

WATERBOUWKUNDIG LABORATORIUM



Flanders
State of
the Art

00_144_2
FHR reports

Scientific Assistance towards a Probabilistic Formulation of Hydraulic Boundary Conditions

Extreme Value Analysis software tool Manual

DEPARTMENT
MOBILITY &
PUBLIC
WORKS

www.flandershydraulicsresearch.be

Scientific Assistance towards a Probabilistic Formulation of Hydraulic Boundary Conditions

Extreme Value Analysis software tool Manual

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

Legal notice

Flanders Hydraulics Research is of the opinion that the information and positions in this report are substantiated by the available data and knowledge at the time of writing.
 The positions taken in this report are those of Flanders Hydraulics Research and do not reflect necessarily the opinion of the Government of Flanders or any of its institutions.
 Flanders Hydraulics Research nor any person or company acting on behalf of Flanders Hydraulics Research is responsible for any loss or damage arising from the use of the information in this report.

Copyright and citation

© The Government of Flanders, Department of Mobility and Public Works, Flanders Hydraulics Research 2019
 D/2017/3241/27

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 software tool Manual. Version 3.0. FHR Reports, 00_144_2. Flanders Hydraulics Research & IMDC: Antwerp.

Reproduction of and reference to this publication is authorised provided the source is acknowledged correctly.

Document identification

Customer:	Flanders Hydraulics Research	Ref.:	WL2019R 00_144_2
Keywords (3-5):	Extreme Value; Flooding; Floodrisk		
Text (p.):	27	Appendices (p.):	7
Confidentiality:	<input checked="" type="checkbox"/> No	<input checked="" type="checkbox"/> Available online	

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

Control

	Name	Signature
Reviser(s):	Pereira, F.; Nossent, J.	Getekend door: Fernando Pereira (Signature) Getekend op: 2019-11-29 15:23:19 +01:00 Reden: Ik keur dit document goed <i>Fernando Pereira</i>
		Getekend door: Jiri Nossent (Signature) Getekend op: 2019-11-29 15:48:08 +01:00 Reden: Ik keur dit document goed <i>Jiri Nossent</i>
Project leader:	Pereira, F.	Getekend door: Fernando Pereira (Signature) Getekend op: 2019-11-29 15:23:39 +01:00 Reden: Ik keur dit document goed <i>Fernando Pereira</i>

Approval

Head of Division:	Mostaert, F.	Getekend door: Frank Mostaert (Signature) Getekend op: 2019-11-29 15:22:24 +01:00 Reden: Ik keur dit document goed <i>Frank Mostaert</i>
-------------------	--------------	---



Abstract

The Extreme Value Analysis tool (EVA tool) is a Matlab® software package developed in commission of Flanders Hydraulic Research (FHR). The tool is a standalone executable which facilitates and automates the selection of extremes from time series, e.g. wind speed, wave heights, discharges, water levels, etc., the application of frequency analysis and the determinations of the appropriate distributions of these extremes. The tool is part of a suite of software tools to facilitate the probabilistic formulation of hydraulic boundary conditions. This report is a manual for the tool. It assists the user when installing the tool and gives an overview of the workflow. Furthermore it provides a short theoretical overview of the Extreme Value Analysis and an in depth description of the workflow and the functionalities of the tool.

Contents

Abstract	III
Contents	V
List of tables.....	VII
List of figures	VIII
1 Introduction.....	1
2 Software	2
2.1 Installation of MCR	2
3 Workflow	3
4 Short theoretical overview	4
4.1 Marginal vs conditional distributions	4
4.2 selection of the appropriate distribution	6
4.2.1 Excess functions (conditional distributions).....	6
4.2.2 Shape parameter	7
4.2.3 RMSE as an estimate for goodness of fit.....	7
4.2.4 Confidence intervals of the parameters.....	7
4.2.5 Visual control of the return level plot	7
4.3 Confidence intervals.....	8
4.4 Check of the Poisson process	8
5 Part 1: Selection of extremes	9
5.1 Input	9
5.1.1 Load/Unload time series	9
5.1.2 Set outputdir.....	11
5.1.3 Load POT.....	11
5.1.4 Load Block.....	11
5.2 Selection of extremes	12
5.2.1 POT selection	13
5.2.2 Selection of Block maxima.....	14
5.3 Visualization.....	15
5.4 Extra functionalities.....	17
6 Part 2: Conditional distributions.....	20
6.1 Practical use.....	21
6.2 Selecting the appropriate distribution	22

6.3	Optimal threshold.....	23
7	Part 3: marginal distributions.....	24
7.1	Practical use.....	25
7.2	Selecting the appropriate distribution	26
8	References.....	27
Appendix A: Output conditional distribution.....		A1
Figures: Demo.....		A1
Text files: Demo.....		A4
Appendix B Output marginal distribution.....		A5
Figures: Demo.....		A5
Text files: Demo.....		A7

List of tables

Table 4-1: Determination of the appropriate distribution by the shape parameter	7
Table 5-1: Available block ranges and corresponding block start formats	14
Table 6-1: Output files conditional distribution (example in Annex A).....	21
Table 7-1: Output files marginal distribution (examples in Annex B)	25

List of figures

Figure 1-1: Overview of reports, tools and manuals.....	1
Figure 3-1: Workflow of the EVA-tool	3
Figure 4-1: Tail behaviour of GEV with different values of ξ	5
Figure 4-2: Theoretical mean excess functions	6
Figure 5-1: Graphical interface of the EVA-tool: selection of extremes.....	9
Figure 5-2: Graphical interface to load or unload time series.....	10
Figure 5-3: Time management data series.....	11
Figure 5-4: Metadata needed to import POT values stored in a excel file.....	11
Figure 5-5: Metadata needed to import block maxima.	12
Figure 5-6: Pot and Block maxima selection GUI.....	12
Figure 5-7: POT Table	13
Figure 5-8: POT inspection GUI	14
Figure 5-9: Example of a time series with POT values and Block maxima.	15
Figure 5-10: Example of POT properties figure.	16
Figure 5-11: Log file	16
Figure 5-12: Extra menu tabs and info	17
Figure 5-13: Trademark input.....	17
Figure 5-14: Overrule timeframe.....	17
Figure 5-15: Highest extremes special.....	18
Figure 5-16: Directional division of the POT values.....	19
Figure 6-1: Conditional distribution fitting GUI.....	20
Figure 6-2: Conditional distribution control panel	23
Figure 7-1: Marginal distribution fitting GUI	24
Figure A-1: EVA-distribution sheet	A1
Figure A-2: Distribution parameters as function of number of POT	A2
Figure A-3: Stratified sampling	A3
Figure A-4: Distribution parameters.....	A4
Figure A-5: Selected POT values	A4
Figure A-6: Synthetic extremes created by stratified sampling	A4
Figure B-1: EVA-distribution sheet	A5
Figure B-2: Stratified sampling	A6
Figure B-3: Distribution parameters.....	A7
Figure B-4: Selected Block maxima.....	A7
Figure B-5: Synthetic extremes created by stratified sampling	A7

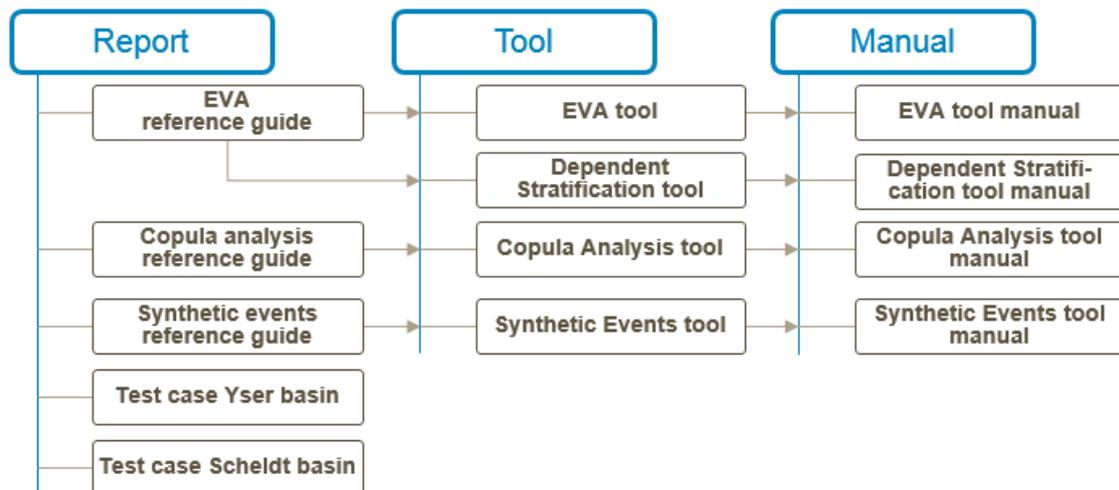
1 Introduction

The Extreme Value Analysis tool (EVA tool) is a Matlab® software package developed in commission of Flanders Hydraulic Research (FHR). The tool is a standalone executable which facilitates and automates the selection of extremes out of a time series, like for example wind speed, wave heights, discharges, water levels, etc, the application of frequency analysis and the determinations of the appropriate distributions of these extremes.

The tool is part of a suite of software tools to facilitate the probabilistic formulation of hydraulic boundary conditions. An overview of the tools and corresponding reports and manuals is presented in Figure 1-1.

The Extreme Value Analysis reference guide gives an overview of the methodology and a summary of the applied formulae. The theoretical techniques used to fit the extreme value distributions are based on international standard literature (Coles, 2001; Kotz, 2000, Nelsen, 2004) and Beirlant's masterpiece (Beirlant, 2004).

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



2 Software

The Extreme Value Analysis tool has been developed by IMDC in a Matlab® environment and compiled into an executable, so there is no software license required to use the toolbox. The tool consists of three main visual interfaces which give access to numerous Matlab® functions.

The user needs to install the Matlab Compiler Runtime (MCR) before the first execution of the EVA-tool. The MCR is a Matlab® copy without the graphical interface that can be deployed royalty free and possesses all the strengths of the full Matlab® environment. You must have administrative privileges to install the MCR on a target machine since it modifies both the system registry and the system path. Running the MCR Installer after the MCR has been set up on the target machine requires only user-level privileges.

2.1 Installation of MCR

The installation of the MCR Installer is guided by an installation GUI which requires the following steps.

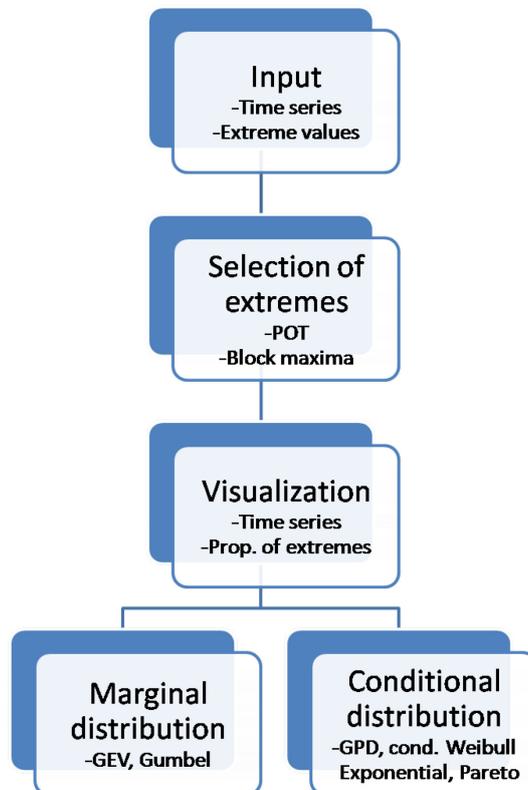
- When the MCR Installer wizard appears, click Next to begin the installation. Click Next to continue.
- In the Select Installation Folder dialog box, specify the location where you want to install the MCR and whether you want to install the MCR for just yourself or others. Click Next to continue.
- Confirm your selections by clicking Next.
- The installation begins. The process takes some time due to the amount of files that are installed.

A more detailed explanation of the MCR can be found on Matlab (2011).

3 Workflow

The workflow of the EVA-tool consists of 4 successive blocks presented in Figure 3-1. Starting with the import of time series followed by the selection of extreme values of the time series. These extreme values can be peak over threshold (POT) values or block maxima, like annual maxima. The tool provides the functionality to visualise the time series in combination with the selected extremes and generates a plot of the properties of these extreme values. This last plot will aid the user in his choice for the appropriate distribution. After the third step a distinction is made between marginal distributions, appropriate for the block maxima, and conditional distributions, for POT values.

