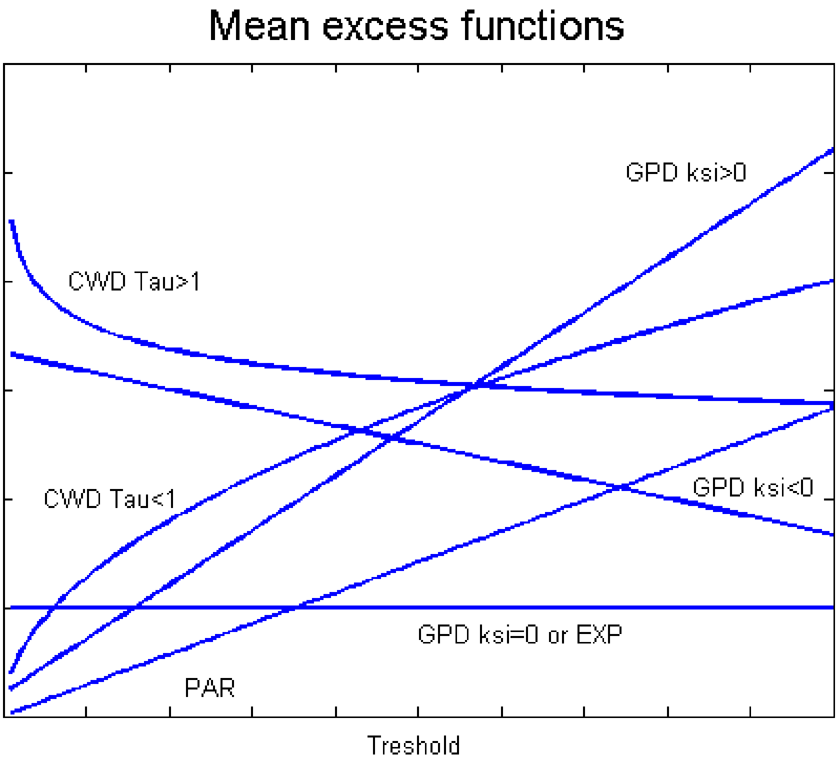
Distribution	Mean excess function
GPD	$E(X-u\|X>u) = \frac{\sigma + \xi u}{1-\xi}$
Exponential distribution	$E(X-u\|X>u) = \frac{1}{\lambda}$
Pareto distribution	$E(X-u\|X>u) = \frac{\lambda u}{1+\lambda}$
Conditional Weibull distribution	$E(X-u\|X>u) = \frac{u^{1-\tau}}{\lambda\tau}$

The empirical mean excess function can be found by:

$$\hat{e}_k = \frac{\sum_{i=1}^{k} X_i * 1_{(u,\infty)}(X_i)}{\sum_{i=1}^{k} 1_{(u,\infty)}(X_i)} - u$$

Where $1_{(u,\infty)}(x_i)$ equals 1 if xi >u, and 0 otherwise (Beirlant 2004).

Figure 4-1: Mean excess functions



4.2 RMSE as a parameter of the goodness of fit

Once the parameters of the selected distribution are determined the root mean square error (RMSE) between the empirical and the model values can be calculated. This is done by assigning an exceedance probability to every observation above the optimal threshold u. After sorting the observations from large to small

$$X_1 \geq X_2 \geq \cdots \geq X_k$$

the exceedance frequencies corresponding to the sorted observations are calculated by:

$$p_i = \frac{i}{k+1}$$

These frequencies are used in the inverse cumulative distribution to calculate the estimated observations M1, …, Mk. The RMSE of X and M gives an estimation of the goodness of fit of the selected distribution.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{k}(X_i - M_i)^2}{k}}$$

4.3 Confidence intervals of the parameters

The confidence intervals of the parameters give an impression of the possible variation. The confidence intervals are calculated based on the standard deviation and mean value of the parameters obtained in the maximum likelihood fit. Very wide confidence intervals will result in a very uncertain return level. A large variation can be caused by a small dataset or an inappropriate choice of distribution.

4.4 Visual control of the return level plot

The goal of an extreme value analysis is the extrapolation to return periods and corresponding return levels higher values than the empirical values. So the return period vs. return level plot is of major importance. There has to be a good similarity between the calculated curve and the empirical values in the low return period domain to give reliable extrapolation values. A visual check is a useful tool to guarantee this similarity.

