Modelling Risk on Losses due to Water Spillage for Hydro Power Generation.

A Verster and DJ de Waal

Department Mathematical Statistics and Actuarial Science

University of the Free State

Bloemfontein

ABSTRACT

Generation of Hydro Power, at two of the major reservoirs in South Africa, experiences major losses due to spillage. Spillage due to unpredicted heavy rainfall in the catchment area occurs when the water levels in the reservoirs are high and the management of the water through the turbines are not efficient enough to prevent spillage. In this paper the annual losses that occurred at one of the reservoirs, taking into account the years without losses, were modelled to be able to predict future losses given that the management of the water stays the same. Due to the occurrence of extreme inflows, an extreme value distribution with inflated zeros was fitted to calculate the risks.

KEYWORDS: GBG; water spillage; losses; inflated zeros; extreme values; predictive distribution

1. INTRODUCTION

The Gariep Dam is the largest reservoir in South Africa and lies in the upper Orange River. At full supply it stores 5943 million cubic meters of water. ESKOM, the main supplier of electricity in South Africa, has a hydro power station at the dam wall consisting of four turbines, each turbine can let through 162 cubic meters per second. If all 4 turbines are operating, the total release of water through the turbines is 648 m^3/s. Spillage over the wall will occur if the dam is 100% full with all 4 turbines running and the inflow into the dam exceeds 648 m^3/s. Part of managing the water is to make sure that there is storage capacity, especially in the rain season, and to lower the level of the water in the dam through power generation. There are however restrictions for ESKOM on generating too much power, for example: The Department of Water Affairs and Forestry set up a control curve above which ESKOM can use as much water for power generation as they like, but if the water level reaches the curve they are only allowed to let water through that is needed for irrigation purposes downstream.

The total loss observed at Gariep due to spillage during 1971 to 2006 is 1.7693×10^{10} million cubic meters and in terms of South African Rand it was calculated as R76, 950, 708. This is a major loss and the question on the risk ESKOM takes in the next 100 years, under the same rules, is a valid question. It is however important to note that out of the 36 years, 23 appeared without losses. Figure 1 shows the spillage during this period.

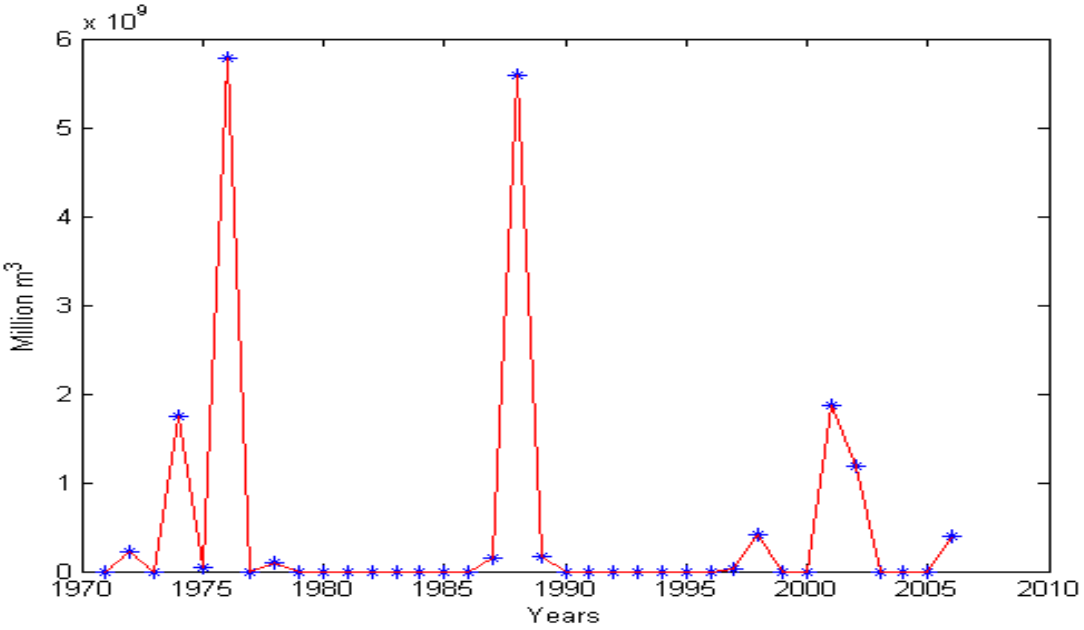

Figure 1: Spillage at Gariep Dam during 1971-2006

It is clear that two data points seem to be quite extreme; therefore we need to look into a Pareto type of distribution to fit these losses. Since we have all the data we will not consider Peaks Over Threshold models, such as the Generalized Pareto (See Beirlant *et al*, 2004). A first choice among the various distributions is the Generalized Burr-Gamma (GBG) distribution (Berlant *et al*, 2002). This distribution is fairly flexible since it has 4 parameters. The way we will proceed is to fit the GBG taking into account the inflated zeros and estimate future tail quantiles and probabilities. We would like to follow the route of the predictive density through the Bayesian approach, but due to heavy numerical integration with respect to the four parameters, we rather use the plug in method as also discussed in Beirlant's book (See for example Beirlant *et al*, 2004, page 156). For interest sake we did follow the predictive route under the Exponential distribution fit (See Beirlant *et al*, 2004, pages 438-442) for purposes of comparison knowing that the Exponential distribution will not be appropriate for this data.