Figure 3-1: Workflow of the EVA-tool



4 Short theoretical overview

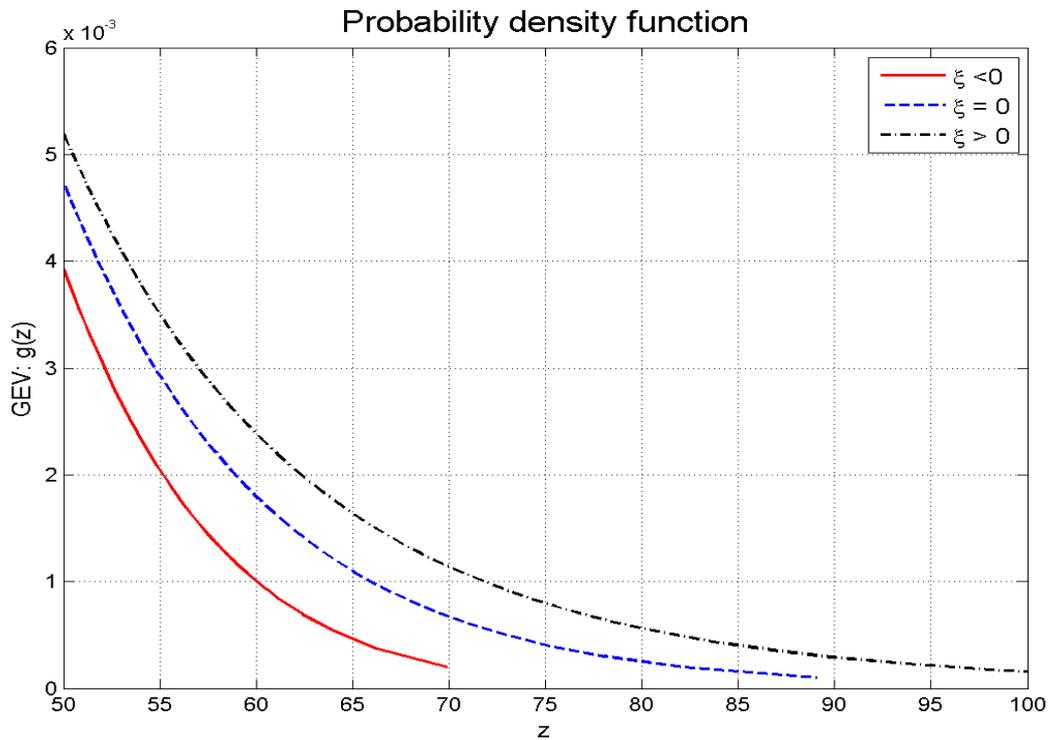
This chapter gives a short theoretical overview of the procedures in the EVA-tool. A more extensive theoretical consideration can be found in the Extreme Value Analysis Reference Guide (IMDC, 2015). Extreme value statistics is unique as a statistical discipline in that it develops techniques and models 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 specialized in the extrapolation to these extreme values combined with a number of tools to choose the appropriate model for the phenomena of interest.

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 (peak over threshold) values. It should be noted that every marginal distribution has a conditional counterpart and vice versa.

4.1 Marginal vs conditional distributions

The first task in extreme value analysis is the selection of appropriate extreme values. Traditionally maximum values of each year (or rather storm season) are selected. These are block maxima with a block range of one year. There is a wide range of extreme value marginal distributions (all of them member of the generalized extreme value (GEV) distribution family) available to quantify the stochastic properties of the block maxima. In case of very long datasets this is a strong and straightforward methodology. The Generalized Extreme Value distribution can be divided in three classes depending on tail behaviour (Figure 4-1). If the distribution has a light tail, $\xi < 0$, it is part of the extreme value Weibull domain. Distributions of this family have an upper boundary z_+ . If $\xi = 0$ the distribution belongs to the Gumbel domain and the tail decreases exponentially to infinity. The third family is the Fréchet family ($\xi > 0$). These distributions have a heavy tail which decreases polynomially to infinity (Coles 2001).

Figure 4-1: Tail behaviour of GEV with different values of ξ



In reality most datasets contain data from 10 up to 40 years or more. Fitting extreme value distributions through 10 up to 40 data points will result in wide confidence intervals and major uncertainty in the extrapolation domain. In most cases each year contains multiple extreme events, for instance several wind storms. In case of block maxima only the largest event will be selected. By selecting all the extreme events the maximum amount of information can be used to determine the parameters of the appropriate distribution. The peak over threshold (POT) values are all independent values that exceed a set threshold. The independency is guaranteed by two additional selection criteria: the inter event level and the time interval. The inter event level is the maximum value the minimum between two POT values may have. It is determined by a factor that has to be multiplied with the minimum of the two POT values. The time interval is the minimum time lag between two successive POT values. The initial threshold has to be kept sufficiently low to select enough extreme events. In a second step the optimal threshold will be determined, i.e. the threshold above which the values are extreme and follow the considered extreme value distribution.

While the block maxima have an approximating marginal distribution, part of the GEV distribution, the POT values have an approximate conditional distribution within the Generalized Pareto family. These conditional distributions are only valid above the selected optimal threshold. Similar to the GEV distribution, the shape parameter ξ is dominant in determining the behaviour of the GPD. The conditional Weibull (CWD) and conditional exponential distribution are members of the conditional Gumbel family with $\xi = 0$, and the conditional Pareto distribution is part of the conditional Fréchet family with $\xi > 0$. They have a tail that decreases exponentially and polynomially respectively.

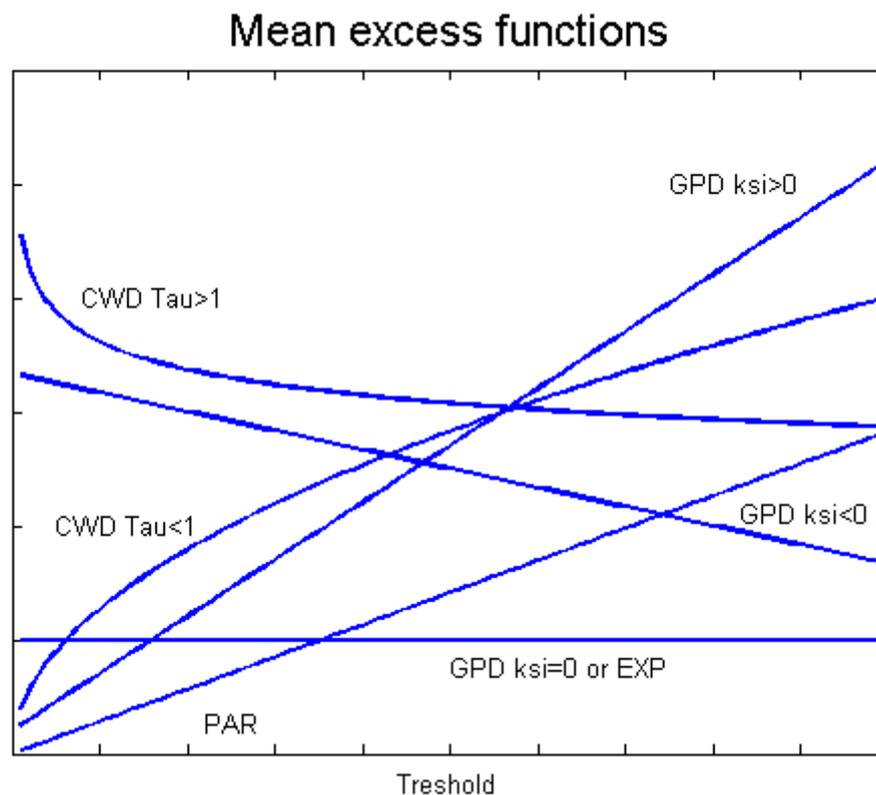
4.2 selection of the appropriate distribution

Tools are implemented to select the proper distribution for a dataset. A maximum likelihood estimation will give the best fit of the selected distribution through a dataset but doesn't guarantee that this distribution is the most appropriate one. The GEV and the GPD distributions cover the entire marginal and the conditional domain respectively. In spite of this broad range it is, in most cases, more convenient to use a more specialized distribution with less parameters or a shape designed for the dataset.

4.2.1 Excess functions (conditional distributions)

By means of the mean excess function it is possible to get an estimate of the proper conditional distribution. The mean excess is the mean over the excess values of all the POT values exceeding the threshold. Every conditional distribution has a theoretical mean excess function which gives the mean excess as a function of a threshold (u). The shape of the empirical mean excess function can be compared with theoretical mean excess functions of the different distributions. If the empirical mean excess function has an increasing trend, the corresponding distribution will belong to GPD ($\xi > 0$): Pareto distribution or conditional Weibull distribution ($\tau < 1$). In case of a horizontal mean excess function the data will follow GPD ($\xi = 0$): the exponential distribution. If the mean excess function is decreasing in function of the threshold the observation set will belong to the GPD ($\xi < 0$), i.e. the CWD ($\tau > 1$). This comparison can be made before fitting the parameters of a distribution (Figure 4-2).

Figure 4-2: Theoretical mean excess functions



4.2.2 Shape parameter

The shape parameter is an additional aid to select the appropriate conditional distribution and an aid to select the appropriate marginal distribution. The shape parameter of the GEV and the GPD distribution gives an indication of the appropriate family. The shape parameter of both the GEV and the GPD distribution will be calculated with 95 % confidence interval. Because of the inherent variation of datasets a shape parameter with value of exactly zero is not likely. However if the confidence interval of the shape parameter contains zero, the distributions corresponding to the $\xi=0$ can be selected. Table 4-1 displays the selection rules based on the shape parameter ξ .

Table 4-1: Determination of the appropriate distribution by the shape parameter

Confidence interval contains	Marginal domain	Conditional domain
$\xi < 0$	GEV	GPD
$\xi = 0$	GEV, Gumbel	GPD, CWD, Exponential
$\xi > 0$	GEV	GPD, Pareto

4.2.3 RMSE as an estimate for goodness of fit

After the parameters of the selected distribution are determined by a maximal likelihood fit, the root mean square error (RMSE) between the empirical and the model values is calculated. This is done by assigning an empirical exceedance probability to each observation above the optimal threshold u . After ranking the observations from large to small

$$X_1 \geq X_2 \geq \dots \geq X_k$$

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

$$p_i = \frac{i}{k + 0.5}$$

These probabilities are used in the inverse cumulative distribution to calculate the estimated observations M_1, \dots, M_k . The RMSE of X and M gives an estimation of the goodness of fit of the selected distribution.

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

4.2.4 Confidence intervals of the parameters

The confidence intervals of the parameters give an impression of the possible variation. Wider confidence intervals will result in uncertain return levels. Large variation can be caused by a too small dataset or an inappropriate choice of distribution.

4.2.5 Visual control of the return level plot

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

4.3 Confidence intervals

The confidence intervals of the return levels in the EVA-tool are calculated by means of the delta method.

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

where V is the variance-covariance matrix of θ MLE (parameter set of the log likelihood estimation) (Coles 2001).

A secondary method to generate confidence intervals is the bootstrap technique. The POT values are resampled to generate a large number (>1000) of POT sets. An extreme value distribution is fitted through each set and the 2.5 and 97.5 percentiles for every return period are the confidence intervals.

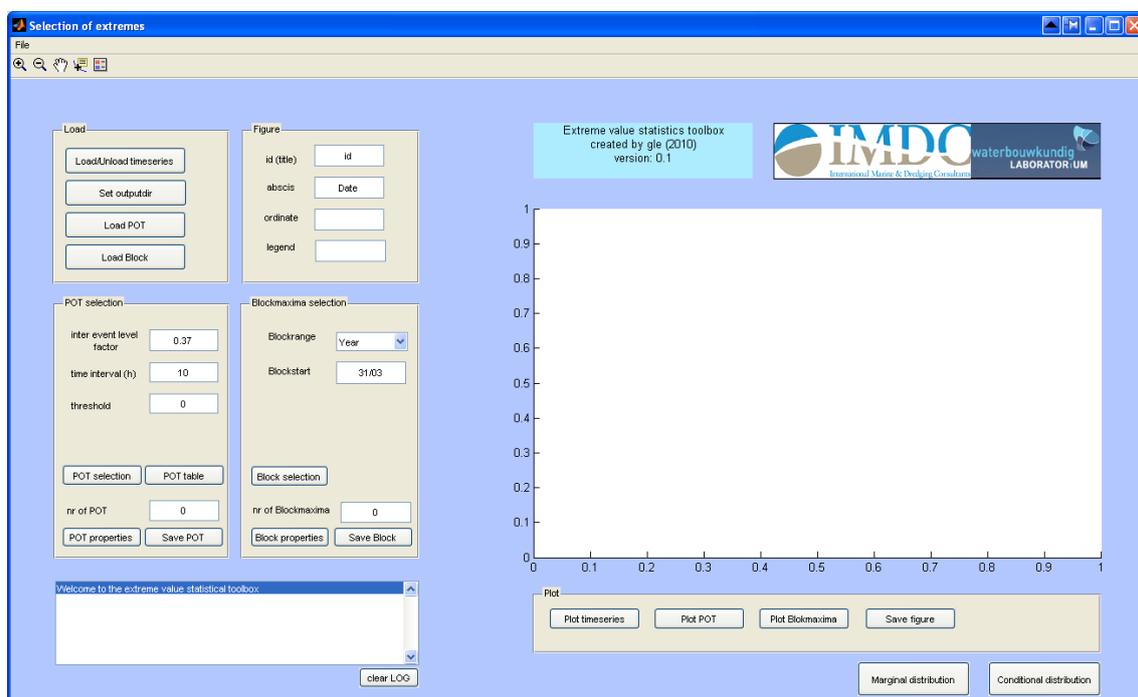
4.4 Check of the Poisson process

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 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 greater than 1 indicates clustering.

