Assessing the Tail Behaviour of Empirical Distributions Using the Extreme Value Theory

Alaa El-Sadek[1]

[1]Water Resources Management Program, Manama, Kingdom of Bahrain, College of Graduate Studies, Arabian Gulf University

Abstract

In this research, an extreme value analysis methodology was used to recognize the anomalies in tail behaviour of the flood frequency distribution in an easy and visual way by means of the so-called Quantile-Quantile plots (Q-Q plots). The type of distribution and the optimal threshold level can be derived with this method in a most efficient way. The methodology was applied to data from Tanzania (catchment of Victoria Lake in Tanzania; which includes Mara, Mwanza and Kagera Regions, daily discharge values from January, 1954 to December, 1986) for testing and first application. Based on the expertise built up for the Tanzanian data, the extreme value analysis was applied using Kenyan discharge data. The methodology is especially interesting because of the visual nature and it can be used afterwards in combination with the traditional Method Of Moments (MOM) and Maximum Likelihood Method (ML) methods. In that way, the advantages of both methods could be combined and an efficient extreme value analysis is made. The research results indicated that, to get a precise result applicable to the hydraulic engineering practice, it is strongly recommended to use daily data (as for Victoria Lake catchment in Tanzania; which includes Mara, Mwanza and Kagera Regions). This is because of the peak flow that will determine the flooding. The monthly values (as only available for the Kenyan stations) are only indirect measures.

The results indicated that, monthly averages should not necessarily follow the extreme value theory as this theory is only valid for maxima of a large number of variables; monthly averages can be considered as 'volumes' and these can also follow distributions other than the Generalized Pareto Distribution (GPD). By analyzing the behavior of the data points in different types of Q-Q plots, the asymptotic behavior of the tail of the distribution can be determined and discrimination can be made between Pareto-, exponential and bounded-type distributions. The discrimination is mainly based on an estimation of the extreme value index as the slope of the linear path in a so-called UH-plot. This slope is estimated by a weighted linear regression. Finally, it can be concluded that, the optimal threshold level can be derived easily as the threshold level that minimizes the mean-squared-error of the regression.

Key words**:** extreme value analysis; Q-Q plots; tail behavior; optimal threshold; flood frequency

1. INTRODUCTION

Over the past several decades, attempts have been made to develop integrated theories (i.e., models) of the water management systems. These models represent approximations to (i.e., simulations of) the actual processes, process interactions, and matter and energy exchanges that take place in the real world. The amount of details contained in simulation models varies widely, depending on the needs and objectives of the projects under which they were developed (Shaffer, 1995). Simulation models attempt to approximate real world processes and their interactions at the mechanistic level. They are extremely important components of decision support systems, and some expert systems may also contain simulation components. Simulation models usually contain logical relationships derived from the subject knowledge base and also may include expert systems (El-Sadek, 2001).

In the immission modeling of receiving waters, the accurate description of extreme surface water states (flooding, deteriorated water quality) and their recurrence rates is of primary importance (Willems, 2000). In this study, an extreme value analysis methodology (Willems, 1998) was used to recognize the anomalies in tail behaviour of the flood frequency distribution in an easy and visual way by means of the so-called Quantile-Quantile plots (Q-Q plots). The type of distribution and the optimal threshold level can be derived with this method in an efficient way. In extreme value analysis the tail of the distribution, describing the probability of occurrence of extreme events, is analysed and modeled by a separate distribution. The considered extremes might exist of extreme rainfall intensities, storm volumes, water levels, discharges, water quality parameters, etc. In practice, its analysis is time consuming as the determination of the 'most highly probable' type of extreme value distribution is difficult. By tradition, some plausible distributions (e.g. Gumbel, Exponential, Generalized Pareto, Weibull, Pearson) are 'tried out' and statistical tests are performed to find the 'best' distribution. The threshold level xt, above which the distribution is calibrated, is chosen arbitrarily. Very inaccurate extrapolations outside the range of observations can result from this anomaly. The anomaly, through a wrong type of distribution, is often caused by a nonoptimal threshold xt. This threshold is often chosen as the minimum threshold that is needed for the application. However, it should be determined in a statistical optimal way by maximizing the validity of the distribution above this threshold. Recently, a methodology has been worked out by Beirlant et al. (1996) to recognise the anomalies in tail behaviour in an easy and visual way by means of the so-called Q-Q plots. The type of distribution and the optimal threshold level can be derived in the same way. These features of the methodology were also stressed by Caers (1996), who applied the methodology to the extreme value analyses of diamond deposits (Willems, 2000).

The methodology was applied to data from Tanzania (Victoria Lake in Tanzania; which includes Mara, Mwanza and Kagera Regions, daily discharge values from January, 1954 to December, 1986) for testing and first application. Based on the expertise built up for the Tanzanian data, the extreme value analysis was applied using Kenyan discharge data (eight station, monthly discharge values from January, 1950 to December, 2000).

2. TWO CASE STUDIES IN THE NILE BASIN

The catchment area of the Nile basin is about 2.9 million squared kilometers, which approximately represents one tenth of the area of Africa. The length of the main stream of the river Nile from its mouth on the Mediterranean Sea to its remote source, at the head of river Luvironza, is nearly 6,500 kilometers. The catchment of the river Nile encompasses parts of many countries, namely: Tanzania, Uganda, Rwanda, Burundi, R D Congo, Kenya, Ethiopia, Eritrea, Sudan and Egypt. The proposed methodology has been applied using 33 years daily discharge data of Victoria Lake in Tanzania (1954-1986); which includes Mara, Mwanza and Kagera Regions and 50 years (1950-2000) discharge data of eight rivers (Kenyan Rivers) draining into Vectoria Lake. Figure 1 shows Vectoria Lake within the Nile Basin.