4.5 Check of the Poisson process

By assigning an empirical probability of i/(1+n) to the POT values an implicit assumption of a stationary Poisson process is made. This means that the occurrence of extreme values follows a Poisson distribution and that the values are not clustered. A check of this assumption is the dispersion coefficient (Vitolo 2009). This is the ratio of the variance and the mean of the number of POT per year. A dispersion smaller than 1 indicates a more regularly process and larger than 1 indicates clustering.

5 Changes in the methodology

This chapter gives an overview of the most important changes in the new methodology in comparison with the 'standard methodology' for Extreme Value Analysis (Willems 2001 & 2002).

5.1 Parameters estimation by the MLE method

The most important change in the new methodology is the parameter estimation. The parameters are estimated by a more reliable maximum likelihood technique in the new methodology, instead of linearization of the QQ-plots, like the Hill estimator. The MLE likelihood allows to estimate the parameters of more distributions like the GEV and GPD distributions while this was not included in the previous methodology.

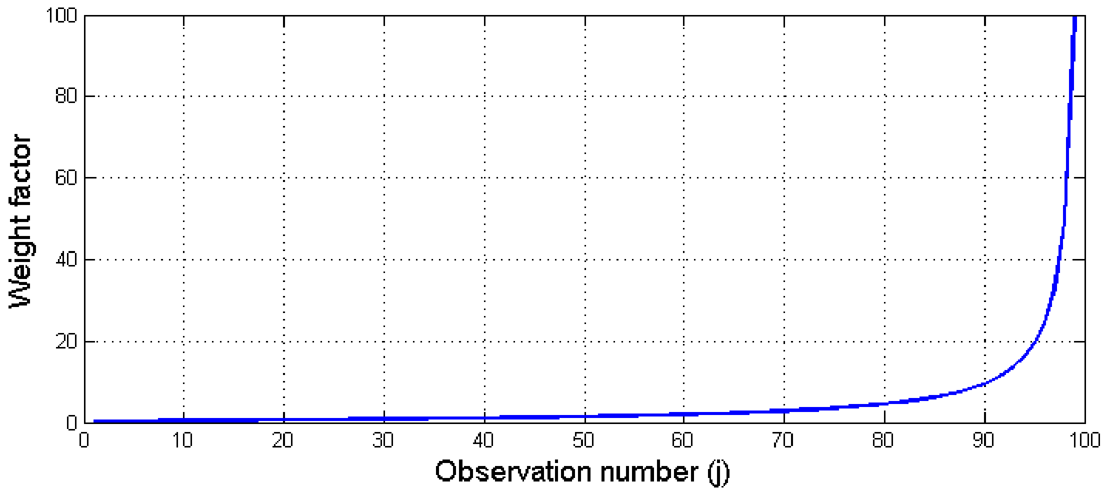
5.2 Weight factor

The weight factor of the 'standard methodology' is not included in the new methodology. This weight factor was used in the parameterization of the Pareto distribution and the calculation of the mean square error. For observations sorted from large to small the weight factor was given by:

$$w_i = -\frac{1}{\log\left(\frac{i}{k}\right)}$$

With i the observations number and k the number of observations. This factor takes the heteroscedasticity, i.e. increasing variance with increasing observation value, into account by assigning a smaller weight to the larger observations (Figure 5-1).

Figure 5-1: Weight factor of the standard methodology



(high POT value have a low weight factor)

The parameterization and the calculation of the RMSE in the new methodology does not use this weighting factor for 2 reasons:

- The very large observations have indeed a larger variation but are on the other hand also more rare. So their impact on the parameter estimation will be rather small.
- The implied weight factor is considered rather strictly for the large observation, because they get a very low weight factor. A low MSE or RMSE calculated with the weighting factor means the distribution has a good fit for the lower observations but not necessarily for the high observation.

6 References

Beirlant J., Goegebeur, Y., Teugels, J., 2004, Statistics of Extremes, Theory and Applications, John Wiley & Sons Ltd.

Coles, S. 2001, An introduction to statistical modeling of extreme values. London: Springer-Verlag

Kotz, S. & Nadarajah, S. 2000, Extreme value distributions, theory and applications. London: Imperial College Press

Nelsen RB 1986, Properties of a one-parameter family of bivariate distributions with specified marginals. Comm Statist Theory Methods 15:3277-3285

Vitolo, R. & Stephenson, D. B., 2009, Serial clustering of intense European storms, Willis Research Network

Willems, P., Berlamont, J., Qvick, A., Vaes, G., Feyen, J., Christiaens, K., 2001, Algemene methodologie voor het modelleren van de waterafvoer in bevaarbare waterlopen in Vlaanderen, studie i.o.v. AWZ, afd. Waterbouwkundig Laboratorium en Hydrologisch Onderzoek.