2. PREDICTIVE DISTRIBUTION UNDER AN EXPONENTIAL FIT WITH INFLATED ZEROS

Let X ~ EXP(λ) with density f(x) = (1/λ)exp(-x/λ), x > 0, then x = 0 is not possible. To accommodate zeros, referred to as inflated zeros, the model is defined as

$$f(x) = \begin{cases} \theta & x = 0 \\ (1-\theta)\exp(-x/\lambda), & x > 0. \end{cases} \quad (1)$$

Let $x_1, x_2, .., x_n$ denote a random sample of n observations from this distribution and let $\bar{x}_r$ denotes the mean of the r positive observations. Assume that the joint prior for θ and λ is given by $\pi(\theta,\lambda) \propto \frac{1}{\lambda}$. The joint posterior becomes the product of a Beta(θ | n-r+1,r+1) and an inverse gamma, IG($\lambda \mid r, \frac{1}{r\bar{x}_r}$). The predictive density of a future Z is given by

$$\text{pred}(z|data) = \begin{cases} E(\theta|\text{data}), & z = 0 \\ E\big((1-\theta)\big|\text{data}\big) \cdot E\left(\frac{1}{\lambda}e^{-\frac{z}{\lambda}}\right), & z > 0 \end{cases} \quad (2)$$

$$= \begin{cases} \frac{n-r+1}{n+2}, & z = 0 \\ \frac{r+1}{n+2} \cdot \frac{1}{x_r}\left(1+\frac{z}{r\bar{x}_r}\right)^{-r-1}, & z > 0 \end{cases}$$

The last expression is the density of a GPD(1/r, $\bar{x}_r$). From this the tail quantile, at a tail probability of p, is given by

$$z = \left\{\left(\frac{n+2}{r+1}p\right)^{-1/r} - 1\right\} r\bar{x}_r \ . \quad (3)$$

Suppose we need to predict the maximum annual loss during the next 100 years. From (2) E(θ|data) = 0.6316 given n = 36 and r = 13, therefore the probability that no losses will occur in the next 100 years is 0.6316. The probability that losses will occur in the next 100 years is 0.3684. Therefore we predict 37 years ahead during which losses will occur. Let p = 37/101 (we take 100+1 in the denominator to prevent an infinite number in the calculation) in (3), then z =

5.675×10^9 million cubic meter. This is the equivalent of an annual risk of R24.68 million. If we compare this with the maximum annual spillage observed, namely 5.7889×10^9 million cubic meter (or R25.182 million), then our prediction seems to be too low. This is due to the fact that the Exponential does not give a good fit to the data. The QQ-plot between the observed and predicted quantiles (according to (3)) in figure 2 clearly shows that the two extreme values are not accommodated. We will therefore consider the fit of a more flexible distribution, like the GBG.