The physiography of the catchments comprises the highland zone (mountains, scarps, hills volcanic foot ridges, footslopes, uplands, plateaus erosional plainsand lowland zones (piedmont plains, river valleys, alluvial plains, and lakeside swamps). The elevation in the catchments ranges from 1130 m, on the lake shore, and 3030 m in the mountains. The slopes are commonly within the range between 0.5% and 30%. The areas have humid to subhumid climate within a mean annual rainfall range between 1000 mm and 1600 mm. The rainfall is trimodal with long and short rains peak periods in March - May, and October - December and the third peak is in August. The mean temperature is 23º C that ranges between 17º and 25º C. The research selected the catchment of Victoria Lake in Tanzania; which includes Mara, Mwanza and Kagera Regions, collectively referred to as the Zone Lake, lying between 1-4º S Lat. and 30-35º E Long.

The soil types, climate, land forms determine the vegetation cover of the areas. Land use varies with topography and agro-climatic conditions. The dominant land uses in the plains include sugarcane growing, both estate and small scale, rice under irrigation and dry season crops like maize, tomatoes, onions, etc. The other major activity is harvesting of papyrus and other species for making mats, seats, fish traps, and thatching materials. On the plains, the vegetation is mainly shrubs and herbs that have adapted to seasonal water logging. The rapid population growth, pollution from the agricultural land use and frequent heavy storms have caused environmental degradation. Forests are being cleared for fuel, timber and agriculture. Soil erosion is high in the higher slopes. There is progressive increase of water demand for the various uses and biodiversity, such as fish, are under threat. Figure 2, shows the Kenyan Rivers that drains into Vectoria Lake.

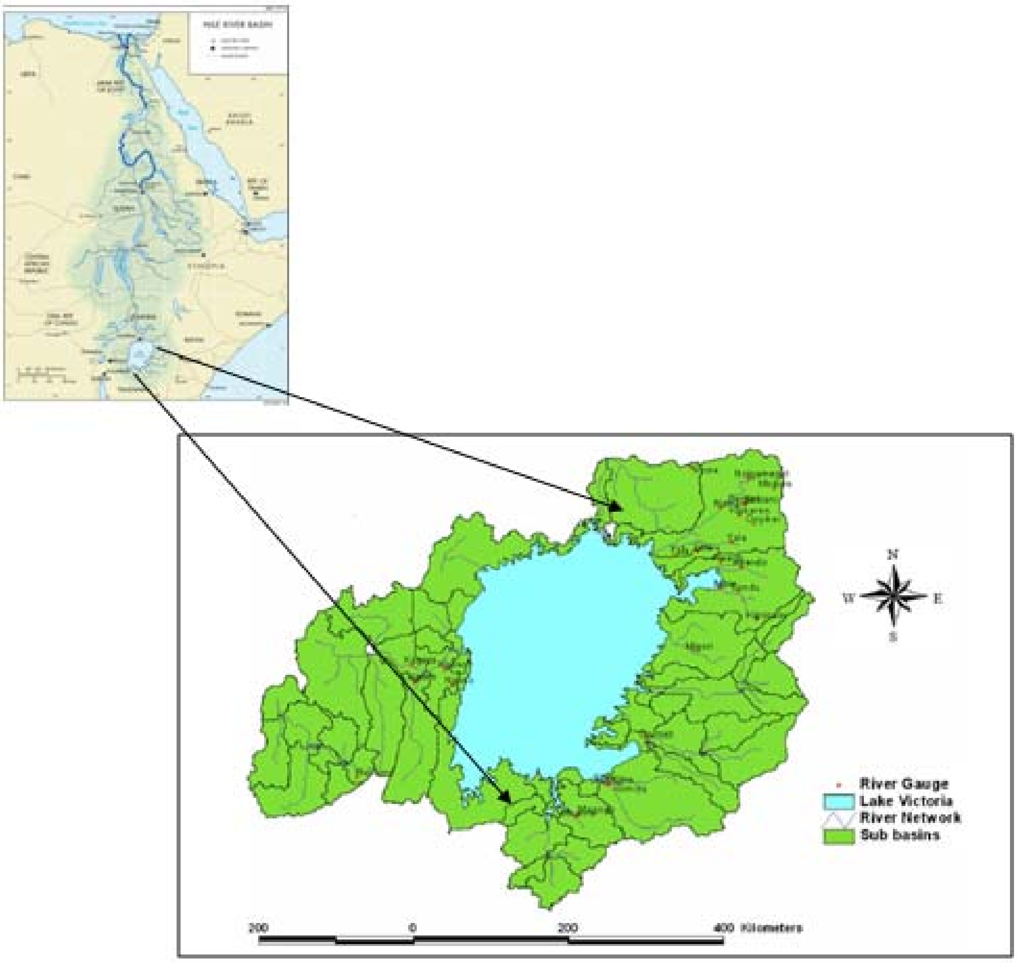

Figure 1: Victoria Lake within the Nile Basin

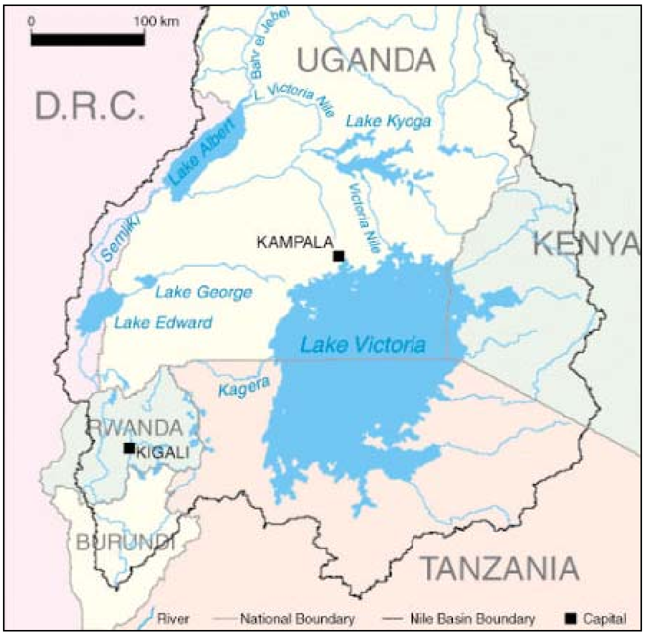

Figure 2: Kenyan and Tanzanian Rivers draining into Victoria Lake

3. MATERIALS AND METHOD

The Quantile-Quantile (Q-Q) plot is a graphical technique for determining if two data sets come from populations with a common distribution. A Q-Q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value. A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the displacement from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions. The advantages of the Q-Q plot are given after (SEMATECH, 2004), as follows:

1. The sample sizes do not need to be equal.
2. Many distributional aspects can be simultaneously tested. For example, shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot. For example, if the two data sets come from populations whose distributions differ only by a shift in location, the points should lie along a straight line that is displaced either up or down from the 45-degree reference line.

In a Q-Q plot, empirical quantiles are shown against theoretical quantiles. The empirical quantiles match the observed extremes xi, i=1,……………,m (x1……………..xm), with pi=i/(m+c) as their corresponding empirical probabilities of exceedance. In this study, the scores c (0 c 1) are given value 1, corresponding to the so-called Weibull plotting position of a quantile plot. For each empirical quantile xi, the theoretical is defined as F-1(1-pi). The function F(x) is the cumulative distribution that is tested in the Q-Q plot and the Q-Q plot is named according to this distribution. In some Q-Q plots, logarithmic transformed quantile values are plotted on both axes. If the observations agree with the considered distribution F(x), the points in the Q-Q plot approach the biosector (Willems, 2000). In practice, however, one wants to test the validity of the distribution F(x) without knowledge of the parameter values. Adapted Q-Q plot are therefore used. In these adapted Q-Q plot, the so-called `quantile function` U(p) is plotted instead of the inverse distribution F-1(1-p). The quantile function U(p) is defined as the simplest function that is linearly dependent on F-1(1-p) and independent on the parameter values of F(x). Quantile functions do not exit for all types of distributions. Expressions to draw exponential, Pareto and Weibull Q-Q plots, are listed hereafter in terms of (U(p);x) or (ln(U(p)); ln(x)), corresponding to the adapted form:

- Exponential Q-Q plot : (-ln(i/m+1);xi), i=1,…………,m (1)
- Pareto Q-Q plot : (-ln(i/m+1); ln(xi)), i=1,…………,m (2)
- Weibull Q-Q plot : (ln(-ln(i/m+1)); ln(xi)), i=1,…………,m (3)

The Q-Q plot is similar to a probability plot. For a probability plot, the quantiles for one of the data samples are replaced with the quantiles of a theoretical distribution. By looking for the Q-Q plot in which the largest observations behave in an asymptotic linear way, the type of the distribution of the studied extremes can be derived. By performing weighted linear regressions in this plot, also the optimal threshold can be derived.

The optimal threshold is the value above which the calibrated extreme value distribution is the most accurate one. The methodology is applied in the Nile basin to data from Tanzania (catchment of Victoria Lake in Tanzania; which includes Mara, Mwanza and Kagera Regions, daily discharge values from January, 1954 to December, 1986) for testing and first application. Based on the expertise built up for the Tanzanian data, the extreme value analysis was applied using Kenyan discharge data.

4. RESULTS

4.1 Catchment of Victoria Lake in Tanzania

First, the daily discharge series was processed to extract the independent Peak-Over-Threshold (POT) values from the time series. Based on an analysis of these POT-values in the exponential Q-Q plot, the Pareto Q-Q plot and the generalized quantile plot (UH-plot), the sign of the extreme value index γ was firstly estimated. The sign was not significantly different from zero. This was concluded from all 3 types of Q-Q plots. From the UH-plot of Figure 3, it is clear that the extreme value index shows a fluctuation around the zero. At the most optimal threshold for estimation of the index (at rank number t=86), the index is almost zero (0.04). This conclusion was confirmed in the other plots. In the exponential Q-Q plot of Figure 4, the points showed a linear tail behaviour. The slope of the points in the exponential Q-Q plot indeed became stable for the higher threshold levels (Figure 5). In the Pareto Q-Q plot, the points showed a continuous bending downwards (Figure 6), with a slope that was continuously decreasing to higher thresholds (Figure 7).