Willems P., Vaes G., Berlamont J., 2002, Onderzoek naar het effect van composiethydrogrammen voor de berekening van overstromingsvlaktes met eendimensionale hydraulische modellen, eindrapport, KULeuven

Appendix A: Extreme value model formulae

GEV

Function	Formula	Properties
Cumulative density function	$\begin{cases} G(z) = exp\left\{-\left[1+\xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-1/\xi}\right\} & \xi \neq 0 \\ G(z) = exp\left\{-\left[-\left(\frac{z-\mu}{\sigma}\right)\right]\right\} & \xi = 0 \end{cases}$	$\left\{z : 1 + \frac{\xi(z-\mu)}{\sigma} > 0\right\}$ $-\infty<\mu<\infty$, $\sigma>0$ $-\infty<\xi<\infty$
Survival function	$\begin{cases} 1-G(z) = 1-exp\left\{-\left[1+\xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-1/\xi}\right\} & \xi \neq 0 \\ 1-G(z) = 1-exp\left\{-\left[-\left(\frac{z-\mu}{\sigma}\right)\right]\right\} & \xi = 0 \end{cases}$	$-\infty<\mu<\infty$, $\sigma>0$ $-\infty<\xi<\infty$
Probability density function	$g(z) = \frac{1}{\sigma}\left[1+\xi\left(\frac{z-\mu}{\sigma}\right)\right]^{\left(-\frac{1}{\xi}\right)-1} exp\left\{-\left[1+\xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-1/\xi}\right\}$	$-\infty<\mu<\infty$, $\sigma>0$ $-\infty<\xi<\infty$
Return level	$z_p = \begin{cases} \mu - \frac{\sigma}{\xi}\left[1-\{-log(1-p)\}^{-\xi}\right] & \xi \neq 0 \\ \mu - \sigma log\{-log(1-p)\} & \xi = 0 \end{cases}$	p=probability $-\infty<\mu<\infty$, $\sigma>0$ $-\infty<\xi<\infty$

Gumbel distribution

Function	Formula	Properties
Cumulative density function	$G(z) = exp\left(-exp\left(-\frac{z-\mu}{\sigma}\right)\right)$	$-\infty<z<\infty$
Survival function	$1-G(z) = 1-exp\left(-exp\left(-\frac{z-\mu}{\sigma}\right)\right)$	$-\infty<z<\infty$
Probability density function	$G(z) = \frac{exp\left(-\frac{z-\mu}{\sigma}\right)}{\beta} exp\left(-exp\left(-\frac{z-\mu}{\sigma}\right)\right)$	$-\infty<z<\infty$
Return level	$z_p = \mu - \sigma log\{-log(1-p)\}$	p=probability $-\infty<z<\infty$

GPD (y=x-u)

Function	Formula	Properties
Cumulative density function	$\begin{cases} G(z) = 1 - \left[1 + \frac{\xi y}{\sigma}\right]^{-1/\xi} \\ G(z) = 1 - exp\left(\frac{-y}{\sigma}\right) \end{cases}$	$y > 0\,, (1 + \frac{\xi y}{\sigma}) > 0$ $y = z - u$
Survival function	$\begin{cases} 1 - G(z) = \left[1 + \frac{\xi y}{\sigma}\right]^{-1/\xi} \\ 1 - G(z) = exp\left(\frac{-y}{\sigma}\right) \end{cases}$	$y > 0\,, (1 + \frac{\xi y}{\sigma}) > 0$ $y = z - u$
Probability density function	$g(z) = \frac{1}{\sigma}\left[1 + \xi\left(\frac{y}{\sigma}\right)\right]^{\left(-\frac{1}{\xi}\right)-1}$	$y > 0\,, (1 + \frac{\xi y}{\sigma}) > 0$

Return level	$x_m = \begin{cases} u + \frac{\sigma}{\xi}\left(\left(\frac{T * k}{A}\right)^{\xi} - 1\right) & \xi \neq 0 \\ u + \sigma log\left(\frac{T * k}{A}\right) & \xi = 0 \end{cases}$	T= return period
		 A= # years of data
		 k= # POT above u

