Selection of Flood Frequency Model in Tanzania Using L-Moments and the Region of Influence Approach

Gerald MUHARA

Kenya meteorological department, P.O.Box 30259, Nairobi.
E:Mail: Muhara@lion.meteo.go.ke

ABSTRACT

This study aims at establishing the underlying statistical distributions for various sites, derive hydrologically homogenous regions in Tanzania and establish the regional statistical distributions, based on the L-Moments diagrams. The parameters of these distributions are estimated. Linear regression models for various regions and eventually the whole country are also established. The region of influence (ROI) method of defining hydrologically homogenous regions has been used in this study. The method gives satisfactory results but only after coupling it with a statistical tool for checking regional homogeneity. The regions derived using this method compares well with those defined earlier using the method of Principal Component Analysis (PCA). Of the Twelve regions in Tanzania 2 were described by Log Logistics (LLG), 1 by Generalized Pareto (GPA), 4 by 3 Parameter Lognormal (LN3), and 3 by Pearson type three (P3). One region did not have sufficient data for analysis. The study revealed that, instantaneous annual maximum flood in Tanzania can be estimated by P3. Regional parameters of P3 for standardized instantaneous annual maximum flow data in Tanzania are: QMEAN=1.00, QSTDV=0.58 and QSKEW=1.51 The linear regression model for the country takes the form:

Log Q(bar)=0.58 Log A + 0.88 Log R + 0.14 Log S + Log C

With R squared = 0.48 and standard error of Q(bar) estimate = 0.42 and Log C=-2.84.

Thus the model is a poor estimator of Q(bar), however regional multiple linear regression models show high efficiency.

Key words *L-moments, PWM, regionalization, homogeneity, heterogeneity, regression, quantiles*

1 Introduction

Before designing a variety of engineering works in water resources planning and other water related projects, engineers require flood estimates at a particular site of interest or project location. This includes the design of spillways, bridges, culverts, water supply system and other diversion structures. Flood volume estimates are required for the design of flood control structures, storage reservoirs and for flood plain zoning. A common problem encountered in many aspects of the water resources engineering is that of estimating the return period of rare events such as extreme floods or precipitation for a site or a group of sites (Hosking and Wallis,1993, Cunane,1985)

In flood frequency analysis, the objective is to establish a flow magnitude (Q) corresponding to any required return period (T) of occurrence. That is, a past record is fit with a statistical distribution function which is then used to make inferences about future events. Identification of the true statistical distributions for the various hydrologic and meteorological data sets (annual flood peaks and annual maximum daily rainfall) continues to be a major question facing engineers and scientists. An even greater problem facing hydrologists and meteorologists is the identification of the distribution form for regional data. The regional data question is easier to answer for precipitation owing to the grater number of rainfall stations and the continuity of the weather causing systems unlike floods.

1.1 Background

Flood frequency model selection relates to, choice of distribution and method of parameter estimation and also, use of at site data and regional data when available among other things. Gumbel (1941), Moran (1957), Bensoan (1968), did a lot of research concerning the choice of distribution. However Fiering (1967), observed that the treatment of choice of distribution and parameter estimation separately is not justified.

Due to existence of extremely large values (*outliers*), most of the statistical flood frequency methods under-estimates the occurrence of very large floods. If these outliers are omitted from the analysis, the resulting quantiles are more accurate but the sample can no longer be regarded as random and unbiased. Cunane (1989), observed that the effect of these outliers is negligible if the appropriate methods for parameter estimation are applied. Whereas the methods of conventional moments has been applied by many researchers in identification of the underlying distributions, it has one shortcoming in that it involves raising the raw data to the powers of 2,3 and 4 in order to calculate the 2nd ,3rd and 4th moments namely standard deviation, skewness and kurtosis. Thus it involves non linear transformation of data,(Pearson,1991).

Recently introduced by Hoskings are the L-Moments which are equivalent to probability weighted moments (PWM). L-Moments equates linear combinations of the sample data to the corresponding theoretical expressions involving parameters of the statistical distributions in order to specify these parameters. They avoid non-linear transformations of data. Recent hydrological literature on statistical theories for dealing with annual maximum flood series (Hosking et al 1985a, Lettenmaier and Potter, 1985, Hoskings, 1990, Hoskings and Wallis, 1990) have shown that probability weighted moments and L-Moments are often superior to standard estimation techniques particularly for regional studies.