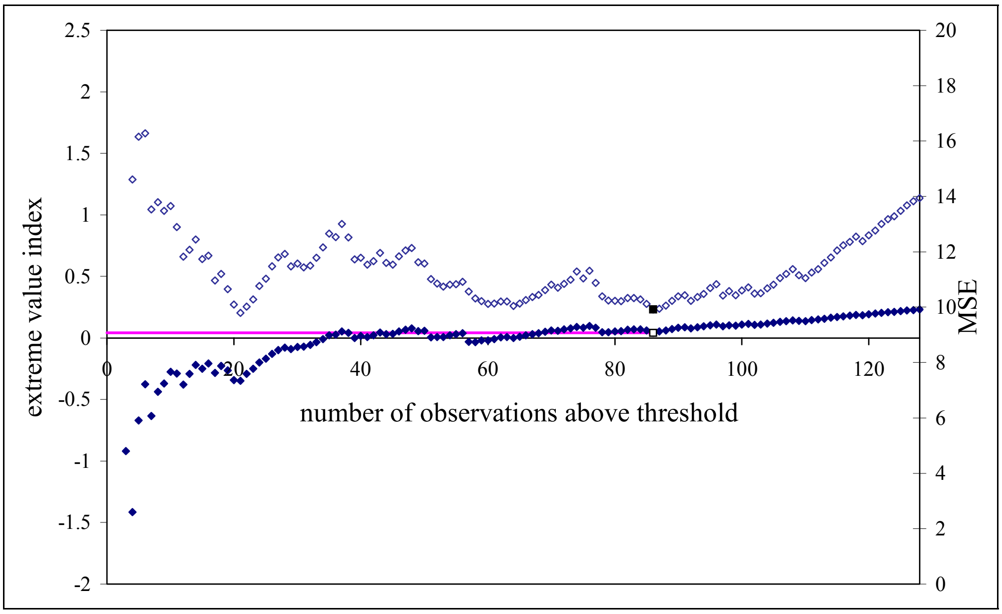

Figure 3: UH-estimation extreme value index; (a) left vert. axis (♦): Hill-type estimation of the slope in the exponential Q-Q plot, (b) right vert. axis (◊): Mean Squared Error of Hill-type regression in the exponential Q-Q plot

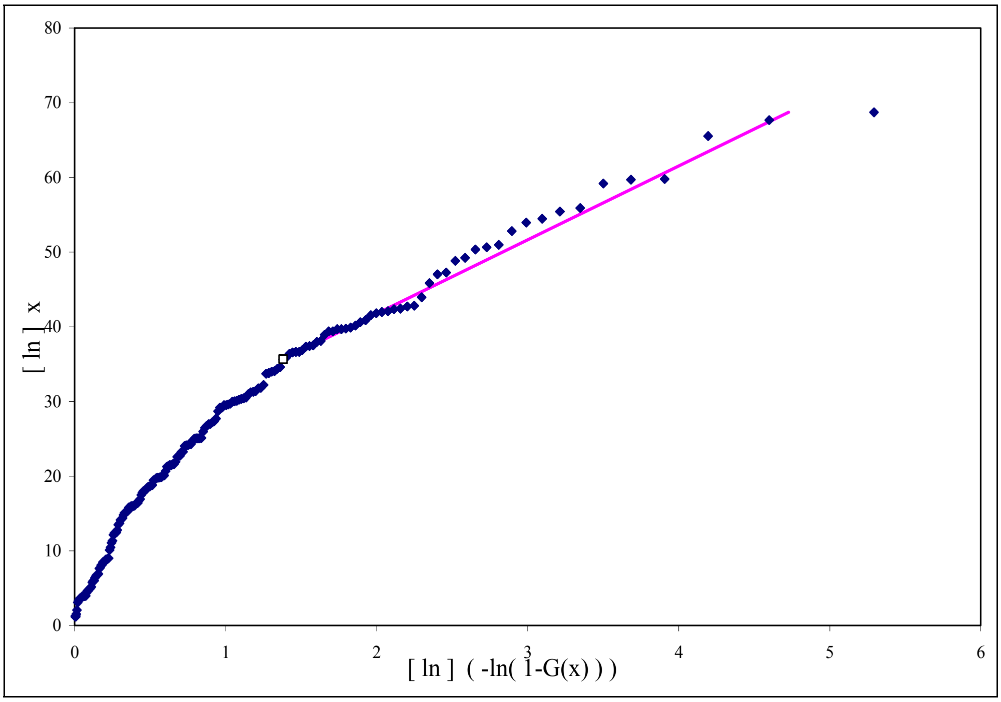

Figure 4: Hill-type regression above the optimal threshold t=50 in the exponential Q-Q plot

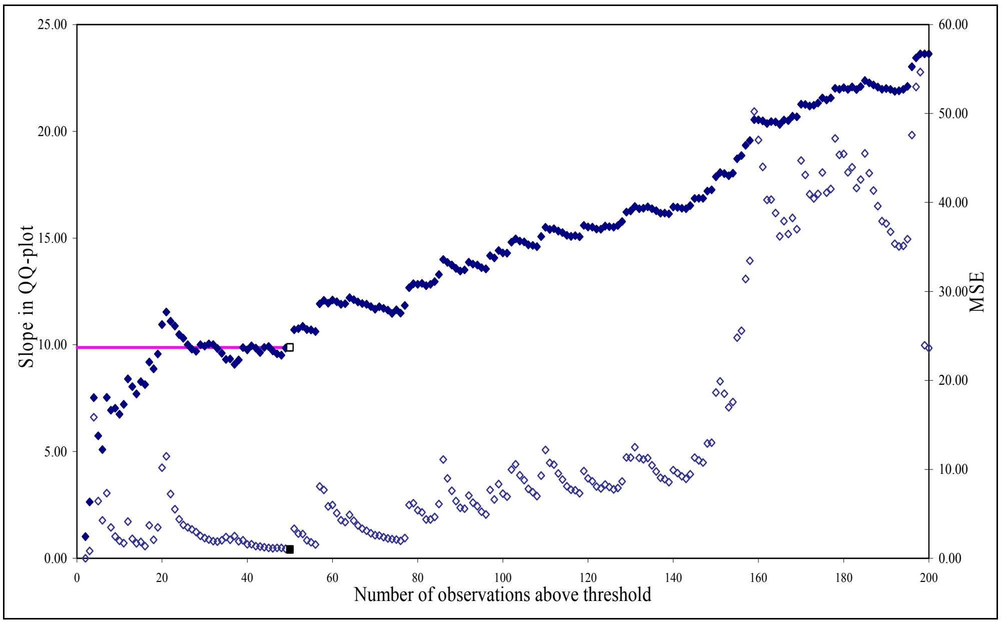

Figure 5: Hill-type estimation parameters exponential distribution; (a) left vert. axis (◆): Hill-type estimation of the slope in the exponential Q-Q plot, (b) right vert. axis (◊): Mean squared error of Hill-type regression in the exponential Q-Q plot