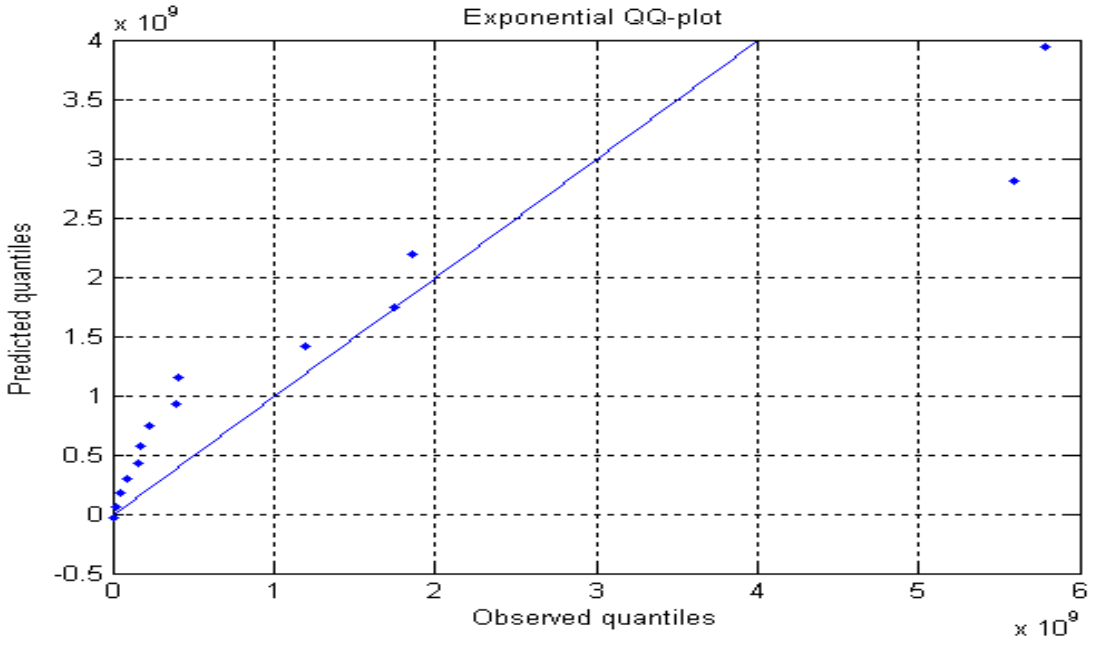

Figure 2: QQ-plot between predicted and observed qauntiles on an Exponential fit

3. THE GENERALIZED BURR-GAMMA (GBG) DISTRIBUTION

The generalized Burr-Gamma class of distributions includes many of the well known extreme value of distributions, such as the Gumbel, Weibull, Burr, Generalized Extreme Value and generalized Pareto distributions to name a few. The GBG distribution contains four parameters, k, μ, σ, ξ, where ξ is known as the extreme value index. μ is called a location parameter although it is the mean of $Y = -\log(X)$, where X is GBG distributed, only if $\xi = 0$. Similarly σ is called the standard deviation of Y only if $\xi = 0$. Since $\xi = 0$ implies that the distribution belongs to the Gumbel class with no extremes, it means that if extremes do exist and is deleted, μ and σ can be estimated from the rest of the data. The mean and standard deviation of the GBG are shown to be complex expressions of all four parameters. We will make use of the idea to delete extremes exceeding a threshold to estimate μ and σ as moment estimates and then proceed to estimate the two shape parameters k and ξ in Section 3.3. The GBG distribution models all

the data, also the data in the tail and is given as follows by Beirlant *et al*. 2002. A random variable X is GBG(k, μ, σ, ξ) distributed when the distribution function is given by (Berlant *et al*, 2002)

$$F(x) = P(X \le x) = \frac{1}{\Gamma(k)} \int_0^{\upsilon_\xi(x)} e^{-u} u^{k-1} du \tag{4}$$

where

$$\upsilon_\xi(x) = \frac{1}{\xi} \log\left(1 + \xi \nu(x)\right) > 0$$

and

$$\nu(x) = e^{\left\{\psi(k) + \frac{\log x + \mu}{\sigma}\sqrt{\psi'(k)}\right\}}$$

for

$$1 + \xi v(x) > 1.$$

$\psi(k) = \frac{\partial}{\partial k} \log \Gamma(k)$ and $\psi'(k) = \frac{\partial}{\partial k}\psi(k)$ represent the digamma and trigamma functions respectively.

The parameter space is defined as $\Omega = \{-\infty < \mu < \infty, \sigma > 0, k > 0, -\infty < \xi < \infty\}$.

It is shown by Beirlant *et al*. 2002, p. 115 that $V_\xi \sim \mathsf{GAM}(k,1)$. They also show that for $\xi = 0$, $V \sim GAM(k, 1)$ and X is generalized Gamma distributed with distribution function

$$F(x) = \frac{1}{\Gamma(k)} \int_0^{\upsilon(x)} e^{-u} u^{k-1} du. \tag{5}$$

3.1 PROPERTIES OF THE GBG DISTRIBUTION

For $V_\xi \sim GAM(k, 1)$ and $Y = -logX$, the approximated expected value and variance of Y can be derived by using the delta method (Rice, 1995) as

$$E(Y) \approx \mu - \frac{\sigma}{\sqrt{\psi'(k)}}\left[\left[log\left(\frac{\exp(\xi k)-1}{\xi}\right) + \frac{1}{2}k\left(\frac{-\xi^2\exp(\xi k)}{(\exp(\xi k)-1)^2}\right)\right] - \psi(k)\right] = \mu_Y \quad (6)$$

and

$$Var(Y) \approx \frac{\frac{\sigma^2}{\psi'(k)}k\xi^2\exp(2\xi k)}{(exp(\xi k)-1)^2} = \sigma_Y^2 \quad (7)$$

(Beirlant *et al*. 2002).

For $X = \exp(-Y)$ where $g(Y) = \exp(-Y)$ the delta method can again be applied and the approximations for the expected value and variance of X can be derived as follows:

$$E(X) \approx g(\mu_Y) + \frac{1}{2}\sigma_Y^2 g''(\mu_Y)$$

$$\approx exp\left[\frac{\sigma}{\sqrt{\psi'(k)}}\left[\left[log\left(\frac{\exp(\xi k)-1}{\xi}\right) + \frac{1}{2}k\left(\frac{-\xi^2\exp(\xi k)}{(\exp(\xi k)-1)^2}\right)\right] - \psi(k)\right] - \mu\right] \quad (8)$$

and

$$Var(X) \approx \sigma_Y^2[g'(\mu_Y)]^2$$

$$\approx \frac{\frac{\sigma^2}{\psi'(k)}k\xi^2\exp(2\xi k)}{(\exp(\xi k)-1)^2}\left[-exp\left[\frac{\sigma}{\sqrt{\psi'(k)}}\left[\left[log\left(\frac{\exp(\xi k)-1}{\xi}\right) + \frac{1}{2}k\left(\frac{-\xi^2\exp(\xi k)}{(\exp(\xi k)-1)^2}\right)\right] - \psi(k)\right] - \mu\right]\right]^2 \quad (9)$$