5 Part 1: Selection of extremes

The first graphical interface (Figure 5-1) of the Extreme Value Analysis tool is built to guide the user through the input of the data, the selection of extremes and the visualization of the data and the extremes. The file menu in the upper left corner contains 'save' and 'load workspace' functions as well as the 'new project' function. These functions allow the user to save his work, in order to be able to continue later. The new project function resets the tool to start the analysis for a new dataset.

Figure 5-1: Graphical interface of the EVA-tool: selection of extremes

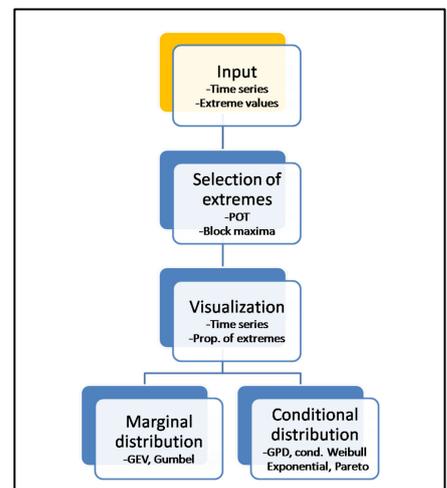


5.1 Input

The panel in the upper left corner of Figure 5-1 will guide the user through the input of time series or extreme values.

5.1.1 Load/Unload time series

The button "Load/Unload time series" will open a new window, "Load_data" (Figure 5-2). The current GUI has the ability to read ascii files in the inv format and binary mat files. An inv file needs to have 2 columns, the first with a timestamp (yyyy-mm-dd HH:MM:SS, yyyy-mm-dd HH:MM or yyyy/mm/dd HH:MM:SS) and a second with data. The mat-file has to contain 1 variable with in its first column the time in Matlab format (serial date starting at Jan-1-0000 00:00:00) and in its second column the data. The missing value is the value assigned to wrong or missing values (ASCII only). The properties panel gives the



start and end date and the minimum and maximum of the time series. If the input consist of multiple time series files the box “Overlap?” will give a warning in case of data with equal timestamps. The last imported dataset will not be accepted in this case. The button “Unload time series” will remove unwanted time series. Once the close button is clicked the program will check if there are data gaps in the time series. If there is missing data for a time span longer than 10 days this is considered a time gap. The total number of years real data is recalculated (Figure 5-3) and the user can use this new number of years in the following analysis (recommended).

Figure 5-2: Graphical interface to load or unload time series

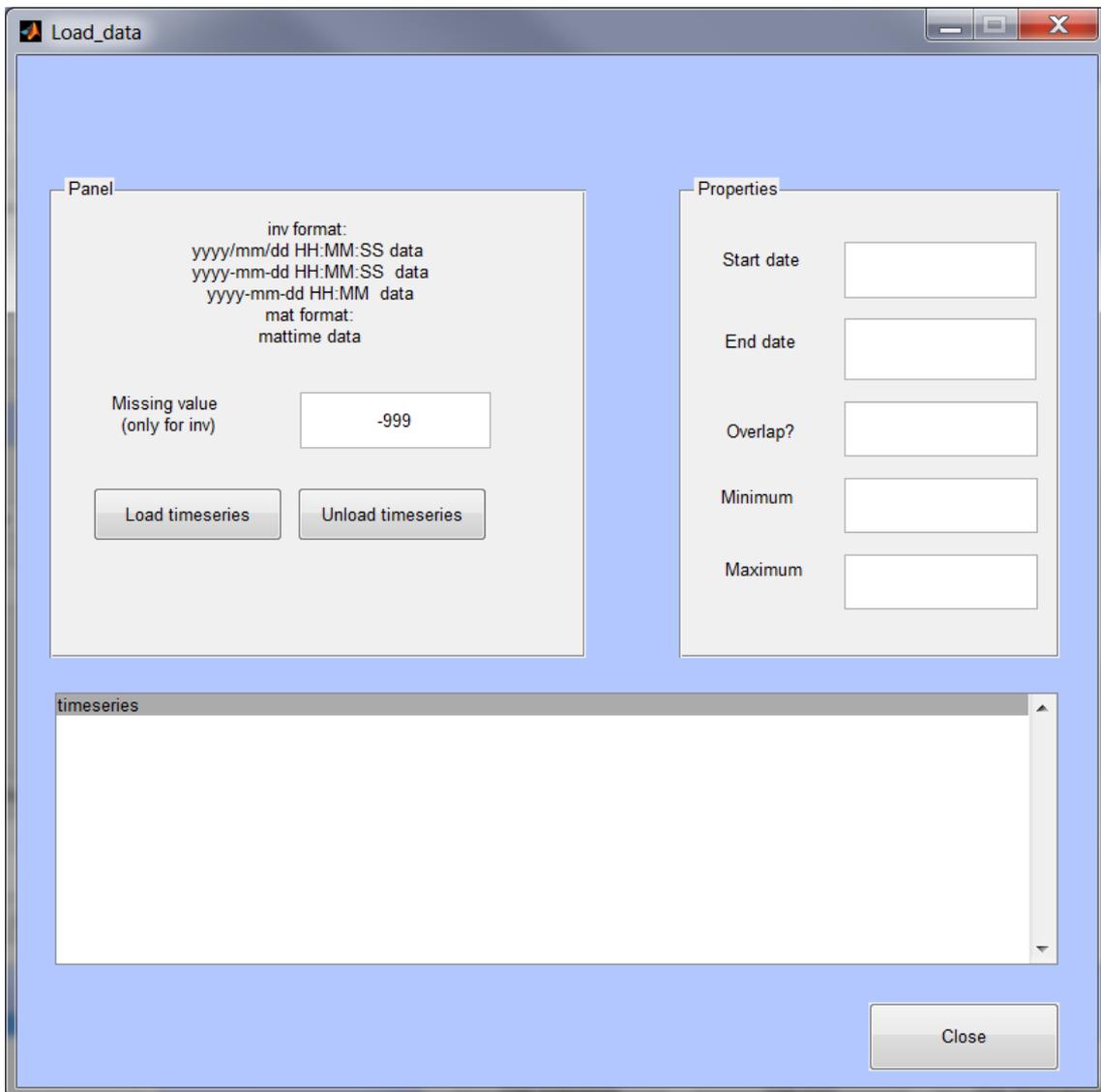
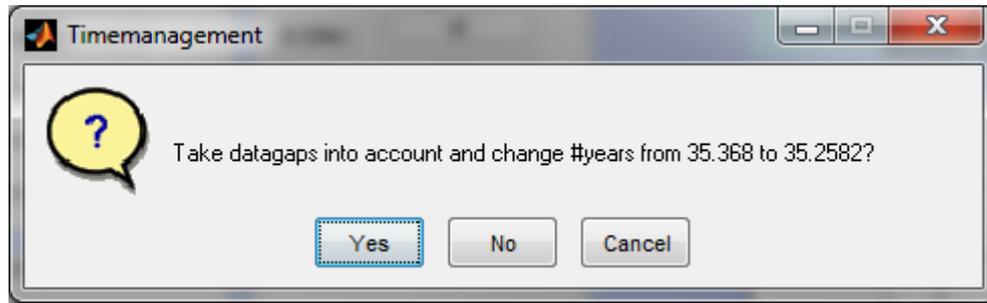


Figure 5-3: Time management data series



5.1.2 Set outputdir

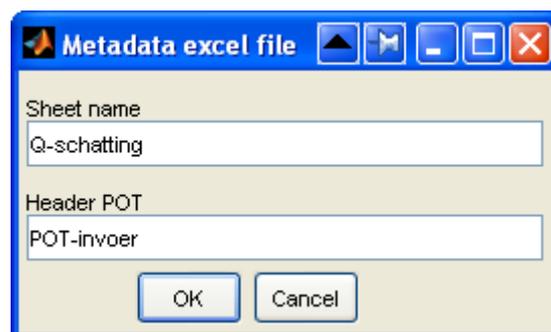
This button allows the user to select the output directory. This directory will contain all graphics and dataset generated by the EVA-tool.

5.1.3 Load POT

POT values selected in previous sessions or projects can be imported by “Load POT”. These POT values have to be in a binary mat file generated by the EVA-tool or an Excel file. In case of an Excel file a dialog window will appear to ask for metadata (Figure 5-4): the sheet name that contains the POT values and the text string in the cell above the POT values.

An important factor in a conditional distribution is the length of the time series from which the POT are obtained. In case there’s no time series available a popup window will occur to ask for these length. Hence a time series is not necessary in the further analysis.

Figure 5-4: Metadata needed to import POT values stored in a excel file



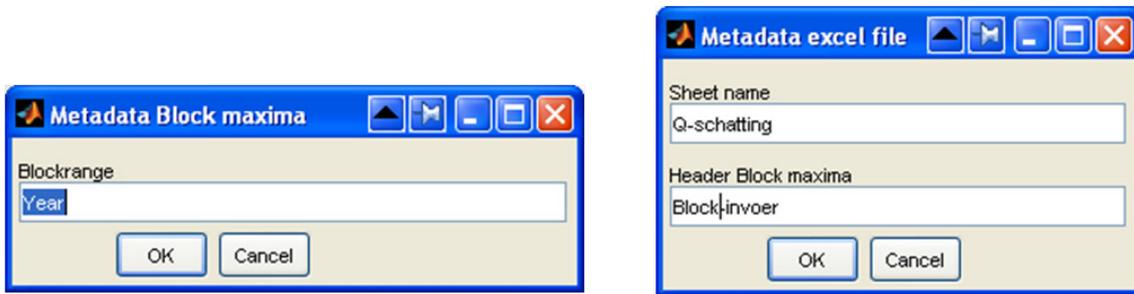
5.1.4 Load Block

The import of block maxima has the same properties as the import of the POT values. The metadata requested for the import of block maxima is the block range and, in case of an Excel file, also sheet name and header of the block maxima (Figure 5-5) are required. The available choices for the block range are:

- Year
- Month
- Day

Other input will result in a warning message and a request to import the proper block range string.

Figure 5-5: Metadata needed to import block maxima.



5.2 Selection of extremes

The selection of extremes can be divided in the selection of POT (peak over threshold) values and block maxima. The selection of POT values is more complicated, but POT values have the advantage to contain more information of the extremes. The smaller a dataset becomes, the more important this advantage will be. Block maxima on the other hand are still commonly used and are included for comparison. The POT and block maxima selection can be done on the same dataset and the results can be compared. The graphical user interface (GUI) for this selection is displayed in Figure 5-6.