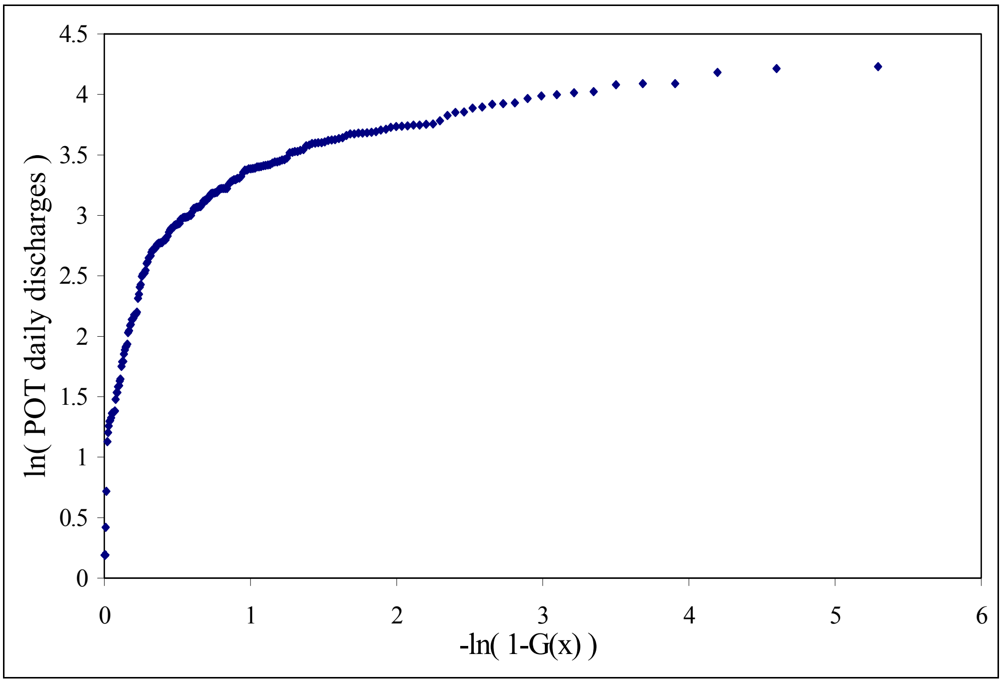

Figure 6: Pareto Q-Q plot (the relation between Generalized Pareto Distribution, G(x) and daily discharges)

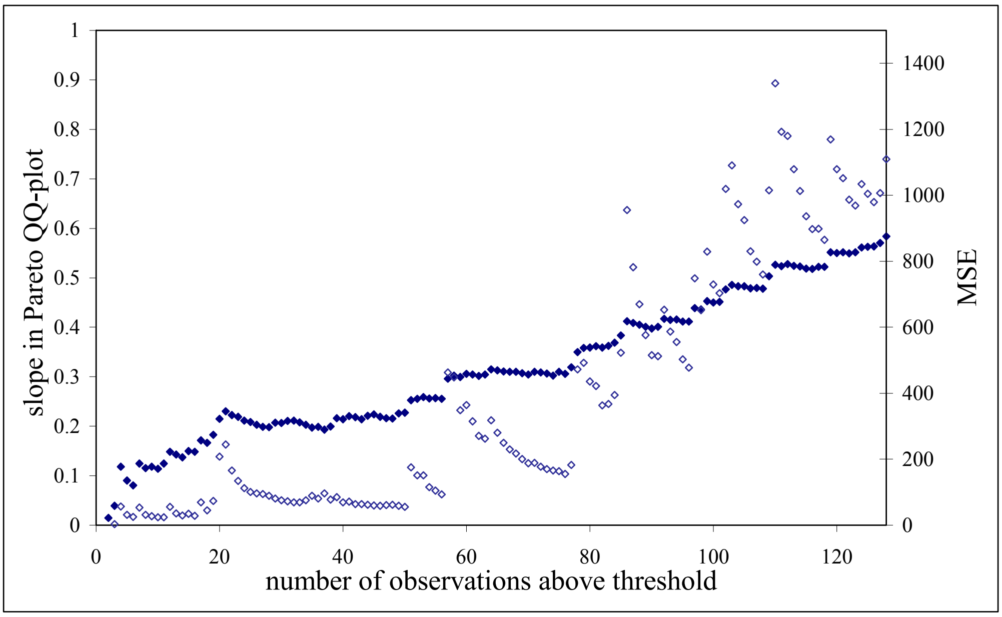

Figure 7: (a) left vert. axis (◆): Hill-type estimation of the slope in the Pareto Q-Q plot, (b) right vert. axis (◊): Mean Squared Error of Hill-type regression in the exponential Q-Q plot

4.2 Kenyan Data (8 stations)

For the Kenyan stations, only monthly averaged data was available. The analysis was done in a similar way as compared to the catchment of Victoria Lake in Tanzania; which includes Mara, Mwanza and Kagera Regions. In total, eight Kenyan stations were considered. A summary of the two most important calibration plots (as explained above) are shown for each of these stations in Figures 8 to 15 for North Awach, South Awach, Nzoia and Sio rivers, respectively. Based on these applications, it has been shown that the methodology has clear advantages. Using traditional methodologies, an extreme value analysis is often time consuming. In most cases, only a few data points of extremes are available. Even when a lot of extremes have been measured, it is difficult to discriminate between different plausible distributions. Mostly, only few distributions can be tested in practice. The uncertainty on the threshold level above which the distribution holds contributes to the time consumption of the analysis.