3.2 INVESTIGATING THE APPROXIMATIONS

To investigate the appropriateness of the approximations for $E(X)$ and $E(Y)$ in (6) and (7), data sets were simulated from a GBG distribution. After simulating a data set the approximated mean and variance are compared to the true mean and variance of Y. This is illustrated in the

following figures. The true mean and variance is indicated by the solid line and the approximated mean and variance is indicated by '*'. For the simulations $\mu = 0$ and $\sigma = 1$ are assumed to be fixed. Figure 3 indicates that for $3 \leq k \leq 6$ and $0 \leq \xi \leq 3$ the approximated mean are close to the true mean of Y.

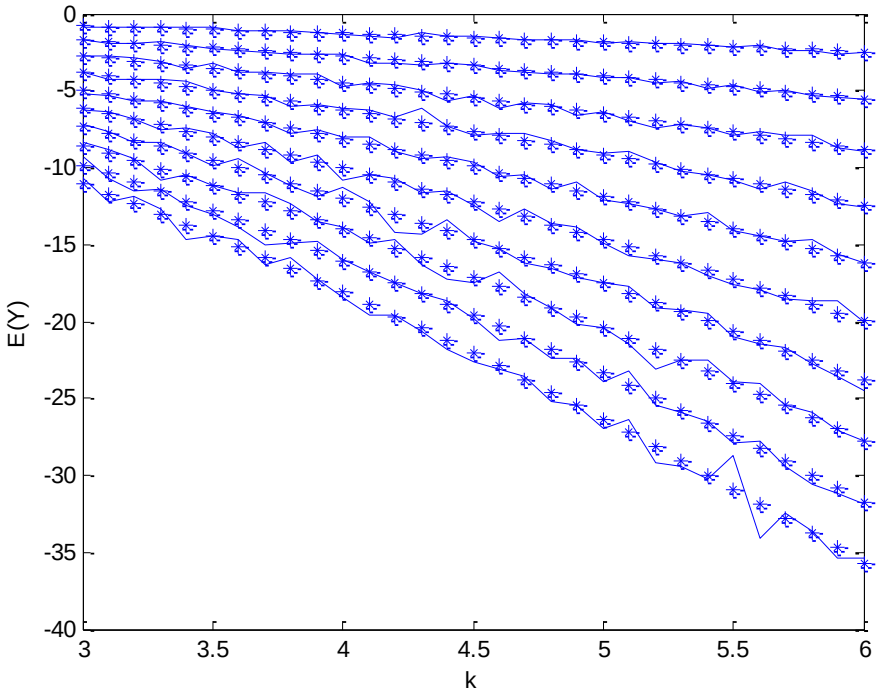

Figure 3 True mean of Y (-) vs. the approximated mean of Y (*) where the values of ξ range from 0.3 (first graph at the top) to 3 (last graph at the bottom)

Figure 4 indicates that for $3 \leq k \leq 6$ and $0 \leq \xi \leq 3$ the approximated variance are close to the true variance of Y.

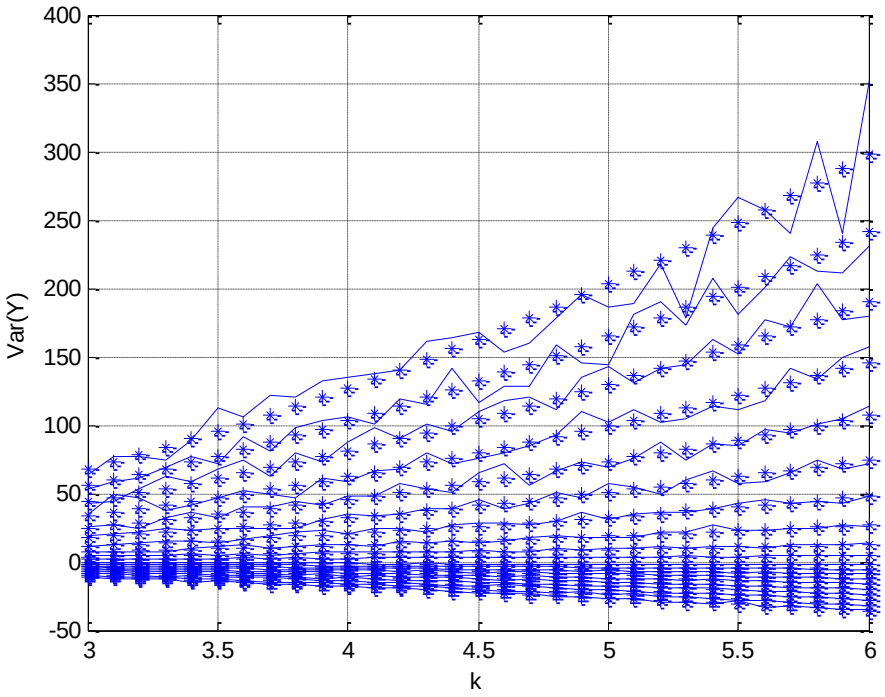

Figure 4 True variance of Y (-) vs. the approximated variance of Y (*) where the value of ξ range from 0.3 (last graph at the bottom) to 3 (first graph at the top)