1.2 Theory of L-Moments and their advantages

L-Moments are analogous to the conventional moments but are estimated by linear combinations of an ordered data set called L-statistics. L-Moments have a theoretical advantage over conventional moments of being able to characterize a wider range of distributions and, when estimated from a sample, of being more robust to the presence of *outliers* in the data. L-Moments and PWM are analogous to ordinary moments in that their purpose is to summarize theoretical probability distributions and observed samples. Similar to ordinary product moments L-Moments can also be used for parameter estimation, interval estimation and hypothesis testing. Although the theory and application of L-Moments parallels those of the conventional moments, L-Moments have several important advantages.

Since sample estimators of L-Moments are always *linear* combinations of the ranked observations, they are subject to less bias than ordinary product moments. Reason is that ordinary product moments estimators such as skewness and kurtosis requires squaring and cubing observations, which causes them to give greater weights to the observations far from the mean, resulting in substantial bias and variance. Hosking (1990) and Stedinger et al (1993) provide a summary of the theory and application of L-Moments. Greenwood et al (1979) also provides the summary of the PWM theory. PWM may be defined as;

$$\mathrm{B}_r = E[X[F_x(x)]^r] \qquad \mathbf{1}$$

$$a_r = E[X[1- F_x(x)]^r] \qquad \mathbf{2}$$

where $F_X(x)$ is a cumulative distribution function of X.

1.3 Biased and unbiased estimators of PWMs and L-moments

Although biased estimates of PWMs and L-Moments often produce quantile with lower root mean square error than the unbiased alternatives, the latter are preferred in goodness of fit evaluation such as L-Moments diagrams. Unbiased estimates have less bias for estimating t_3 and t_4 and are invariant if the data are multiplied by a constant, which is not the case for the biased estimators. Unbiased sample estimates of the PWM for any distribution can be computed from:

$$\boldsymbol{b}_o = m = \frac{1}{n}\sum_{j=1}^{n} x_j \qquad 3$$

$$\boldsymbol{b}_1 = \sum_{j=1}^{n-1} [\frac{(n-j)}{(n(n-1))} x_j \qquad 4$$

$$\boldsymbol{b}_2 = \sum_{j=1}^{n-2} [\frac{(n-j)(n-j-1)}{n(n-1)(n-2)}] x_j \qquad 5$$

$$\boldsymbol{b}_3 = \sum_{j=1}^{n-3} [\frac{(n-j)(n-j-1)(n-j-2)}{n(n-1)(n-2)(n-3)}] x_{(j)} \qquad 6$$

where $x_{(j)}$ represents the ordered statistics with $x_{(1)}$ the largest and so on.
In general,

$$\boldsymbol{b}_r = \frac{1}{n}\sum_{j=1}^{n-r} \frac{\begin{bmatrix}(n-j)\\ r\end{bmatrix}}{\begin{bmatrix}(n-1)\\ r\end{bmatrix}} x_{(j)} \qquad 7$$

For any distribution the first, four L-Moments are easily computed from PWM using;

$$\boldsymbol{l}_1 = \boldsymbol{b}_o \qquad 8$$

$$\boldsymbol{l}_2 = 2\,\boldsymbol{b}_1 - \boldsymbol{b}_o \qquad 9$$

$$\boldsymbol{l}_3 = 6\,\boldsymbol{b}_2 - 6\,\boldsymbol{b}_1 + \boldsymbol{b}_o \qquad 10$$

$$\boldsymbol{l}_4 = 20\,\boldsymbol{b}_3 - 30\,\boldsymbol{b}_2 + 12\,\boldsymbol{b}_1 - \boldsymbol{b}_o \qquad 11$$

for the population, while for the sample are obtained by substituting the sample moments b_1,b_2,b_3,b_4