The applied methodology allows for an easier discrimination between competing distribution models. The methodology is based on quantile-quantile plots (Q-Q plots) and studies the tail behavior of the distribution. By analyzing the behavior of the data points in different types of Q-Q plots, the asymptotic behavior of the tail of the distribution can be determined and a discrimination can be made between Pareto-, exponential- and Weibull-type distributions. The discrimination is mainly based on an estimation of the extreme value index as the slope of the linear path in Q-Q plots (mainly the exponential, Paretao and UH-plots). This slope is estimated by a weighted linear regression. Also the optimal threshold level can be derived easily as the threshold level that minimizes the Mean-Squared-Error (MSE) of the regression. The methodology is interesting because of the visual nature and it can be used afterwards in combination with the traditional Method Of Moments (MOM) and Maximum Likelihood Method (ML) methods. In that way, the advantages of both methods could be combined and an efficient extreme value analysis could be made.

Traditionally, extreme value analysis are performed by testing different ‘typical’ distributions above the threshold levels that are chosen arbitrarily and the parameters are calibrated by a statistical method (e.g. method of moments (MOM), maximum likelihood method (ML), probability weighted moments method). An overview of these statistical methods and their application to the GPD distribution is given by Hosking and Wallis (1987). The traditional methodology has some disadvantages. Only limited number of distributions is tried in practice, possibly all of the same class, and a distribution with a wrong index (extreme value or Weibull index) is often determined. In spite of a good fit in the range of X for which observations are available, an extrapolation outside this range, by means of the distribution, can be erroneous in that case. Extreme events can be strongly overestimated or underestimated. A visualization of the distribution in the Q-Q plot of the class, to which the distribution belongs, makes this error undeniably clear.

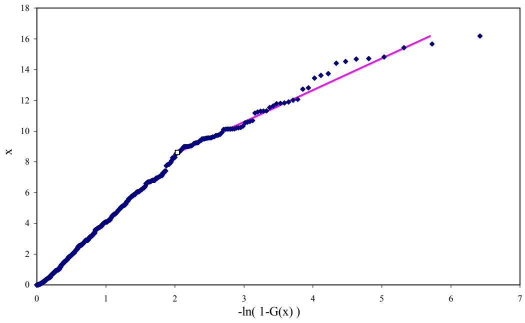

Figure 8: Exponential Q-Q plot (North Awach)

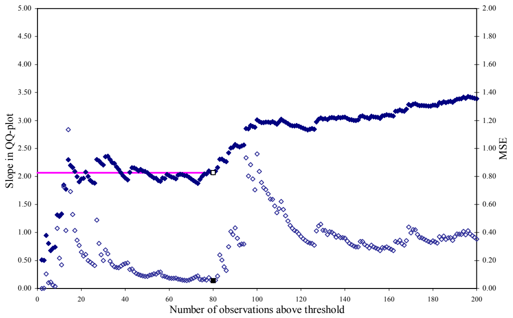

Figure 9: Hill-type estimation slope in exponential Q-Q plot (North Awach)

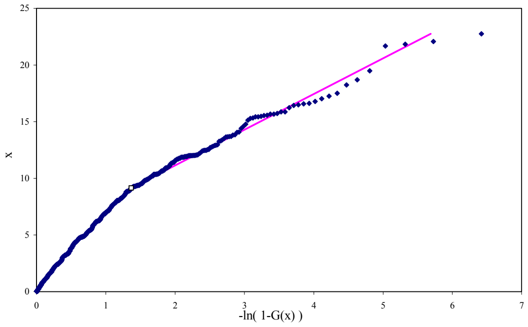

Figure 10: Exponential Q-Q plot (South Awach)

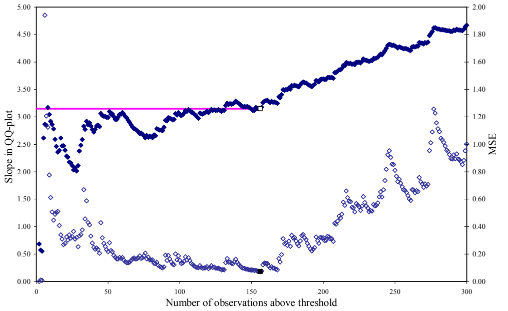

Figure 11: Hill-type estimation slope in exponential Q-Q plot (South Awach)

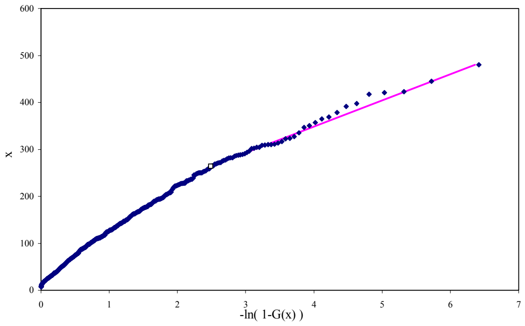

Figure 12: Exponential Q-Q plot (Nzoia)

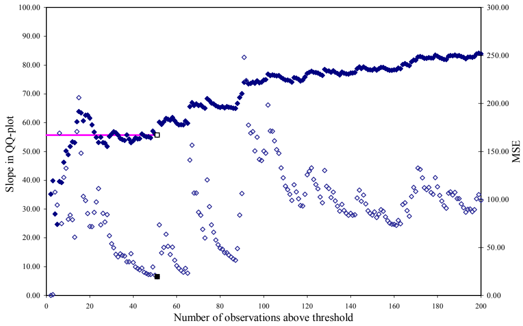

Figure 13: Hill-type estimation slope in exponential Q-Q plot (Nzoia)

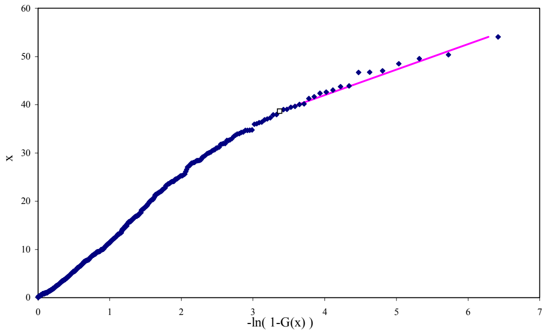

Figure 14: Exponential Q-Q plot (Sio)