3.3 ESTIMATION OF THE GBG PARAMETERS

This section discusses the estimation of the four GBG parameters. First it is shown through simulation studies that the parameters μ and σ can be estimated fairly accurately through the method of moments when only considering the data below a threshold. The method of moments are given by the following equations

$$\hat{\mu} = \sum_{j=1}^{n_t} y_j / n_t \tag{10}$$

and

$$\hat{\sigma} = \sum_{j=1}^{n_t} \left(y_j - \mu \right)^2 / n_t \tag{11}$$

where $y_j, j = 1, \ldots, n_t$ denotes the n_t observed y values below the threshold t.

The threshold is chosen by making use of the Generalized Pareto quantile plot. The Generalized Pareto quantile plot is shown in Beirlant *et al*. 2004. By making use of this method a threshold is chosen where the QQ-plot starts to follow a straight line. This is illustrated in the following simulation.

$n = 500$ values $x_1, \dots, x_n$ were simulated from a GBG with the following set of parameters, $[\mu = 0, \sigma = 1, k = 3, \xi = 0.2]$. Figure 5 shows the simulated data values of X. Let $y_j = -log x_j, j = 1, \dots, n$, the mean and variance of Y is calculated as $\mu = -0.5350$ and $\sigma = 0.8545$ respectively.

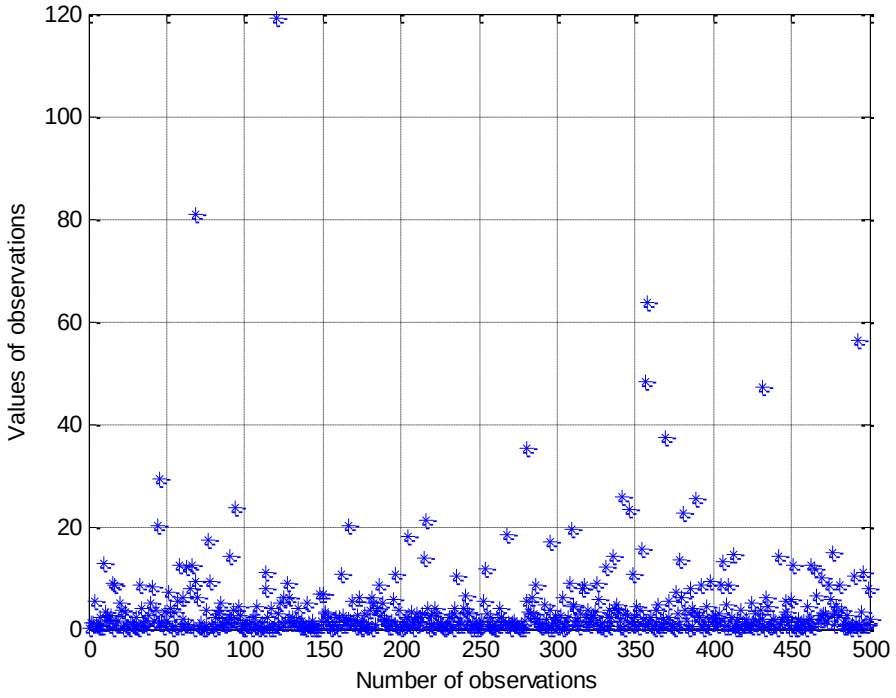

Figure 5 Simulated value of X.

From Figure 5 it is evident that extreme values occur in the data set. A threshold is now selected by using the Generalized Pareto quantile plot shown in Figure 6.