1.4 Conventional moment and l-moment ratio diagrams

Conventional product moment ratio diagrams compare the sample estimates of the product moments cv=s/mean (coefficent of variation) and G (skewness) and K (kurtosis) with their theoretical counterparts σ/μ, γ and, for a range of assumed distributions. L-Moments diagrams compares sample estimates of the dimensionless ratios t_2, t_3, t_4 with their population counter parts, for a range of assumed distributions. An advantage of L-Moments diagrams is that one can compare fit of several distributions using a single graphical instrument (Vogel et al 1992)

Comparison of L-moment and product moment ratio diagrams shows that the latter overlaps more for a range of assumed distributions as compared to the former. This makes the L-Moments more useful in identifying the parent underlying distribution. Also t_3-t_4 diagrams are based on unbiased estimates as compared to c_s and c_k whose bias has to be collected.

1.5 Methods of quantile estimation

Estimation methods depends on data availability and on the amount of regional pooling of data which is to be allowed. Two categories of data availability are:

1) at-site hydrological data are available along with other sites in the region.

2) at-site hydrological data are not available but data at some other sites in the region are available.

If only at site data is available and a flood quantile estimate is needed, the estimation is usually done either by an AM or POT model. In the case of an AM model, it is calibrated from N-years of record length of AM flood peaks at the site of interest. The quantile Q_T is then estimated as $Q_T=a+by_T(c)$ where a,b,c are sample estimates of the location scale and shape parameter respectively, of the selected statistical distribution and $y_T(c)$ is a standardised variate value with return period T of years from the selected distribution. When both at-site and regional data are available, flood quantiles can be estimated based on regional scale. Regional statistics or information from all the sites in the region except the site of interest are pooled together(in this case regional homogeneity is assumed). Several methods of combining this information have been proposed as will be discussed later. Sites from the region should be hydrologically homogeneous.

1.6 Region of influence (ROI) approach

A region is defined as a composition of sites from which extreme flow characteristics are similar and hence when the information of extreme flows is combined, can improve the estimation of extremes at any site in the region (Zrinji and Burn, 1994). Because the gauging of every single stream is impossible, or the record length at the site of interest is short as compared to the desired design period sometimes regionalization techniques are employed to transfer the information from gauged streams to ungauged sites to improve quantile estimates of gauged sites with short records(Gingras et al, 1994).

ROI (Tasker and Slande,1994) involves estimation of unique regression for each ungauged sites. Suggested by Acreman and Wiltshire (1987, The regression equation for a site is computed using data from a unique region called the region of influence, (Burn,1990a,1990b). The unique subset of the gauging stations that comprise ROI for each ungauged sites is made up of the N nearest neighbours. The distance between two sites is measured, rather by euclidean distance but not the physical one. Two catchment characteristics are used for regionalization:

(1) Physical catchment characteristic: Include drainage area (DA), channel slope (SL), mean basin elevation (EL), mean annual precipitation (MAR) mean basin slope CL etc.

(2) Catchment hydrological characteristics:

Include: Mean Discharge (Q), Coefficient of variation (Cv) L-coefficient of variation (Lcv) etc.

The method has one advantage in that the defined region include only stations whose size, slope, and elevation etc are similar to the site of interest.

1.7 Regional homogeneity

Regardless of the way regions are formed, they have to satisfy homogeneity criteria. Each of the regions should have two basic properties, that is, dissimilarity from other regions and, homogeneity of flood characteristics within the region. The importance of regional homogeneity has been demonstrated by Hosking (1985), Wiltshire (1986) and Lettenmaier (1987). There are several tests available to examine regional homogeneity in terms of the hydrologic response of stations in a region, (Zrinji Z, 1996). In order to ensure that the resulting regions are unique internally to a given level of tolerance, homogeneity and heterogeneity test need to be performed. The test uses the L-Moment ratios to check if the sites statistics are sufficiently similar to those of the parent distribution. The $\chi^2(R)$ statistic has been employed by Chowdhury et al (1991)

1.8 Linear regression models

These are models of the form, Q(T)=F(c) where c represents catchment characteristics. The models can be of raw data or log transformed data. The form of such relation adopted for use depends on the amount of physical and climatic data availability as well as on the actual dependence between Q and the catchment characteristics. Examples are,$Q=C_1A^{0.85}$ (Benson 1962,Cole 1966),$Q=C_2A^{0.85}+S^{0.75}$ (Nash and Shaw 1966),$Q=C_3A^{0.99}S^{0.37}G^{0.98}R^{1.19}$.