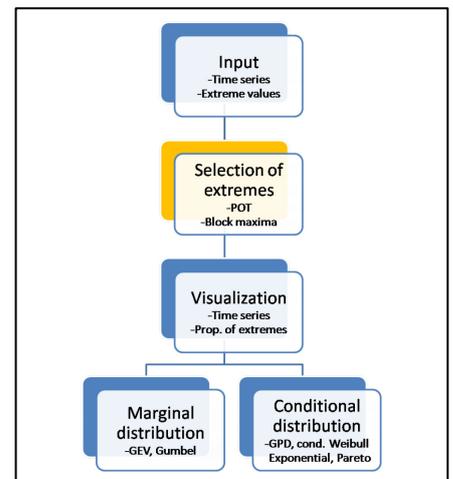
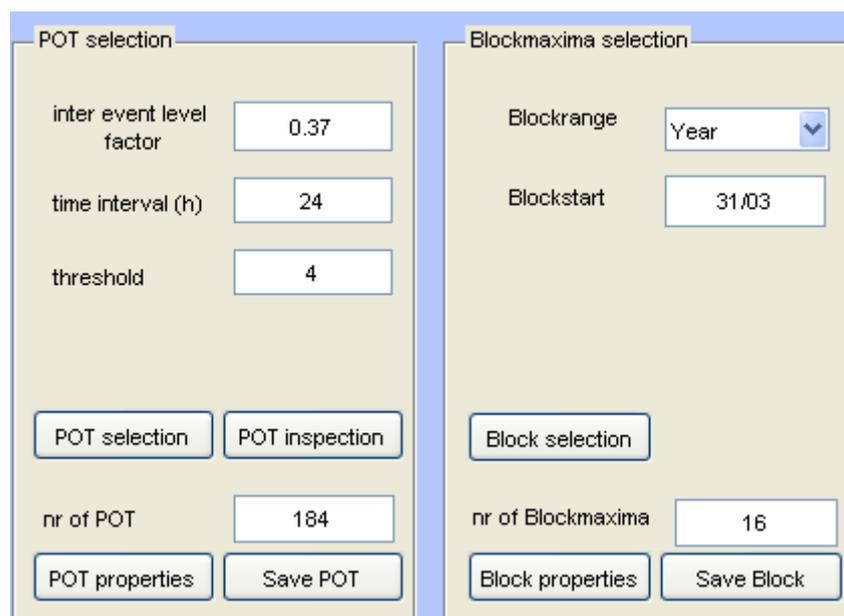


Figure 5-6: Pot and Block maxima selection GUI



5.2.1 POT selection

Peak Over Threshold (POT) selection searches for all extreme events above a threshold. The independency of the events is guaranteed by use of an inter-event level and a time interval. The tool automatically sets the threshold to 60 % of the maximum value. This initial value can be adapted by the user. The initial threshold has to be sufficiently low in order to include enough events. In the next step the initial threshold will be increased to the optimal threshold. Variables with a strong “memory” (autocorrelation), like discharge or water level, will generate multiple successive measurements with extreme values. These extreme values are strongly dependent and belong to 1 (independent) extreme event. The maximum value of one extreme event is a POT value. To make a distinction between events there are 2 extra selection criteria implemented: the inter-event level and a time lag:

- Inter-event level:

The default value of the inter-event level ratio is 0.37. This is the multiplier by which the lowest of two successive possible POT values is being multiplied to get the inter-event level. The minimum value between 2 successive POT values has to be lower than this inter-event level.

- Time lag:

The time interval is the minimum amount of time between 2 POT values.

The number of POT values is displayed in the assigned box and the POT record can be saved in a binary mat format. The POT values can be visualized and checked in a table by the “POT inspection” button. If no time series is available, for example if the POT values are loaded directly, a table will open (Figure 5-7). This table allows the user to deselect unwanted POT values by clicking on the check box next to the POT value and eventually pressing the OK button.

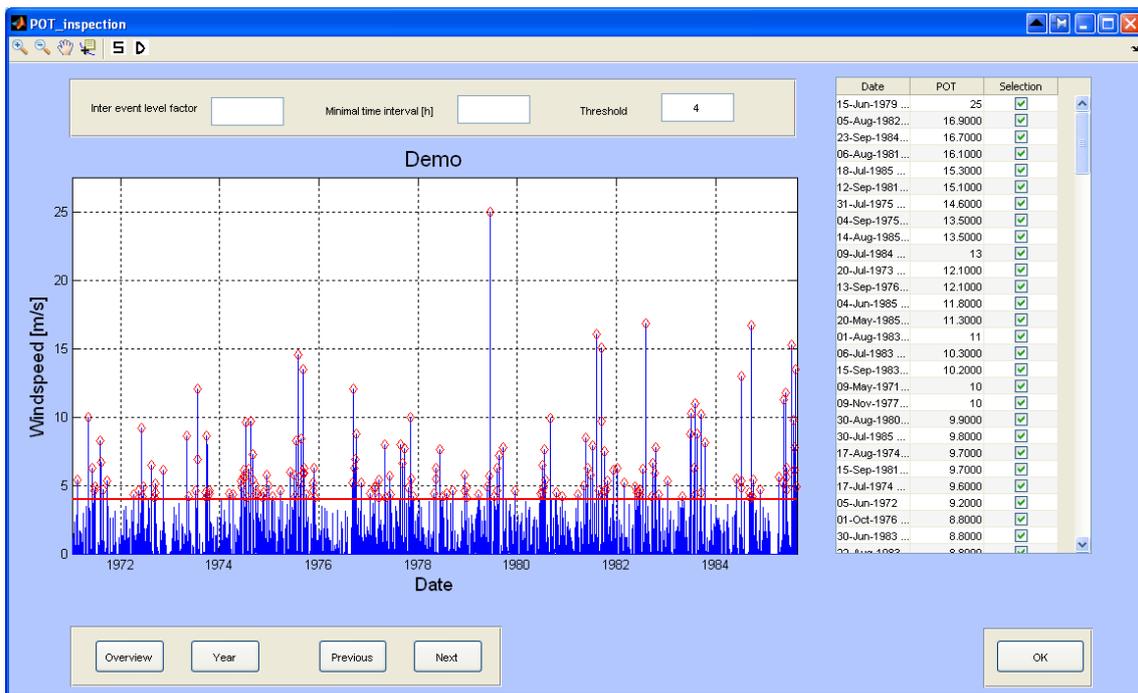
Figure 5-7: POT Table

Date	POT	Selection
15-Jun-1979 19:00:00	25	<input checked="" type="checkbox"/>
05-Aug-1982 18:00:00	16.9000	<input checked="" type="checkbox"/>
23-Sep-1984 21:00:00	16.7000	<input checked="" type="checkbox"/>
06-Aug-1981 19:00:00	16.1000	<input checked="" type="checkbox"/>
18-Jul-1985 19:00:00	15.3000	<input checked="" type="checkbox"/>
12-Sep-1981 15:00:00	15.1000	<input checked="" type="checkbox"/>
31-Jul-1975 23:00:00	14.6000	<input checked="" type="checkbox"/>
04-Sep-1975 08:00:00	13.5000	<input checked="" type="checkbox"/>
14-Aug-1985 13:00:00	13.5000	<input checked="" type="checkbox"/>
09-Jul-1984 18:00:00	13	<input checked="" type="checkbox"/>
20-Jul-1973 20:00:00	12.1000	<input checked="" type="checkbox"/>
13-Sep-1976 12:00:00	12.1000	<input checked="" type="checkbox"/>
04-Jun-1985 18:00:00	11.8000	<input checked="" type="checkbox"/>
20-May-1985 17:00:00	11.3000	<input checked="" type="checkbox"/>
01-Aug-1983 08:00:00	11	<input checked="" type="checkbox"/>
06-Jul-1983 17:00:00	10.3000	<input checked="" type="checkbox"/>
15-Sep-1983 06:00:00	10.2000	<input checked="" type="checkbox"/>
09-May-1971 08:00:00	10	<input checked="" type="checkbox"/>
09-Nov-1977 05:00:00	10	<input checked="" type="checkbox"/>
30-Aug-1980 11:00:00	9.9000	<input checked="" type="checkbox"/>
30-Jul-1985 20:00:00	9.8000	<input checked="" type="checkbox"/>
17-Aug-1974 04:00:00	9.7000	<input checked="" type="checkbox"/>

If a time series is available a new window will open which allows the user for a detailed POT inspection. The time series together with the POT values and the initial threshold can be found in the figure, while the POT

values with date are ranked from largest to smallest in a table. The buttons °Overview° and °Year° will switch from the total time range to a time range of 1 year. The standard Matlab zoom and pan functions are also available in the menu bar. The buttons °Previous° and °Next° will set the display respectively the previous and the next year. A first click on a POT value will enlarge the red diamond while a second click will deselect the POT value. A click on the time series will select a possible POT value by selecting the maximal value in a time range of 3h around the click. The minimal time interval and the inter event level factor of the selected maximum will be displayed. A second click on the big red diamond will add the maximum to the set of POT values.

Figure 5-8: POT inspection GUI



5.2.2 Selection of Block maxima

A block maximum is the maximum value in a fixed time interval, the block range. The block start value gives the start date, month or hour of the blocks. It is recommendable, in case of a yearly storm season, to use a one year time interval setting the block start in between two storm seasons. This way the possibility that one extreme event is selected for the block maximum of two successive years is fairly negligible. The available ranges with corresponding block start formats are displayed in Table 5-1. The number of block maxima is displayed in the assigned box and the block maxima record can be saved in a binary mat format.

Table 5-1: Available block ranges and corresponding block start formats

Block range	Block start
Year	dd/mm
Month	dd
Day	hh

5.3 Visualization

The tool uses the strong Matlab visualisation functionalities. The imported time series can be visualised together with the selected POT values and block maxima. The title and the label of the abscissa and ordinate can be adapted by the figure panel. Toolbars at the top of the GUI contains zoom or pan and a data cursor to obtain the data for a specific point. A useful tool for time series is the horizontal zoom. This function enables the user to zoom on the x-axis leaving the y-axis unchanged. This function is enabled by selecting the zoom function, right clicking on the figure, selecting zoom options and horizontal zoom. The figure can be exported in the .fig and .png format (Figure 5-9).

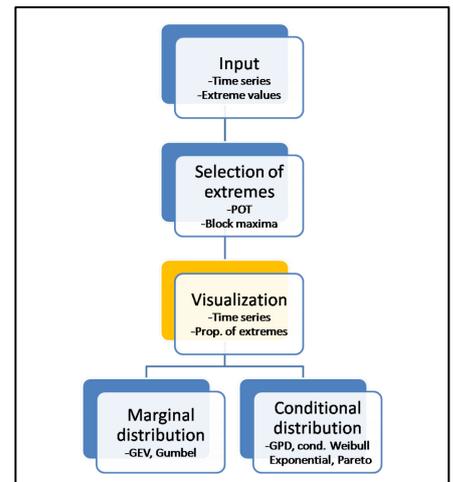
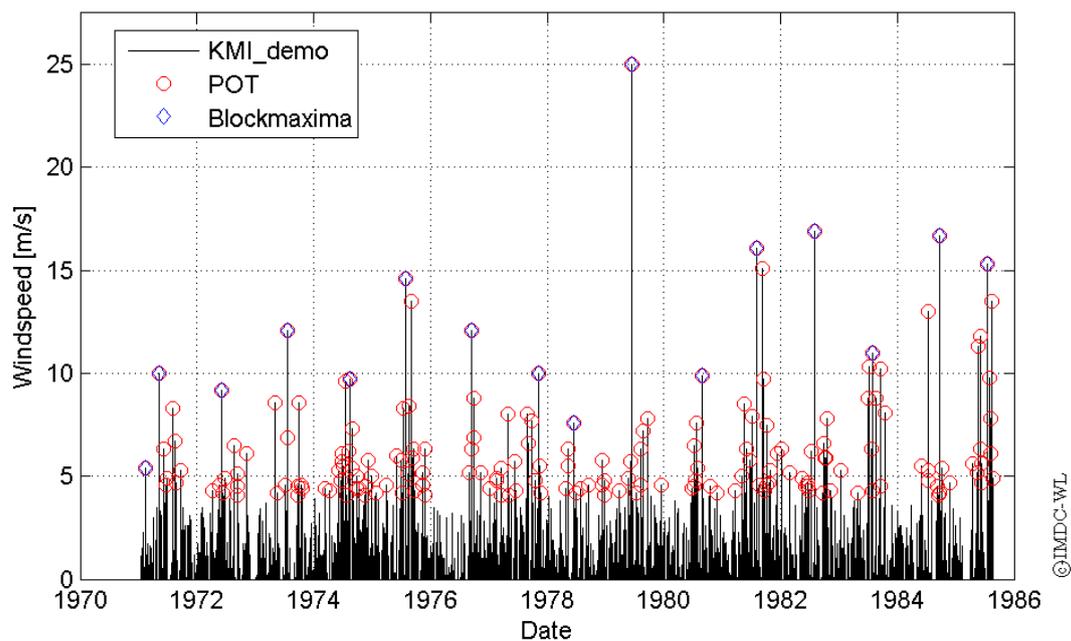


Figure 5-9: Example of a time series with POT values and Block maxima.