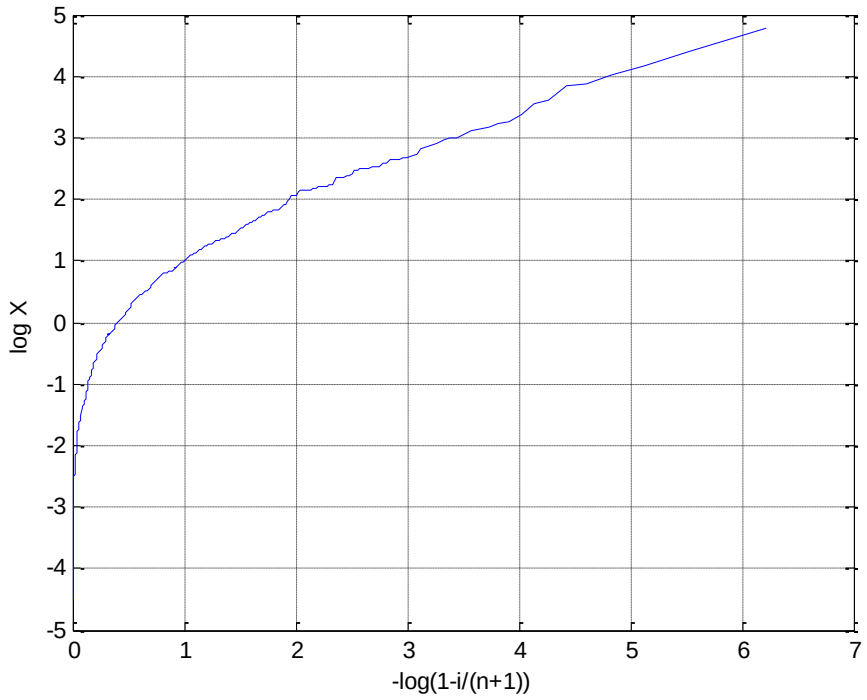

Figure 6 The Generalized Pareto QQ-plot on X, where $i = 1, \dots, n$.

From Figure 6 it seems that the quantlie plot tends to become a straight line when the log of the data exceeds 2. Therefore the threshold is $t = \exp(2) = 7.3891$. Figure 7 shows the simulated X values below the threshold. The mean and the variance of Y is now calculated for the data below the threshold as follows: $\mu = -0.1809$ and $\sigma = 1.1143$.

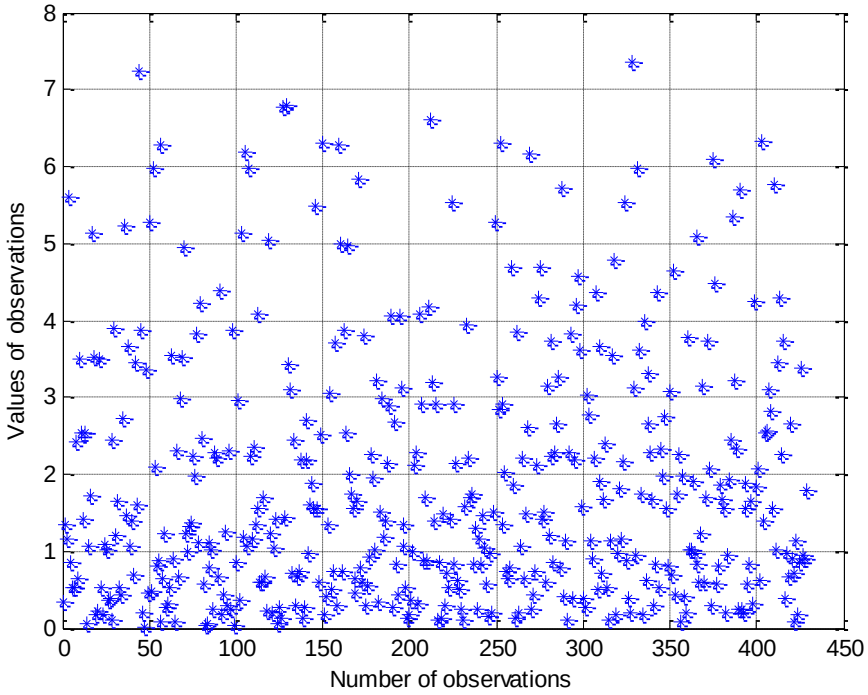

Figure 7 Simulated value of X below the threshold 7.3891

The estimates of μ and σ are closer to the true μ and σ when only the data below the threshold is considered.

The Kolmogorov Smirnov measure, $KS = \max|F_n - F|$ (Conover, 1980) is used to estimate values for k and ξ where F_n denotes the empirical cdf and F the fitted cdf. Since it is known that $V \sim GAM(k, 1)$ the Kolmogorov Smirnov measure calculates the maximum absolute difference between the empirical Gamma function and the cumulative Gamma function for different values of k and ξ. With the Kollmogorov Smirnov measure one can see how well the model fits the data. The minimum value of the different maximum Kolmogorov Smirnov measure values will indicate the best fit. This is illustrated by continuing with the simulation study. For different values of $1.8 < k < 5$ and $0 < \xi < 1.6$ the Kolmogorov Smirnov measure is calculated. The estimates of k and ξ are the values of k and ξ that gives the minimum Kolmogorov Smirnov measure value. The estimated parameter values are shown in Table 1 together with the minimum Kolmogorov Smirnov (KS) measure value. Table 1 also includes the estimated parameter values when the threshold is chosen as the 75th percentile.

Table 1 Estimated vs true parameter values.

	True parameter value	**Estimated parameter value for t = 7.3891 (Generalized Pareto quantile plot)**	**Estimated parameter value for t = 75th percentile (t = 3.9944)**
$\boldsymbol{\mu}$	0	-0.1809	0.0354
$\boldsymbol{\sigma}$	1	1.1143	1.0271
$\boldsymbol{k}$	3	2.7	3.7
$\boldsymbol{\xi}$	0.2	0.2	0.2
KS		0.0162	0.0165

From Table 1 it is clear that the estimated parameter values for both thresholds are close to the true parameter values.