Exponential distribution (conditional form y=x-u)

Function	Formula	Properties
Cumulative density function	$G(z) = 1 - exp(-\lambda y)$	$y > 0$
Survival function	$1 - G(z) = exp(-\lambda y)$	$y > 0$
Probability density function	$g(z) = \lambda * exp(-\lambda y)$	$y > 0$

Return level	$x_m = u + \frac{1}{\lambda} log\left(\frac{T * k}{A}\right)$	T= return period
		 A= # years of data
		 k= # POT above u

Pareto distribution (conditional form y=x/u)

Function	Formula	Properties
Cumulative density function	$G(z) = 1 - y^{-\lambda}$	$y > 1$
Survival function	$1 - G(z) = y^{-\lambda}$	$y > 1$
Probability density function	$g(z) = \lambda u^{\lambda} x^{-\lambda-1}$	

Return level	$x_m = u\left(\frac{T * k}{A}\right)^{1/\lambda}$	T= return period
		A= # years of data
		k= # POT above u

Conditional Weibull distribution (y=x-u)

Function	Formula	Properties
Cumulative density function	$G(z) = 1 - exp(-\lambda y^{\tau})$	$y > 0, \tau > 0$
Survival function	$1 - G(z) = exp(-\lambda y^{\tau})$	$y > 0, \tau > 0$
Probability density function	$g(z) = \lambda \tau y^{\tau-1} exp(-\lambda y^{\tau})$	$y > 0, \tau > 0$

Return level	$x_m = u + \left(\frac{1}{\lambda} log\left(\frac{T * k}{A}\right)\right)^{1/\tau}$	T= return period
		A= # years of data
		k= # POT above u

Appendix B: Validation of conditional distributions

GPD

Let X have the distribution function F, and for a sufficiently large number of values n:

$$F^n(z) \approx exp\left\{-\left[1+\xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-1/\xi}\right\}$$

With the parameter restriction by $\left\{z : 1+\frac{\xi(z-\mu)}{\sigma} > 0\right\}$, where -∞<μ<∞, σ>0 en -∞<ξ<∞.

$$\mathrm{n}\log F(z) \approx -\left[1+\xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-1/\xi}$$

For high values of z a Taylor expansion can be applied:

$$\log F(z) = \log(1-(1-F(z)) \approx -\{1-F(z)\}$$

After substitution

$$1-F(z) \approx \frac{1}{n}\left[1+\xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-1/\xi}$$

$$1-F(u) \approx \frac{1}{n}\left[1+\xi\left(\frac{u-\mu}{\sigma}\right)\right]^{-1/\xi}$$

$$1-F(u+y) \approx \frac{1}{n}\left[1+\xi\left(\frac{u+y-\mu}{\sigma}\right)\right]^{-1/\xi}$$

So the conditional probability is given by (y=X-u):

$$Pr\{X > u+y|X > u\} = \frac{1-F(u+y)}{1-F(u)}, \qquad y > 0$$

$$Pr\{X > u+y|X > u\} \approx \frac{\frac{1}{n}\left[1+\xi\left(\frac{u+y-\mu}{\sigma}\right)\right]^{-1/\xi}}{\frac{1}{n}\left[1+\xi\left(\frac{u-\mu}{\sigma}\right)\right]^{-1/\xi}}, \qquad y > 0$$

$$Pr\{X > u+y|X > u\} = \left[1+\frac{\xi\left(\frac{y}{\sigma}\right)}{1+\xi\left(\frac{u-\mu}{\sigma}\right)}\right]^{-1/\xi}, \qquad y > 0$$

$$Pr\{X > u+y|X > u\} = \left[1+\frac{\xi y}{\tilde{\sigma}}\right]^{-1/\xi}, \qquad y > 0$$

With:

$$\tilde{\sigma} = \sigma + \xi(u-\mu)$$

The cumulative distribution formula is given by:

$$G(z) = 1-\left[1+\frac{\xi y}{\sigma}\right]^{-1/\xi}$$

In the case of ξ=0 this formula can be transformed by taking the limit of $\xi \to 0$:

$$G(z) = 1-\exp\left(\frac{-y}{\sigma}\right)$$

Exponential distribution

The same derivation can be applied for the exponential distribution.