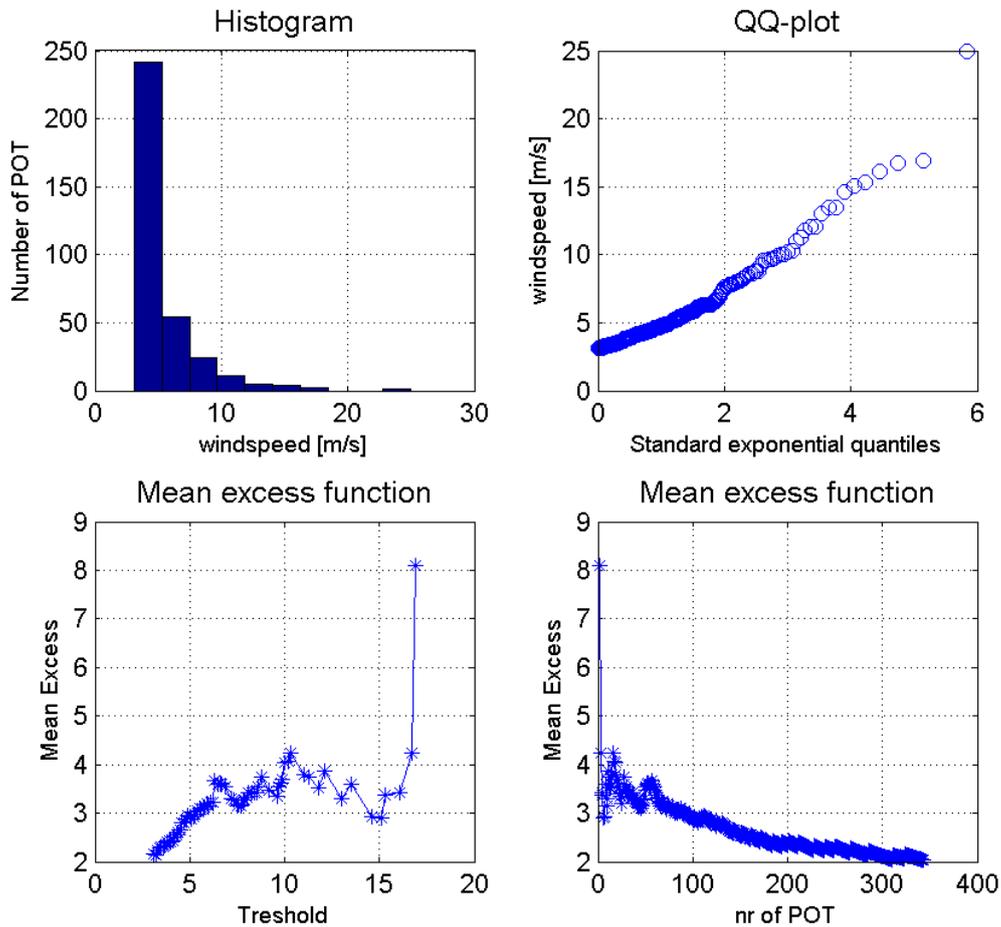


To get more insight in the selected extreme values, the POT values or block maxima, 4 property plots are made by clicking buttons “POT properties” and “Block properties” (see example in Figure 5-10). These plots are a histogram, a QQ plot with the standard exponential quantiles, a mean excess function in function of the threshold and one in function of the number of POT. The QQ plot and the mean excess function will provide a first indication with respect to the appropriate distribution. If the QQ plot is linear the exponential distribution will be an appropriate approximation of the tail behaviour of the extreme values. In case of a steeper increase than linear the appropriate distribution will have a heavier tail than the exponential distribution (GPD $\xi > 0$, conditional Weibull $\tau < 1$ or Pareto distribution). In case of a smaller increase the appropriate tail will be lighter (GPD $\xi < 0$, conditional Weibull $\tau > 1$). The distribution can also be determined by comparing the mean excess function in the lower left corner of Figure 5-10 with the theoretical mean excess functions in Figure 4-2.

The mean excess function above the highest thresholds will usually be strongly influenced by the sample variation and therefore not reliable with respect to the overall trend.

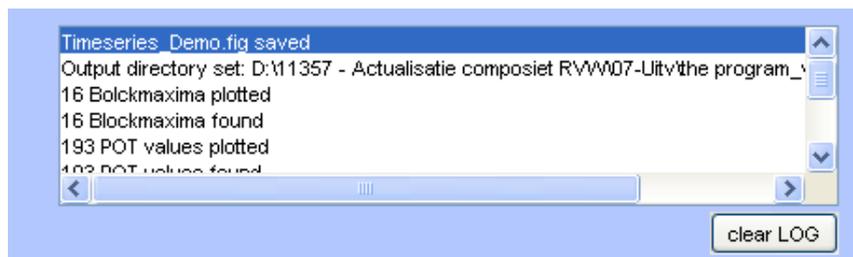
The example in Figure 5-10 displays a linear trend above a threshold around 7 m/s. This inclination is also noticeable in the QQ plot. The optimal threshold will most likely be situated around 7 m/s. The sample variation is responsible for the deviation of the mean excess functions above the threshold of 16 m/s.

Figure 5-10: Example of POT properties figure.



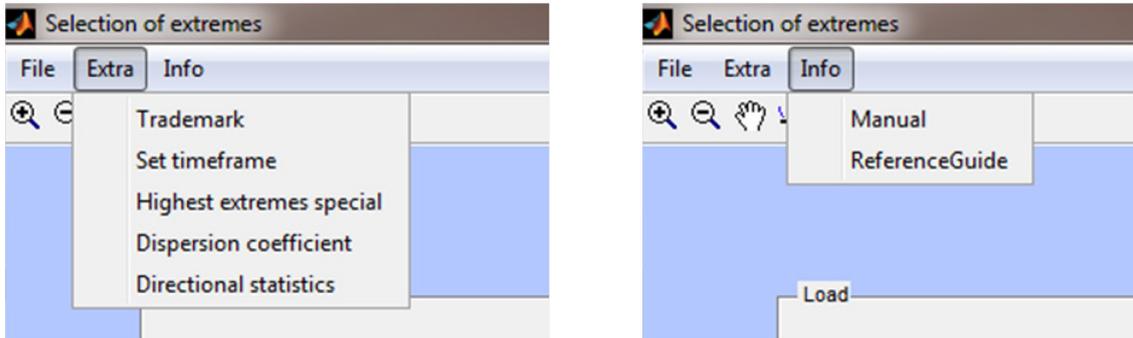
All important manipulations are logged and visualized in the left corner. These manipulations include data import, selection of extremes, visualizations, etc. (Figure 5-11)

Figure 5-11: Log file



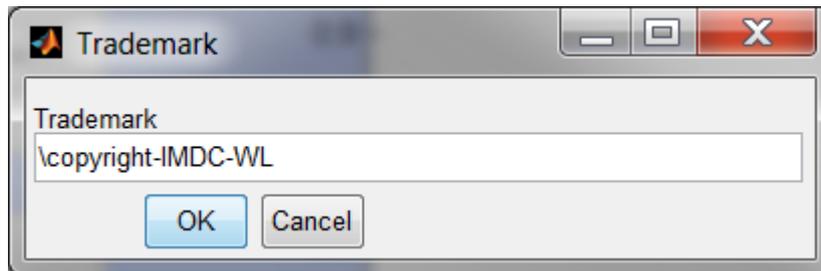
5.4 Extra functionalities

Figure 5-12: Extra menu tabs and info



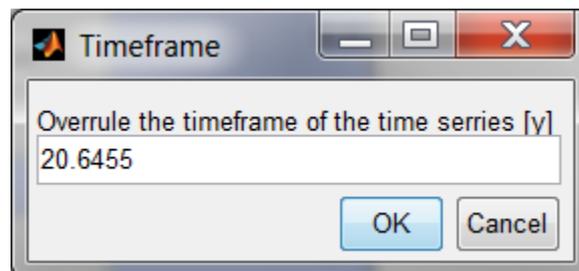
There are some extra functionalities implemented in the menu tab 'Extra' (Figure 5-13). The first one 'Trademark' (Figure 5-13) allows the user to change the trademark included in the figures. The default is `\copyright-IMDC-WL` which gives ©-IMDC-WL in a TEX interpreter.

Figure 5-13: Trademark input



The second option set timeframe allows the user to overrule the automatic calculated timeframe of the time series (Figure 5-14). This automatic calculation uses the first and last timestamp in the dataset. In case of a time series with large gaps this will give an overestimation of the total time frame which is important to calculate the return period.

Figure 5-14: Overrule timeframe

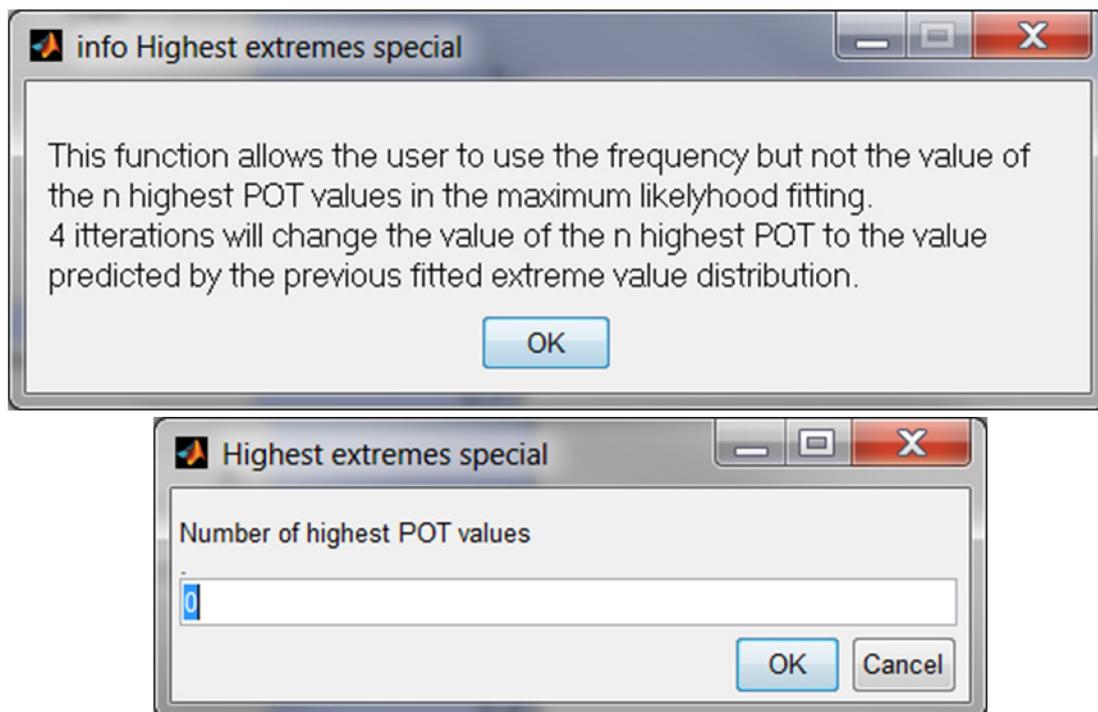


The third option should only be used by experts in extreme value distributions. The 'highest extremes special' allows the user to take the frequency but not the value of the n highest POT values into account (Figure 5-15).

Though an iterative procedure the value of the n highest POT values is replaced by the value predicted by the extreme value distribution for the same frequency as the POT value. Due to the iterative procedure the fitting of the extreme value distribution will take more time.

A fourth option is the visualization of the dispersion coefficient in function of the threshold. . The POT values should follow a Poisson point distributions which has a dispersion coefficient of 1. If the dispersion coefficient is less than 1 the process is under-dispersed. The pattern of occurrence is more regular than the randomness associated with a Poisson process. If the dispersion coefficient is higher than 1 the process is over-dispersed. The pattern of occurrence is more irregular than the randomness associated with a Poisson process. The data is clustered in certain intervals.

Figure 5-15: Highest extremes special



The POT values can be divided into different directions by the fifth option under extra, Directional statistics. This function displays the tab in Figure 5-16. A filter time series with dimensions [time value direction] in mat or inv format is needed to assign directions to the POT values. The time range of this filter has to be at least the range of the POT values. The number of directional categories has a default value of 16. The button 'Filter' divides the POT in the categories. The number of POT in each category is visualized in the table under the column amount. This table can be exported with the button 'Save filter properties'. Only the selected categories will be taken into account in the further analysis.