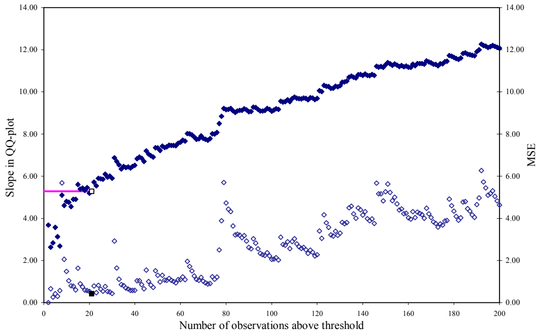

Figure 15: Hill-type estimation slope in exponential Q-Q plot (Sio)

5. DISCUSSION

To get a precise result, applicable to the hydraulic engineering practice, it is strongly recommended to use daily data (as applied for data from the catchment of Victoria Lake in Tanzania; which includes Mara, Mwanza and Kagera Regions). This is because of the peak flow that will determine the flood. The monthly values, only available for the Kenyan stations, are only indirect measures. Monthly averages should not necessarily follow the extreme value theory, because this theory is only valid for maxima of a large number of variables; monthly averages can be considered as 'volumes' and these can also follow distributions other than the Generalized Pareto Distribution (GPD). So, the fact that the data shown in Figures 8 to 15 fit exponential distributions well should not be considered as unconditional. In some cases, there was a clear influence of flooding (sudden bending down of the distribution) (like North Awach above 8.6, South Awach above 9.2). For Nzoia and Sio, it might be doubtfull whether the extreme value index is not negative (upper limit keeps bending down; more and more flooding influence). For this, it would be important to have a discussion with the focal persons or to contact local water engineers in each river catchment to receive a better understanding of the underlying physics. For stations with a clear flooding influence above a given level, a distinction can be made between the flood frequency distribution of the river discharges (flooding influence considered) and the distribution of the upstream runoff discharges (based on the lower discharge points which are not subject to flooding). The difference between these two types of distributions is given on Figures 16 and 17 for the North Awach station.

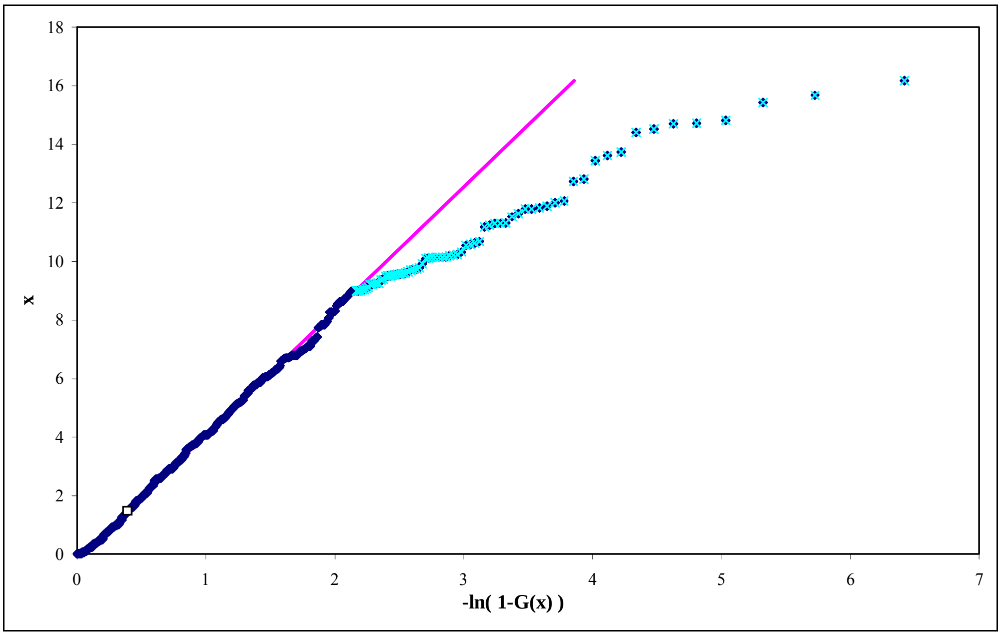

Figure 16: Flood frequency distribution of the upstream runoff discharges (based on the lower points that are not influenced by flooding conditions)

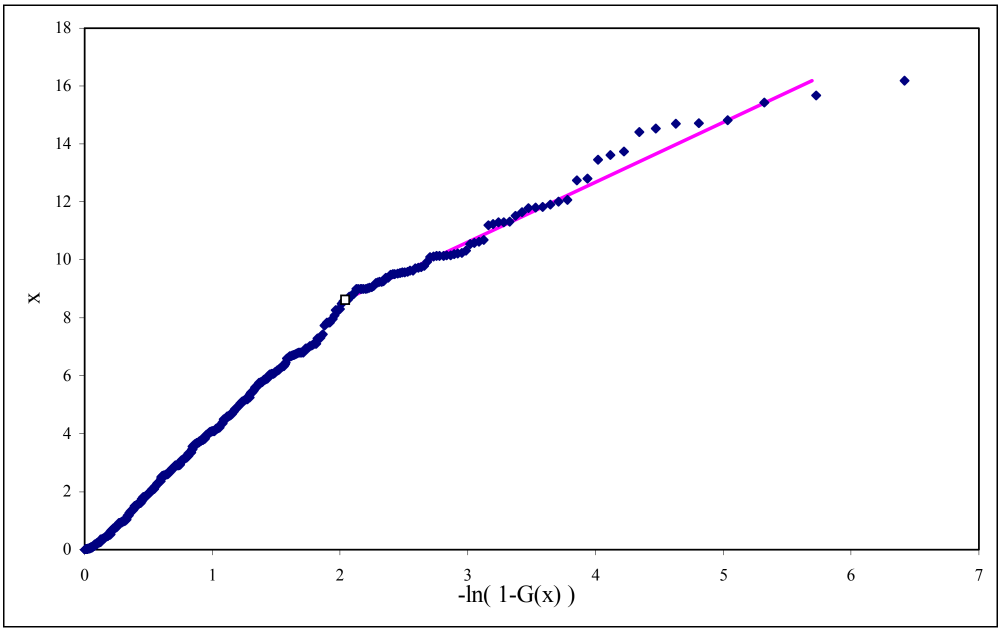

Figure 17: Flood frequency distribution of the river discharges influenced by flooding conditions