$$Pr\{X > u + y | X > u\} = \frac{1 - F(u + y)}{1 - F(u)}, \qquad y > 0$$

$$Pr\{X > u + y | X > u\} = \frac{\exp(-\lambda(u + y)}{\exp(-\lambda u)} = \exp(-\lambda y), \qquad y > 0$$

So the conditional formula of the exponential distribution is exactly the same as the marginal formula. The cumulative distribution formula is given by:

$$G(z) = 1 - exp(-\lambda y)$$

Pareto distribution

The conditional formula of the Pareto distribution can be found by:

$$Pr\{X > u + y | X > u\} = \frac{1 - F(u + y)}{1 - F(u)}, \qquad y > 0$$

$$Pr\{X > u + y | X > u\} = \frac{(\mathrm{u} + \mathrm{y})^{-\lambda}}{(\mathrm{u})^{-\lambda}} = \left(\frac{\mathrm{u} + \mathrm{y}}{\mathrm{u}}\right)^{-\lambda} = \left(\frac{\mathrm{X}}{\mathrm{u}}\right)^{-\lambda}, \qquad y > 0$$

This formula can be simplified by the substitution of y=X/u. It is important to keep in mind that the y in the Pareto distribution is not the same as the y in the other distributions. This yields in the familiar conditional formula of the Pareto distribution:

$$G(z) = 1 - y^{-\lambda}$$

Conditional Weibull distribution

The Conditional Weibull distribution, as it is implied in various reports, is not the conditional formula of the Weibull formula but the Weibull formula used as a conditional formula. This may be a subtle difference but it is important to keep it in mind. Because of the wide range of applications of this conditional Weibull distribution it is advised to keep it in its current form with y=x-u:

$$G(z) = 1 - exp(-\lambda y^{\tau})$$

Appendix C: Derivation of the return level

GEV

The return level of a marginal distribution is obtained by inversing the cumulative density function.

$$\begin{cases} G(z) = exp\left\{-\left[1+\xi\left(\frac{z_p-\mu}{\sigma}\right)\right]^{-1/\xi}\right\} = 1-p & ,\xi \neq 0 \\ G(z) = exp\left\{-\left[-\left(\frac{z_p-\mu}{\sigma}\right)\right]\right\} = 1-p & ,\xi = 0 \end{cases}$$

Rearranging gives the return level in function of the exceedance probability p:

$$z_p = \begin{cases} \mu - \frac{\sigma}{\xi}\left[1-\{-log(1-p)\}^{-\xi}\right] & ,\xi \neq 0 \\ \mu - \sigma log\{-log(1-p)\} & ,\xi = 0 \end{cases}$$

The probability can be expressed in terms of the return period T (years).

$$p = \begin{cases} \frac{1}{T} & blockmaxima\ per\ year \\ \frac{1}{12 * T} & blockmaxima\ per\ month \\ \frac{1}{365.25 * T} & blockmaxima\ per\ day \end{cases}$$

Gumbel distribution

The Gumbel distribution is equal to the GEV distribution with ξ=0. This results in a return level given by:

$$z_p = \mu - \sigma log\{-log(1-p)\}$$

The exceedance probability (p) can be expressed in terms of the return period T (years).

$$p = \begin{cases} \frac{1}{T} & blockmaxima\ per\ year \\ \frac{1}{12 * T} & blockmaxima\ per\ month \\ \frac{1}{365.25 * T} & blockmaxima\ per\ day \end{cases}$$

GPD

The return level of a corresponding return period can be derived from the survival function

$$Pr\{X > x|X > u\} = \left[1 + \xi\frac{x-u}{\sigma}\right]^{-1/\xi} \quad ,y > 0\,,(1 + \frac{\xi y}{\tilde{\sigma}}) > 0$$

It follows that:

$$Pr\{X > x\} = \zeta_u\left[1 + \frac{\xi y}{\sigma}\right]^{-1/\xi} \quad ,y > 0\,,(1 + \frac{\xi y}{\sigma}) > 0$$

Where ζ_u = Pr{X>u}. Hence, the level x_m that is exceeded on average once every m observations is the solution of:

$$\zeta_u\left[1 + \xi\left(\frac{x_m - u}{\sigma}\right)\right]^{-1/\xi} = \frac{1}{m}$$