Figure 5-16: Directional division of the POT values

The screenshot shows a 'Filter' dialog box with a control panel on the left and a data table on the right. The control panel includes buttons for 'Load filter', 'Filter', 'Select all', 'Deselect all', 'Save filter properties', and 'OK'. It also has input fields for '# categories' (16) and 'Missing Value' (-999). The data table has the following structure:

	min	max	amount	selected
1	348.7500	11.2500	0	<input checked="" type="checkbox"/>
2	11.2500	33.7500	36	<input checked="" type="checkbox"/>
3	33.7500	56.2500	0	<input checked="" type="checkbox"/>
4	56.2500	78.7500	0	<input checked="" type="checkbox"/>
5	78.7500	101.2500	0	<input checked="" type="checkbox"/>
6	101.2500	123.7500	0	<input checked="" type="checkbox"/>
7	123.7500	146.2500	0	<input checked="" type="checkbox"/>
8	146.2500	168.7500	1	<input checked="" type="checkbox"/>
9	168.7500	191.2500	8	<input checked="" type="checkbox"/>
10	191.2500	213.7500	14	<input checked="" type="checkbox"/>
11	213.7500	236.2500	25	<input checked="" type="checkbox"/>
12	236.2500	258.7500	97	<input checked="" type="checkbox"/>
13	258.7500	281.2500	24	<input checked="" type="checkbox"/>
14	281.2500	303.7500	3	<input checked="" type="checkbox"/>
15	303.7500	326.2500	0	<input checked="" type="checkbox"/>
16	326.2500	348.7500	0	<input checked="" type="checkbox"/>

6 Part 2: Conditional distributions

Four conditional distributions are implemented in the EVA-tool:

- GPD
- Pareto
- Conditional Weibull
- Exponential

In this part of the analysis the appropriate distribution with the optimal threshold value has to be determined. Six graphics are available to aid this choice: the root mean square error (RMSE) as a function of the number of POT values, the parameter values with confidence intervals as a function of the number of POT values, a probability plot, a QQ plot and the resulting return level plot (Figure 6-1).

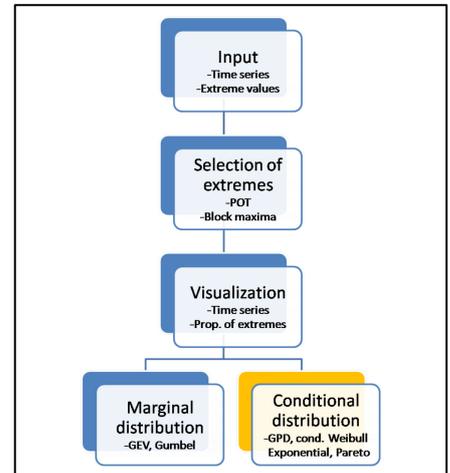
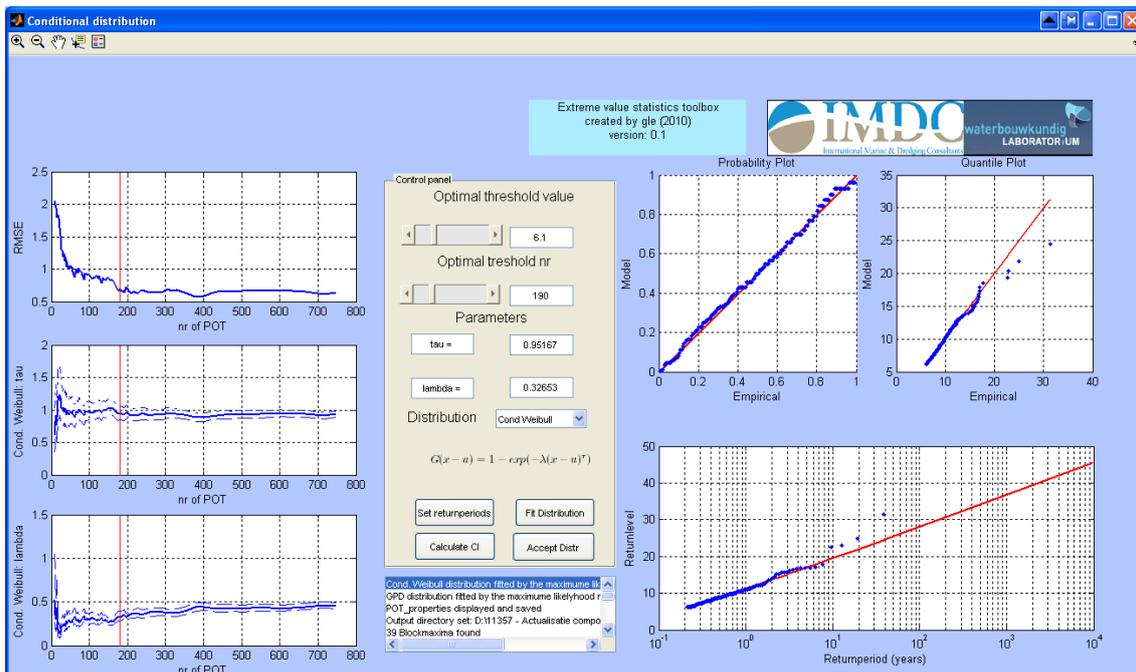


Figure 6-1: Conditional distribution fitting GUI



6.1 Practical use

The control panel in the middle of the Conditional distribution tool contains all control buttons and the obtained parameter values.

- **Listbox distributions:**
This allows the user to select a distribution.
- **Set return period:**
This button allows the user to change the values of the return periods. The default return periods are 1, 2, 5, 10, 25, 50, 100, 500 1000, 2500, 4000 and 10000 years. The return level corresponding to these return periods will be included in the output.
- **Fit distribution:**
This button makes a maximum likelihood fit of the selected distribution for every possible threshold. The results of the parameter estimations and RMSE are visualized in the left-hand graphics. The result of the automatically selected optimal threshold is visualized in the right-hand figures. This can be a time consuming process for large sets of POT values.
- **Calculate Confidence Interval:**
The confidence interval of the return level is calculated with A bootstrap method (default) or by the use of the parameter confidence intervals determined in the maximum likelihood estimation (delta method).
- **Accept distribution:**
This button will start the output generation module. This module will generate 10 output files (Table 6-1). The files will overwrite eponymous files in the output directory.

Table 6-1: Output files conditional distribution (example in Annex A)

File	Description	Extension
Distribution fiche	Formulas, parameters, return level plot, probability plot, QQ-plot, histogram and return level table	.png
Parameter fiche	RMSE as a function of nr of POT Parameter as a function of nr of POT	.png
Stratified sampling of synthetic event	Return period as a function of return level Return level as a function of return period Return level as a function of freq of exceedance Synthetic event as a function of freq.	.png
Distribution parameters	Cdf formula Return level formula Parameter values Return level table	.txt and .mat
Selected POT values	Sorted POT above optimal threshold	.txt and .mat
Synthetic extremes	Table with synthetic extremes	.txt
Copula/ synthetic events input	Time series POT above optimal threshold Distribution parameters	.mat

6.2 Selecting the appropriate distribution

Some guidelines are available to select the appropriate distribution. One should keep in mind that the end goal is to obtain a distribution with a realistic estimation of return levels for the high return periods. The red line in the lower right corner of the Conditional distribution fitting GUI (see Figure 6-1) has to be a good approximation of the data and needs to have 'realistic' values in the extrapolation domain. The tool allows to easily swap between the different distributions. When in doubt one can try to fit all of them for comparison.

In a standard analysis it is recommendable to fit the GPD distribution as a first step. The GPD distribution covers the entire conditional domain. The drawback is consequently a large confidence interval (Coles 2001). The estimation of the GPD ξ parameter will give a good estimation of the tail behaviour. This should be confirmed in the mean excess function, as explained in the section 4.2.1. If the confidence interval of ξ contains zero, a conditional Weibull or exponential distribution is likely.

The same argumentation can be adopted for the Conditional Weibull distribution. This distribution covers the conditional domain with tails that decrease exponentially. If the confidence interval of the parameter τ contains the value 1, it is recommendable to use the exponential distribution. The exponential distribution has less parameters to fit; accordingly the confidence interval will be smaller.

6.3 Optimal threshold

Once the appropriate distribution is selected, the optimal threshold has to be determined. The optimal threshold of a distribution is the threshold above which the POT value behaviour is in best confirmation with the distribution. This optimal threshold has to be selected in an area with constant parameter value, small parameter confidence intervals and a local minimum of the RMSE. The tool will automatically select the optimal threshold corresponding to the minimum RMSE. The optimal threshold with corresponding optimal threshold number (the rank of the POT value equal to the optimal threshold) and the parameter values are displayed in the Control panel shown in Figure 6-2.

Figure 6-2: Conditional distribution control panel

The screenshot shows a software control panel for conditional distribution analysis. It features several input fields and buttons. The 'Optimal threshold value' is set to 15.2692, and the 'Optimal threshold nr' is 92. Under 'Parameters', 'ksi' is -0.15112 and 'sigma' is 1.9937. The 'Distribution' is set to GPD. The 'Confidence interval method' is Bootstrap. The GPD distribution function is displayed as $G(x - w) = 1 - (1 + \frac{\xi(x-w)}{\sigma})^{-\frac{1}{\xi}}$. At the bottom, there are four buttons: 'Set returnperiods', 'Fit Distribution' (highlighted with a dashed border), 'Calculate CI', and 'Accept Distr'.

7 Part 3: marginal distributions

Two marginal distributions are implemented in the EVA-tool:

- GEV
- Gumbel

These marginal distributions can be fitted through the block maxima. The goodness of fit is evaluated by a probability plot, a QQ plot, a histogram and a return level-return period plot with corresponding table (Table 7-1). The GUI to fit the marginal distribution with the demo data is visualised in Figure 7-1.

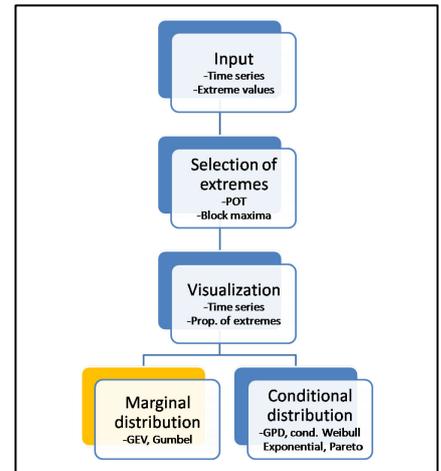
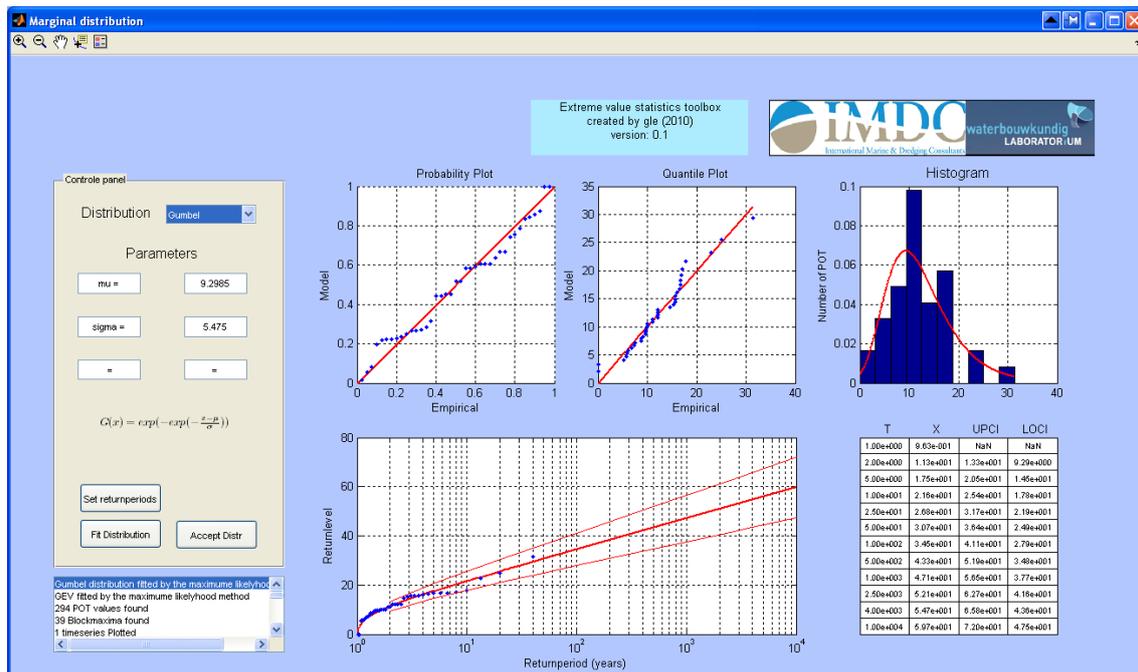


Figure 7-1: Marginal distribution fitting GUI



7.1 Practical use

The control panel of the Marginal distribution GUI is situated on the left hand side. The GUI will automatically fit the GEV distribution with confidence intervals. The fit of a marginal distribution is easier because the optimal threshold is irrelevant. There are four user controls:

- **Distribution popup menu:**
This menu allows the user to choose between the GEV and the Gumbel distributions. The fit will be executed when selecting a distribution.
- **Set return periods:**
The return periods can be manually adapted. The default levels are 1, 2, 5, 10, 25, 50, 100, 500, 1000, 2500, 4000 and 10000.
- **Fit distribution:**
This control button will refit the chosen distribution, e.g. after a change of return levels.
- **Accept Distribution:**
This button will start the output generation module. This module will generate 7 output files (Table 7-1). These files will overwrite eponymous files in the output directory.