A methodology that has been developed recently by Beirlant et al. (1996) allows for an easier discrimination between competing distribution models. The methodology is based on quantile-quantile plots (Q-Q plots) and studies the tail behavior of the distribution. By analyzing the behavior of the data points in the different types of Q-Q plots, the asymptotic behavior of the tail of the distribution can be determined and a discrimination can be made between Pareto-, exponential and bounded-type distributions. The discrimination is mainly based on an estimation of the extreme value index as the slope of the linear path in a so-called UH-plot. This slope is estimated by a weighted linear regression. Also the optimal threshold level can be derived easily as the threshold level minimizes the mean-squared-error of the regression. The methodology is of special interest because of the visual nature and it can be used in combination with the traditional MOM and ML methods. In that way, the advantages of both methods are combined and an efficient extreme value analysis can be made.

6. CONCLUSIONS AND RECOMMENDATIONS

An extreme value analysis methodology was used to recognize the anomalies in tail behavior of the flood frequency distribution in an easy and visual way by means of the so-called Q-Q plots. The type of distribution and the optimal threshold level can be derived with this method in a most efficient way. Q-Q plots are very useful to study the tail behaviour of empirical distributions. By looking for the Q-Q plot, in which the largest observations behave in an asymptotic linear way, the type of the distribution of the studied extremes could be derived. By performing weighted linear regressions in this plot, the optimal threshold can also be derived. The methodology was applied to the data from Tanzania (catchment of Victoria Lake in Tanzania; which includes Mara, Mwanza and Kagera Regions, daily discharge values from January, 1954 to December, 1986) for testing and first application. Based on the expertise built up for the Tanzanian data, the extreme value analysis was applied using Kenyan discharge data. The methodology is of special interest because of the visual nature and it can be used afterwards in combination with the traditional MOM and ML methods. In that way, the advantages of both methods could be combined and an efficient extreme value analysis could be made. To get a precise result, applicable to the hydraulic engineering practice, it is strongly recommended to use daily data (as for Tanzanian data). This is because of the peak flow that determines the flood. The monthly values, as only available for the Kenyan stations, are only indirect measures. Monthly averages should not necessarily follow the extreme value theory, because this theory is only valid for maxima of a large number of variables; monthly averages can be considered as 'volumes' and these can also follow distributions other than the GPD. By analyzing the behavior of the data points in different types of Q-Q plots, the asymptotic behavior of the tail of the distribution can be determined and a discrimination can be made between Pareto-, exponential and bounded-type distributions. The discrimination is mainly based on an estimation of the extreme value index as the slope of the linear path in a so-called UH-plot. This slope is estimated by a weighted linear regression. Also the optimal threshold level can be derived easily as the threshold level minimizes the mean-squared-error of the regression.

7. ACKNOWLEDGEMENTS

The author wishes to express his great thanks to Prof. Patrick Willems from Catholic University of Leuven, Belgium for his valuable discussions and support.

8. REFERENCES

1. Beirlant, J., Teugels, J.L. and Vynckier, P., 1996, *Practical analysis of extreme values*, Leuven University Press, Leuven, Belgium.
2. Caers, J., 1996, *Statistical and geostatistical valuation of diamond deposits,* PhD thesis, Faculty of Engineering, K.U.Leuven, Belgium.
3. El-Sadek, A., 2001, *Engineering approach to water quantity and quality modelling at field and catchment scale,* PhD thesis, Faculty of Engineering, K.U.Leuven, Belgium, 251pp.
4. Hosking, J.R.M. and Wallis, J.R., 1987, *Parameter and quantile estimation for the generalized Pareto distribution,* Technometrics, vol. 29 (3): 339-349.
5. National Institute of Standards and Technology International SEMATECH, 2004, *Engineering Statistics Handbook, An online tool to help scientists and engineers incorporate statistical methods in their work,* www.itl.nist.gov/div898/handbook.
6. Shaffer, M.J., 1995, *Fate and transport of nitrogen, what models can and cannot do,* Working paper no. 11, USDA, Agricultural Research Service, Great Plains Systems Research Unit, Fort Collins, Colorado, USA.
7. Willems, P., 1998, '*Hydrological applications of extreme value analysis', In: Hydrology in a changing environment*, H. Wheater and C. Kirby (Eds), John Wiley & Sons, Chichester, vol. III, pp. 15-25; (ISBN 0-471-98680-6).
8. Willems, P., 2000, *Probabilistic immission modelling of receiving surface waters,* PhD thesis, Faculty of Engineering, K.U.Leuven, Belgium.

AUTHOR BIOGRAPHY

Dr. Alaa El-Sadek is Associate Professor at the National Water Research Center, Ministry of Water Resources and Irrigation, Egypt. He received a PhD and MSc with Distinction in Civil Engineering from Catholic University of Leuven, Belgium. Dr. El-Sadek was awarded the Encouragement State Prize in Engineering from Academy of Scientific Research and Technology, Egypt in 2004. He also got the UNESCO-MAB young scientists award in 2003. The research focused on hydrology, water quality, desalination, virtual water and water resources project management.