Figure 8 shows the QQ-plots for the estimated parameters in Table 1. The QQ-plot gives an indication of the goodness of fit of the GBG to the simulated data. If the QQ-plot follows more or less a straight line it indicates a good fit.

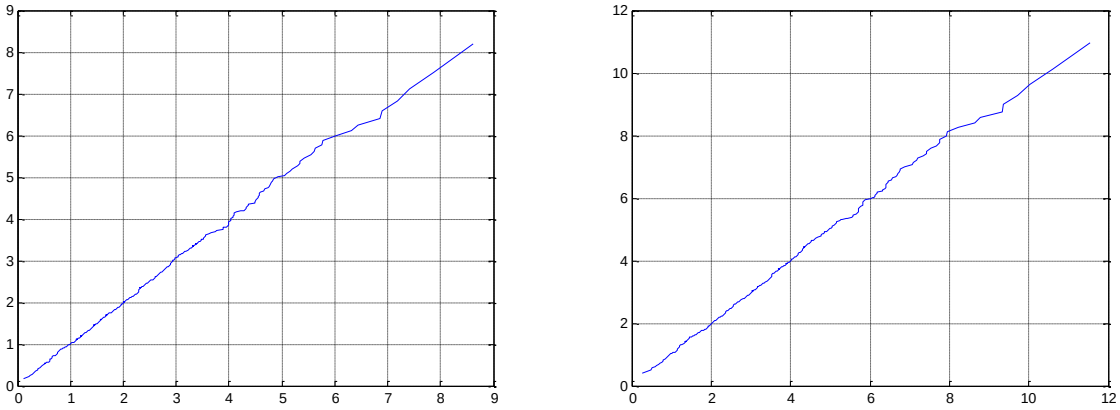

Figure 8 QQ-plots for $t = 7.3891$ and $t = 3.9944$ respectively

From Figure 8 it can be seen that the GBG is a good fit to the data for both thresholds.

4. PREDICTION OF LOSSES THROUGH THE GBG

The Generalized Pareto quantile plot is shown in Figure 9. It seems fairly obvious to delete the largest two losses to fit the GBG and to estimate μ and σ. The threshold is chosen at $t = \exp(22.445) = 5594 \times 10^6$. A GBG distribution is fitted to the data, where μ and σ are estimated from the values below the threshold, k and ξ are estimated where the Kolmogorov

Smirnov measure for goodness of fit, reaches a minimum for different values of k and ξ. The estimated parameter values are shown in Table 2 and the QQ-plot is shown in Figure 10.

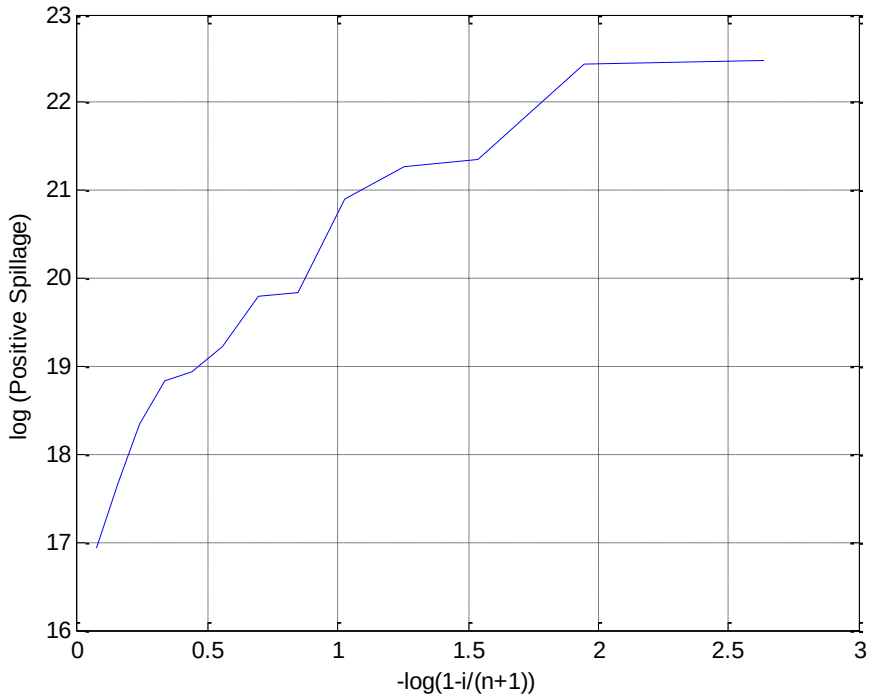

Figure 9 The Generalized Pareto QQ-plot on the positive spillage values

Table 2 Estimated parameter values

μ	σ	k	ξ
-19.3732	1.4375	5.2	0.1

Figure 10 shows a QQ-plot of the GBG fit. The QQ-plot gives an indication of the goodness of fit of the GBG to the simulated data.