Table 7-1: Output files marginal distribution (examples in Annex B)

File	Description	Extension
Distribution fiche	Formulas, parameters, return level plot, probability plot, QQ-plot, histogram and returnlevel table	.png
Stratified sampling of synthetic event	Return period as a function of return level Return level as a function of return period Return level as a function of freq of exceedance Synthetic event as a function of freq.	.png
Distribution parameters	Cdf formula Return level formula Parameter values Return level table	.txt and .mat
Selected block maxima	Block maxima	.txt and .mat
Synthetic extremes	Table with synthetic extremes	.txt

7.2 Selecting the appropriate distribution

There are some guidelines to select the appropriate distribution. One should keep in mind that the end goal is to obtain a distribution with a realistic estimation for the return levels for the high return periods. The thick red line in the return level plot of the distribution fitting GUI (see Figure 7-1) has to be a good approximation of the data and needs to have 'realistic' values in the extrapolation domain. The tool allows for easily swapping between the different distributions. When in doubt one can try to fit all of them for comparison. The wide domain of the GEV distribution causes wider confidence intervals for return levels (Coles 2001). This drawback makes the GEV less appropriate for applications which take these return levels into account (probabilistic design, risk analysis, ...).

8 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 modelling of extreme values. London: Springer-Verlag DC 2011

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

Matlab, 2011, <http://www.mathworks.com/help/toolbox/compiler/f12-999353.html#br2jauc-33> (retrieved on 15/05/2011).

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

Appendix A: Output conditional distribution

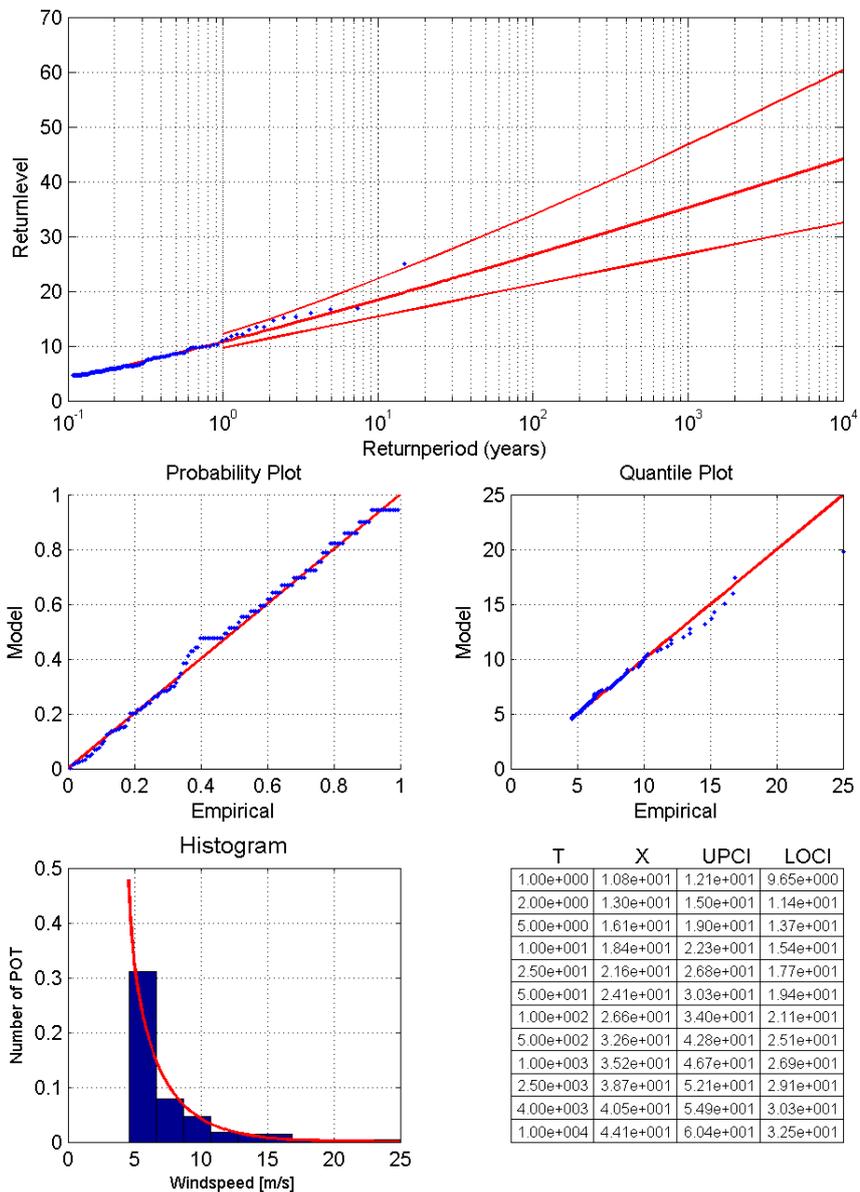
Figures: Demo

Figure A-1: EVA-distribution sheet

Cond. Weibull distribution

$$cdf : 1 - Pr(x > u + y | x > u) = 1 - exp(-\lambda(x - u)^\tau) \quad \begin{matrix} \tau = 0.88572 \\ \lambda = 0.44029 \\ u = 4.5 \\ A = 14.6213 \\ k = 137 \end{matrix}$$

$$Returnlevel : X = u + (\frac{1}{\lambda} \log(\frac{T+k}{A}))^{1/\tau}$$



©IMD-C-WL

Figure A-2: Distribution parameters as function of number of POT

Cond. Weibull distribution Parameters ifo the number op POT values

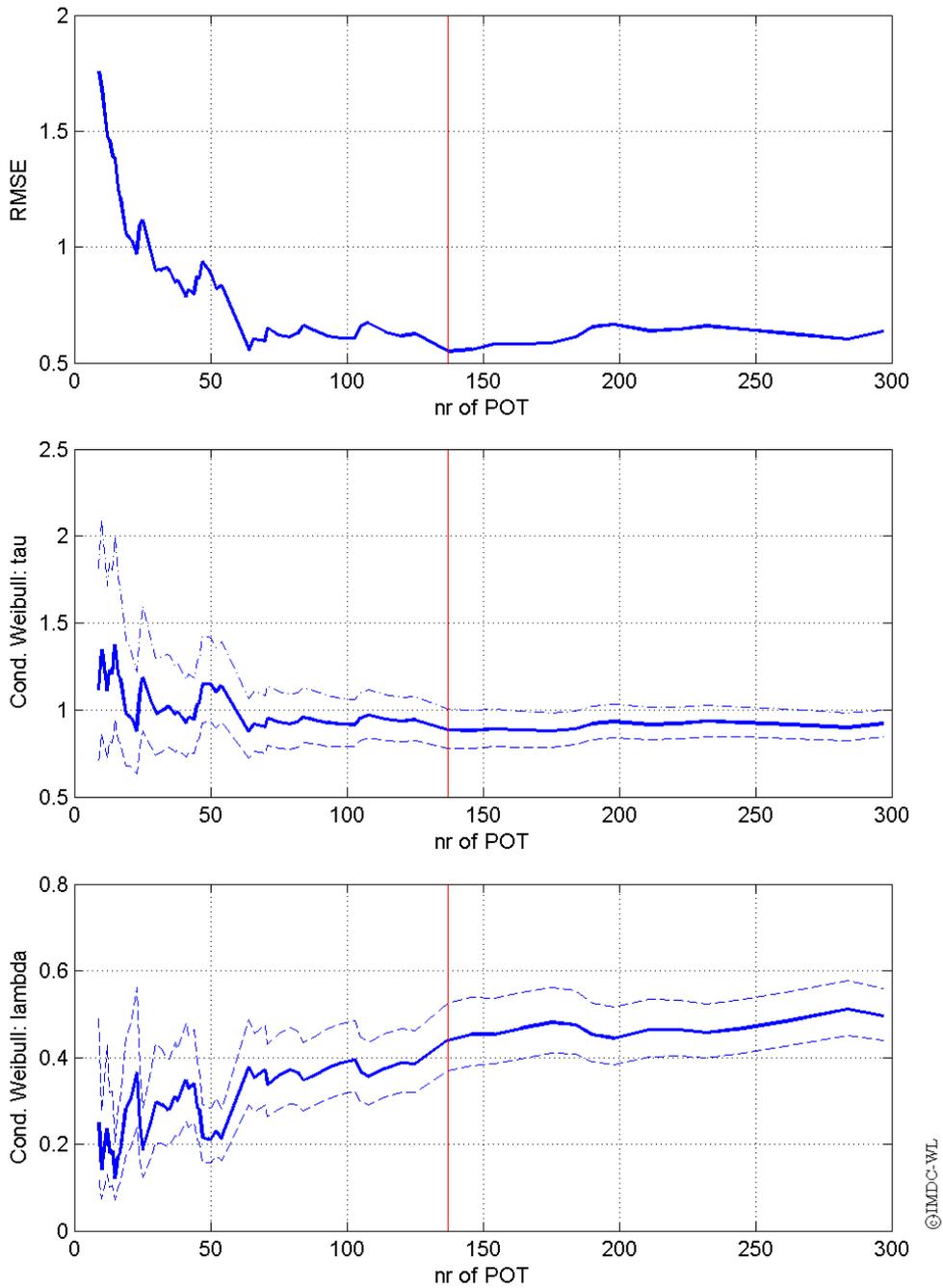
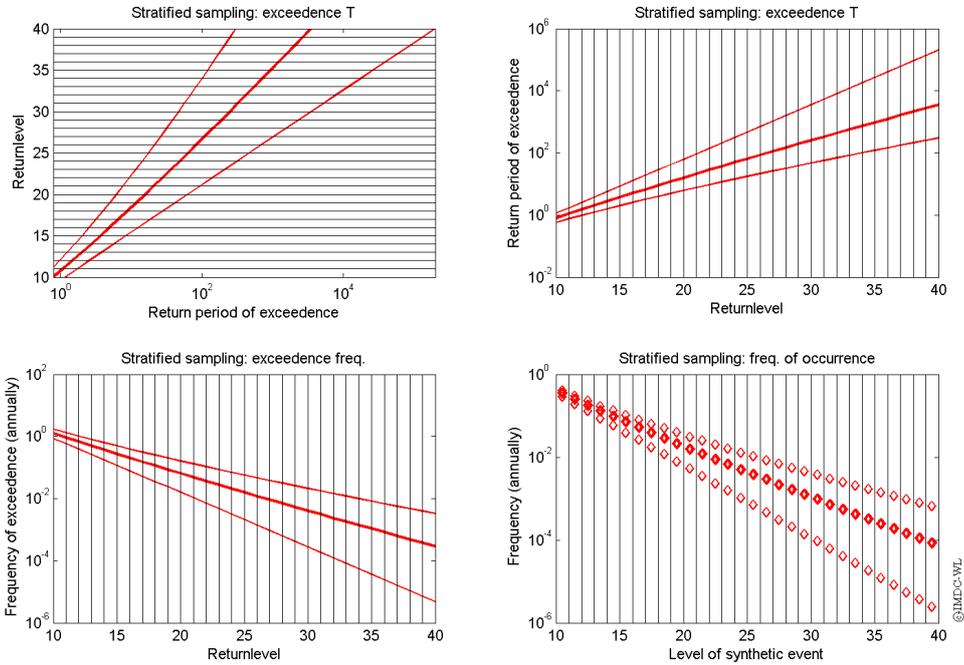


Figure A-3: Stratified sampling



Text files: Demo

Figure A-4: Distribution parameters

```

$cdf: 1-Pr(x>u+y|x>u)=1-exp(-\lambda (x-u)^{\tau})$
$returnlevel: X=u+(\frac{1}{\lambda}log(\frac{T*k}{A}))^{(1/\tau)}$
\tau=0.88572$
\lambda=0.44029$
u=4.5$
A=14.6213$
k=137$
T      X      UPCI      LOCI
1.000000 10.767888      12.125624      9.651896
2.000000 13.000486      14.985243      11.391425
5.000000 16.056869      19.041090      13.685090
10.000000 18.433889      22.286612      15.415697
25.000000 21.648028      26.779322      17.698556
50.000000 24.126942      30.314636      19.422420
100.000000 26.642243      33.956666      21.144074
500.000000 32.605927      42.782520      25.134564
1000.000000 35.221710      46.728134      26.850585
2500.000000 38.718807      52.065821      29.116978
4000.000000 40.528876      54.854812      30.278667
10000.000000 44.087385      60.386369      32.541919
    
```