Rearranging gives (y=x-u),

$$x_m = u + \frac{\sigma}{\xi}\left((m\zeta_u)^{\xi} - 1\right)$$

Provided m is sufficiently large to ensure that x_m>u. In case ξ=0 a similar derivation gives:

$$x_m = u + \sigma log(m\zeta_u)$$

Again provided m is sufficiently large. The natural estimator of ζ_u is given by:

$$\hat{\zeta}_u = \frac{k}{n}$$

With n the total number of observations and k the number of observations exceeding the threshold u. The return level is usually expressed in function of return period T. By expressing m in function of T:

$$m = T * n_y$$

With T the return period [year] and n_y the number of observations per year:

$$x_m = \begin{cases} u + \frac{\sigma}{\xi}\left(\left(\frac{T * n_y * k}{n}\right)^{\xi} - 1\right) & ,\xi \neq 0 \\ u + \sigma log\left(\frac{T * n_y * k}{n}\right) & ,\xi = 0 \end{cases}$$

Under the assumption of an evenly distribution of the observations over the years, the number of observation years (A) can be used as an estimator of $\frac{n}{n_y}$:

$$A = \frac{n}{n_y}$$

This results in:

$$x_m = \begin{cases} u + \frac{\sigma}{\xi}\left(\left(\frac{T * k}{\mathrm{A}}\right)^{\xi} - 1\right) & ,\xi \neq 0 \\ u + \sigma log\left(\frac{T * k}{A}\right) & ,\xi = 0 \end{cases}$$

Exponential distribution

The return level of the exponential distribution can be derived from the survival function distribution.

$$Pr\{X > x|X > u\} = \exp(-\lambda y)$$

It follows that:

$$Pr\{X > x\} = \zeta_u \exp(-\lambda y)$$

Where ζ_u = Pr{X>u}. Hence, the level x_m that is exceeded on average once every m observations is the solution of:

$$\zeta_u \exp(-\lambda y) = \frac{1}{m}$$

Rearranging gives (y=x-u),

$$x_m = u - \frac{1}{\lambda} log\left(\frac{1}{m\zeta_u}\right)$$

Provided m is sufficiently large to ensure that x_m>u. The natural estimator of ζ_u is given by:

$$\hat{\zeta}_u = \frac{k}{n}$$

With n the total number of observations and k the number of observations exceeding the threshold u. The return level is usually expressed as a function of return period T. By expressing m in function of T:

$$m = T * n_y$$

With T the return period [year] and n_y the number of observations per year:

$$x_m = u + \frac{1}{\lambda} log\left(\frac{T * n_y * k}{n}\right)$$

Under the assumption of an evenly distribution of the observations over the years, the number of observation years (A) can be used as an estimator of $\frac{n}{n_y}$:

$$A = \frac{n}{n_y}$$

This results in:

$$x_m = u + \frac{1}{\lambda} log\left(\frac{T * k}{A}\right)$$

Pareto distribution

The return level of the exponential distribution can be derived from the survival distribution.

$$Pr\{X > x | X > u\} = \mathrm{y}^{-\lambda}$$

It follows that:

$$Pr\{X > x\} = \zeta_u y^{-\lambda}$$

Where ζ_u = Pr{X>u}. Hence, the level x_m that is exceeded on average once every m observations is the solution of:

$$\zeta_u y^{-\lambda} = \frac{1}{m}$$

Rearranging gives (y=x/u),

$$x_m = u\left(\frac{1}{m\zeta_u}\right)^{-1/\lambda} = u(m\zeta_u)^{1/\lambda}$$

Provided m is sufficiently large to ensure that x_m>u. The natural estimator of ζ_u is given by:

$$\hat{\zeta}_u = \frac{k}{n}$$

With n the total number of observations and k the number of observations exceeding the threshold u. The return level is usually expressed as a function of return period T. By expressing m in function of T:

$$m = T * n_y$$

With T the return period [year] and n_y the number of observations per year:

$$x_m = u\left(\frac{T * n_y * k}{n}\right)^{1/\lambda}$$

Under the assumption of an evenly distribution of the observations over the years, the number of observation years (A) can be used as an estimator of $\frac{n}{n_y}$:

$$A = \frac{n}{n_y}$$

This results in:

$$x_m = u\left(\frac{T * k}{A}\right)^{1/\lambda}$$

Conditional Weibull distribution

A similar derivation can be applied for the CWD.