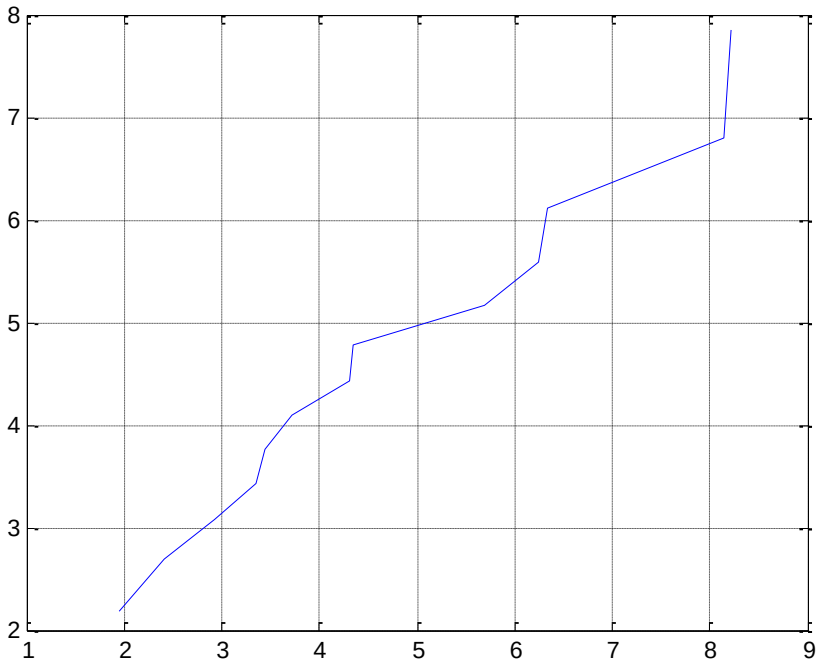

Figure 10 QQ-plot for the estimated parameter values

The value for KS is given by KS $= 0.087$ which is highly significant and confirms the fit of the GBG.

The maximum loss that can be expected during the next 100 years, can now be estimated from the tail quantile function

$$U(p) = \exp\left(-\hat{\mu} + \frac{\hat{\sigma}\psi(\hat{k})}{\sqrt{\psi'(\hat{k})}}\right)\left\{exp\left(\left(\hat{\xi}\Gamma_{\hat{k}}^{-1}(p) - 1\right)/\hat{\xi}\right)\right\}^{\frac{\hat{\sigma}}{\sqrt{\psi'(\hat{k})}}} \tag{12}$$

by letting $p = \frac{n+2}{r+1}\frac{37}{101}$ where $n = 36$ and $r = 13$. $\Gamma_k^{-1}(p)$ denotes the $(1-p)$ th quantile for a gamma$(k, 1)$ distribution. This gives an estimate of 6.3948×10^9 million m^3 loss which is equivalent to a risk of R 28.7 million. Comparing this with risks calculated in Section 2 we note a substantial increase. The estimate of θ (probability of no spillage in a year) is taken here similar to the estimation in Section 2.

5 CONCLUSION

We belief that the GBG is the appropriate model to fit to the dataset. This work includes the calculation of inflated zeros. Instead of using the predictive density of the GBG to calculate future losses, we considered the "plug in" method; this is due to numerical and analytical difficulty of calculating the predictive density. Further research can be done in this area. When compared to the Exponential distribution the estimates of the GBG seems acceptable. Unless the management of the water is changed the risk of future losses due to spillage at the Gariep Dam is large. To be able to manage the water better, a prediction of the rainfall in the catchment area will be necessary. The catchment area is mountainous and lies in Lesotho, known as the Country in the Sky, with almost none rainfall gauges. The soil moisture also plays an important role in predicting the stream flow. This forms part of a long term research plan and is under investigation. An interesting aspect of this research is the prediction of the Southern Oscillation Index (SOI) which has been discussed in the literature (see for example Salisbury & Wimbush, 2002) as a challenging problem.

ACKNOLEGDEMENT

We want to thank Mr. Robert Kydd from ESKOM for his assistance in providing the data.

REFERENCE

[1] Beirlant J, De Waal DJ & Teugels JL. 2002. The generalized Burr-Gamma family of distribution with applications in extreme value analysis. *Limit Theorems in Probability and Statistics* **I**: 113-132.

[2] Beirlant J, Goedgebeur Y, Segers J & Teugels J. 2004. Statistics of Extremes Theory and Applications; Wiley & Sons: England.

[3] Conover WJ. 1980. Practical Nonparametric Statistics. Second Edition; Weily & Sons: United States of America.

[4] Meng XL. 1997. The EM algorithm and medical studies: a historical link. *Statistical Methods in Medical Research* **6:** 3-23.

[5] Rice JA. 1995. Mathematical Statistics and Data Analysis. Second edition; Duxbury Press: California.

[6] Salisbury JI & Wimbush M. 2002. Using modern time series analysis techniques to predict ENSO events from the SOI time series. *Nonlinear Processes in Geophysics* **9**: 341-345.