Figure A-5: Selected POT values

```

** POT values selected by EU-tool
** IMDC - WL
**
** -----
** 2011/06/24 16:07:32
** -----
**
25.000
16.900
16.700
16.100
15.300
15.100
14.600
13.500
13.500
13.000
12.100
12.100
11.800
11.300
11.000
10.300
10.200
10.000
10.000
10.000
    
```

Figure A-6: Synthetic extremes created by stratified sampling

Level	Freq. (annually)	Freq. upper limit	Freq. lower limit
1.050000e+001	3.476972e-001	4.039161e-001	2.860914e-001
1.150000e+001	2.492609e-001	3.012261e-001	1.922618e-001
1.250000e+001	1.800458e-001	2.273294e-001	1.291610e-001
1.350000e+001	1.308661e-001	1.732937e-001	8.671973e-002
1.450000e+001	9.562666e-002	1.332403e-001	5.818830e-002
1.550000e+001	7.019897e-002	1.032059e-001	3.902091e-002
1.650000e+001	5.174189e-002	8.046035e-002	2.615321e-002
1.750000e+001	3.827559e-002	6.308665e-002	1.752029e-002
1.850000e+001	2.840618e-002	4.971654e-002	1.173187e-002
1.950000e+001	2.114391e-002	3.935936e-002	7.852738e-003
2.050000e+001	1.578083e-002	3.128898e-002	5.254351e-003
2.150000e+001	1.180738e-002	2.496732e-002	3.514590e-003
2.250000e+001	8.854735e-003	1.999188e-002	2.350170e-003
2.350000e+001	6.654670e-003	1.605903e-002	1.571097e-003
2.450000e+001	5.011237e-003	1.293799e-002	1.050012e-003
2.550000e+001	3.780739e-003	1.045218e-002	7.015858e-004
2.650000e+001	2.857417e-003	8.465657e-003	4.686717e-004
2.750000e+001	2.163180e-003	6.873181e-003	3.130141e-004
2.850000e+001	1.640193e-003	5.592884e-003	2.090120e-004
2.950000e+001	1.245505e-003	4.560775e-003	1.394918e-004
3.050000e+001	9.471365e-004	3.726624e-003	9.309495e-005
3.150000e+001	7.212188e-004	3.050849e-003	6.213032e-005
3.250000e+001	5.498983e-004	2.502144e-003	4.146494e-005
3.350000e+001	4.197919e-004	2.055666e-003	2.767314e-005
3.450000e+001	3.208404e-004	1.691633e-003	1.846868e-005
3.550000e+001	2.455041e-004	1.394252e-003	1.232575e-005
3.650000e+001	1.880580e-004	1.150876e-003	8.226039e-006
3.750000e+001	1.442051e-004	9.513494e-004	5.489948e-006
3.850000e+001	1.106902e-004	7.875007e-004	3.663917e-006
3.950000e+001	8.529453e-005	6.527361e-004	2.445249e-006

Appendix B Output marginal distribution

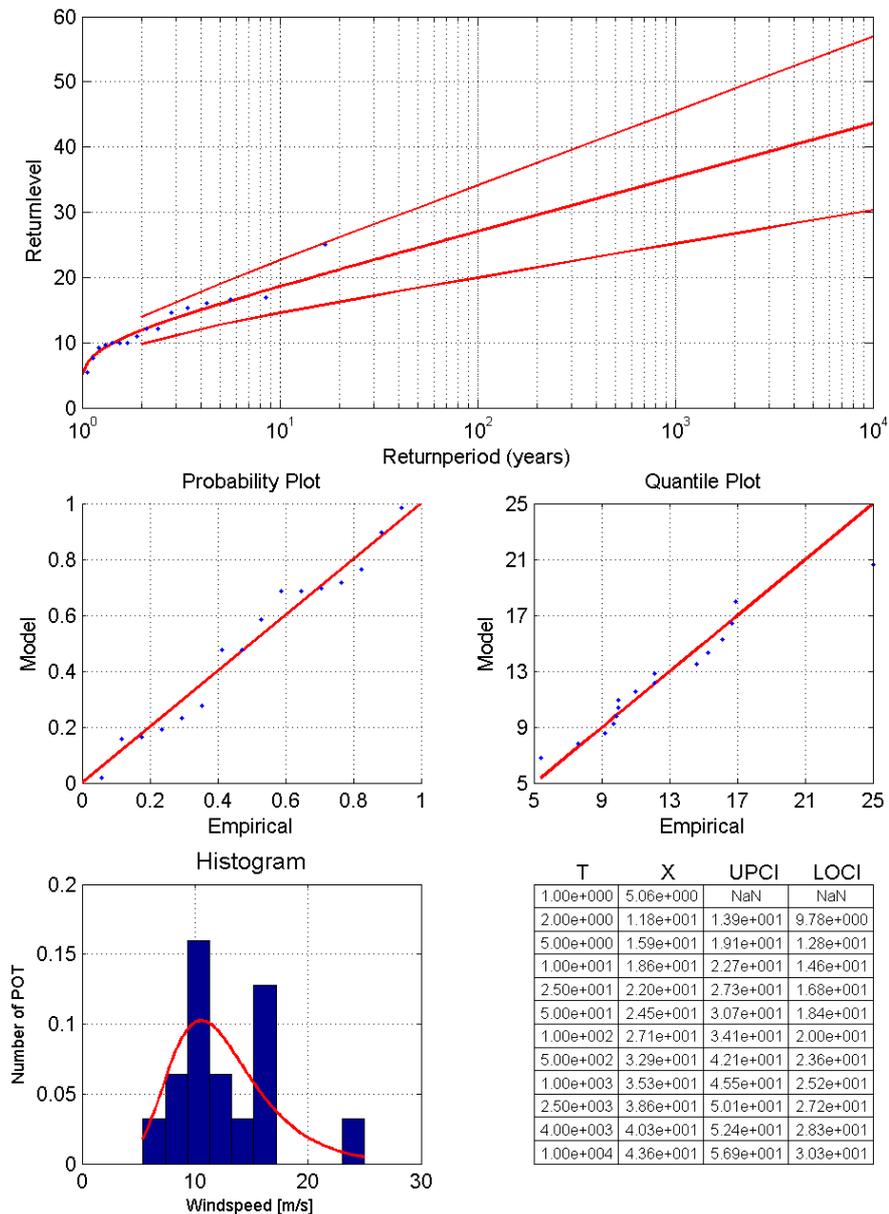
Figures: Demo

Figure B-1: EVA-distribution sheet

Gumbel distribution

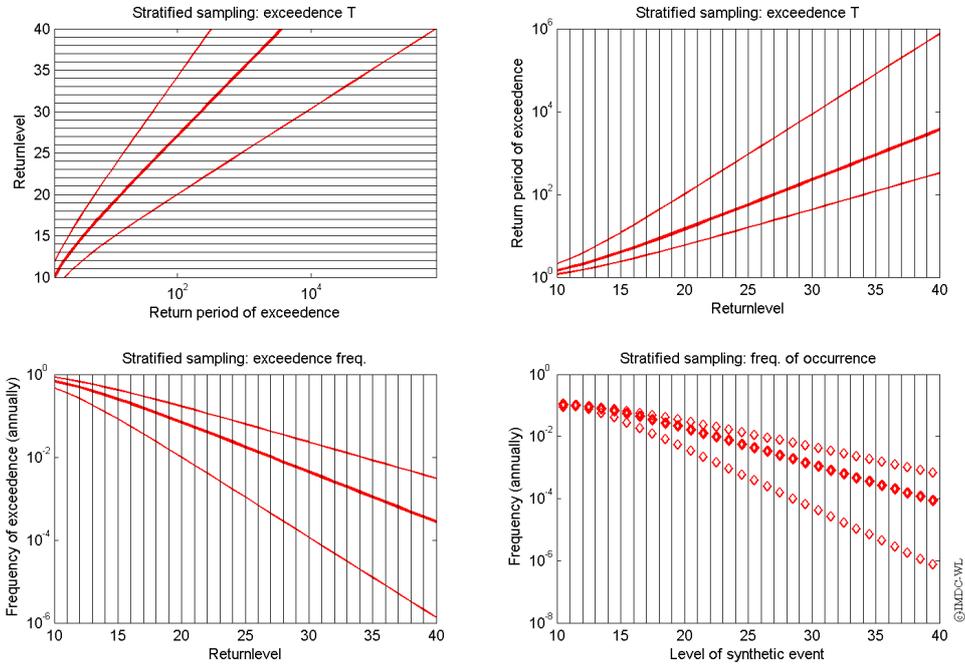
$$cdf : G(x) = G(x) = \exp(-\exp(-\frac{x-\mu}{\sigma})) \quad \mu = 10.5334$$

$$Returnlevel : x = \mu - \sigma \log(-\log(1 - \frac{1}{T})) \quad \sigma = 3.592$$



©IMDC-WL

Figure B-2: Stratified sampling



Text files: Demo

Figure B-3: Distribution parameters

```

$cdf:      G(x)=G(x)=exp(-exp(-\frac{x-\mu}{\sigma}))$
$returnlevel: x=\mu -\sigma log(-log(1-\frac{1}{T}))$
$\mu=10.5334$
$\sigma=3.592$
T          X          UPCI      LOCI
1.000000  5.064415          NaN      NaN
2.000000  11.849900        13.915850    9.783950
5.000000  15.921191        19.087155    12.755227
10.000000 18.616740        22.675920    14.557560
25.000000 22.022573        27.280172    16.764974
50.000000 24.549215        30.721536    18.376893
100.000000 27.057200        34.149822    19.964578
500.000000 32.852771        42.098801    23.606742
1000.000000 35.344367        45.523127    25.165607
2500.000000 38.636779        50.052035    27.221522
4000.000000 40.325310        52.376059    28.274561
10000.000000 43.616912        56.908495    30.325330
    
```

Figure B-4: Selected Block maxima

```

** Block values selected by EU-tool
** IMDC - WL
**
** -----
** 2011/06/24 16:27:19
** -----
**
** 25.000
** 16.900
** 16.700
** 16.100
** 15.300
** 14.600
** 12.100
** 12.100
** 11.000
** 10.000
** 10.000
** 9.900
** 9.700
** 9.200
** 7.600
** 5.400
    
```

Figure B-5: Synthetic extremes created by stratified sampling

Level	Freq. (annually)	Freq. upper limit	Freq. lower limit
1.050000e+001	1.020783e-001	8.978566e-002	1.145393e-001
1.150000e+001	9.884800e-002	9.386882e-002	9.861141e-002
1.250000e+001	9.018380e-002	9.074468e-002	7.808958e-002
1.350000e+001	7.864201e-002	8.384836e-002	5.803000e-002
1.450000e+001	6.626584e-002	7.528644e-002	4.119987e-002
1.550000e+001	5.440473e-002	6.625571e-002	2.833631e-002
1.650000e+001	4.379548e-002	5.743589e-002	1.906850e-002
1.750000e+001	3.473292e-002	4.920649e-002	1.264171e-002
1.850000e+001	2.723607e-002	4.176112e-002	8.295552e-003
1.950000e+001	2.117531e-002	3.517430e-002	5.405147e-003
2.050000e+001	1.635686e-002	2.944500e-002	3.504441e-003
2.150000e+001	1.257298e-002	2.452661e-002	2.264153e-003
2.250000e+001	9.628589e-003	2.034769e-002	1.459134e-003
2.350000e+001	7.353008e-003	1.682594e-002	9.385984e-004
2.450000e+001	5.603279e-003	1.387726e-002	6.029250e-004
2.550000e+001	4.263036e-003	1.142116e-002	3.868910e-004
2.650000e+001	3.239408e-003	9.383828e-003	2.480606e-004
2.750000e+001	2.459296e-003	7.699461e-003	1.589442e-004
2.850000e+001	1.865744e-003	6.310596e-003	1.017289e-004
2.950000e+001	1.414697e-003	5.167816e-003	6.510942e-005
3.050000e+001	1.072261e-003	4.229105e-003	4.167191e-005
3.150000e+001	8.124668e-004	3.459059e-003	2.667123e-005
3.250000e+001	6.154760e-004	2.828046e-003	1.707035e-005
3.350000e+001	4.661665e-004	2.311399e-003	1.092552e-005
3.450000e+001	3.530318e-004	1.888671e-003	6.992649e-006
3.550000e+001	2.673273e-004	1.542968e-003	4.475497e-006
3.650000e+001	2.024137e-004	1.260369e-003	2.864447e-006
3.750000e+001	1.532540e-004	1.029427e-003	1.833329e-006
3.850000e+001	1.160286e-004	8.407429e-004	1.173384e-006
3.950000e+001	8.783845e-005	6.866121e-004	7.509994e-007

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