$$Pr\{X > x|X > u\} = \exp(-\lambda y^{\tau})$$

It follows that:

$$Pr\{X > x\} = \zeta_u exp(-\lambda y^{\tau})$$

Where ζ_u = Pr{X>u}. Hence, the level x_m that is exceeded on average once every m observations is the solution of:

$$\zeta_u \exp(-\lambda(x_m - u)^{\tau}) = \frac{1}{m}$$

Rearranging gives (y=x-u),

$$\mathrm{x_m} = \mathrm{u} + \left(-\frac{1}{\lambda}\log\left(\frac{1}{\mathrm{m}\zeta_\mathrm{u}}\right)\right)^{1/\tau}$$

Provided m is sufficiently large to ensure that x_m>u. The natural estimator of ζ_u is given by:

$$\hat{\zeta}_u = \frac{k}{n}$$

With n the total number of observations and k the number of observations exceeding the threshold u. The return level is usually expressed as a function of return period T. By expressing m as a function of T:

$$m = T * n_y$$

With T the return period [year] and n_y the number of observations per year:

$$x_m = u + \left(\frac{1}{\lambda} log\left(\frac{T * n_y * k}{n}\right)\right)^{1/\tau}$$

Under the assumption of an evenly distribution of the observations over the years, the number of observation years (A) can be used as an estimator of $\frac{n}{n_y}$:

$$A = \frac{n}{n_y}$$

This results in:

$$x_m = u + \left(\frac{1}{\lambda} log\left(\frac{T * k}{A}\right)\right)^{1/\tau}$$

Appendix D: Maximum likelihood estimations

Gumbel distribution

The Gumbel log likelihood function is given by:

$$l(\mu,\sigma) = -n * log(\sigma) - \sum_{i=1}^{n}\left(\frac{z_i-\mu}{\sigma}\right) - \sum_{i=1}^{n} exp\left(-\left(\frac{z_i-\mu}{\sigma}\right)\right)$$

The obtained parameters can be used to calculate the maximum likelihood return level.

$$z_m = \mu - \sigma log\{-log(1-p)\}$$

The variance estimated by the delta method is given by:

$$Var(z_m) \approx \nabla z_m{}^{\mathrm{T}} * \mathrm{V} * \nabla z_m$$

Where V is the variance-covariance matrix of (μ,σ) and

$$\nabla x_m{}^T = \left[\frac{\partial z_m}{\partial \mu}, \frac{\partial z_m}{\partial \sigma}\right] = [1, -log(-log(1-p))]$$

Exponential distribution

The Likelihood function of the exponential distribution is given by:

$$L(\lambda) = \prod_{i=1}^{k} \lambda * exp(-\lambda y_i) = \lambda^k exp(-k\lambda\bar{y})$$

With

$$\bar{y} = \frac{1}{k}\sum_{i=1}^{k} y_i$$

The derivative of the log likelihood function is:

$$\frac{d}{d\lambda} l(\lambda) = \frac{d}{d\lambda} log(L(\lambda)) = \frac{k}{\lambda} - \sum_{i=1}^{k} x_i$$

This leads to the maximum likelihood estimator of

$$\hat{\lambda} = \frac{1}{\bar{y}}$$

Pareto distribution

The logarithmic likelihood function of the Pareto distribution is given by

$$l(\lambda) = k * log(\lambda) + k * \lambda * log(u) - (\lambda + 1) \sum_{i=1}^{k} log(x_i)$$

The derivative of the log likelihood function is:

$$\frac{d}{d\lambda} l(\lambda) = \frac{k}{\lambda} + k * log(u) - \sum_{i=1}^{k} log(x_i)$$

This yields the maximum likelihood estimator of λ:

$$\hat{\lambda} = \frac{k}{\sum_{i=1}^{k} log(x_i/u)}$$

Conditional Weibull distribution

The maximum likelihood estimator for the Conditional Weibull distribution is calculated using its similarity with the Gumbel distribution. If an observation set T has a Weibull distribution with parameters λ and τ than log(T) has a Gumbel distribution with parameters:

$$\tau = -\frac{1}{\sigma}$$

$$\lambda = \left(\frac{1}{exp(\mu)}\right)^{\tau}$$

DEPARTMENT **MOBILITY & PUBLIC WORKS**
Flanders hydraulics Research

Berchemlei 115, 2140 Antwerp
T +32 (0)3 224 60 35
F +32 (0)3 224 60 36
waterbouwkundiglabo@vlaanderen.be
www.flandershydraulicsresearch.be