2 Methodology

United Republic of Tanzania is located in East Africa, south of the equator . It lies between latitudes 0 and 12° South and longitude 29° East and 41° East and covers an approximate area of 940,000 square kilometres. It has a population of 26 million people (census of 1990). Besides a number of inland water bodies, it is surrounded by the great lakes of Eastern Africa and to the east is the Indian Ocean. To the south is lake Nyasa, to the east is lake Tanganyika, to the north is lake Victoria the second largest fresh water lake in the world. Interestingly, the highest mountain in Africa is located to the East of the country.

2.1 Selection of at site distributions

An important use of L-Moments calculated from a random sample is to identify the distribution from which the sample is drawn. It is possible to determine this because the L-Moments l_r exists if and only if E[X] exists and therefore a distribution whose mean exists is uniquely characterised by L-Moments, even if some of the conventional moments of the distribution do not exist. An appropriate distribution can be detected by plotting the sample t_3 vs t_4 values on an L-Moment ratio diagram along those of commonly used distributions.

2.2 Estimation of sample PWM and L-moments

For a given ordered random sample $x_1,x_2,..x_n$, which has been drawn from a probability distribution whose mean exists , then

$$a_r = \frac{1}{n}\sum_{i=1}^{n}\frac{\begin{bmatrix}(n-i)\\ r\end{bmatrix}}{\begin{bmatrix}(n-1)\\ r\end{bmatrix}} * x_i$$

12

r=0,1,2..n-1, and,

$$b_r = \frac{1}{n}\sum_{i=1}^{n}\frac{\begin{bmatrix}(i-1)\\ r\end{bmatrix}}{\begin{bmatrix}(n-1)\\ r\end{bmatrix}} * x_i$$

13

where by convention, k combination j = 0 if k<j

It follows that a_r is unbiased estimator of α_1 and b_r is unbiased estimator of β_1. Special case of these estimators include the sample mean and $a_0 = b_0$, and the extreme data values, $x_1=na_{n-1}$ and $x_n=nb_{n-1}$. In general a_r and b_r are linear combinations of the x_i with the weights which are polynomials of degree r in i, the first for b_r or the last for a_r, r weights being zero. The a_r and b_r are related the same way as their population counterparts a_r and b_r. It follows that,

$a_0=b_0$	$b_0=a_0$	
$a_1=b_0-b_1$	$b_1=a_0-a_1$	
$a_2=b_0-2b_1+b_2$	$b_2=a_0-2a_1+a_2$	
$a_3=b_0-3b_1+3b_2-b_3$	$b_3=a_0-3a_1+3a_2-a_3$	**13a**

Because the L-Moments l_r are linear combinations of the a_r or b_r, we can therefore construct estimators of the l_r which are the linear combinations of the a_r or b_r.

$$\begin{aligned} l_1&=a_0 && =b_0 \\ l_2&=a_0-2a_1 && =2b_1-b_0 \\ l_3&=a_0-6a_1+6a_2 && =6b_2-6b_1+b_0 \\ l_4&=a_0-12a_1+30a_2-20a_3 && =20b_3-30b_2+12b_1-b_0 \end{aligned} \qquad \textbf{13b}$$

In this case a_r and b_r are sample PWM and l_r are sample L-Moments.

2.3 Estimation of L-moments ratios

The L-Moments ratios:

$$t_r = \frac{l_r}{l_2} \qquad \textbf{14}$$

are naturally estimated by,

$$t_r = \frac{l_r}{l_2} \qquad \textbf{15}$$

based on the sample L-Moments. In this case t_3 and t_4 are the sample L-skewness and L-kurtosis. Cunane (1986a) pointed out that, although l_r is unbiased estimator of *lr*, t_r is not unbiased estimator of τ however it is its consistent estimator.

Following the above procedure the L-Moment and the corresponding ratios for each site sample were calculated. The sample ratios, t_3 and t_4 were then compared with theoretical curves, t_3 and t_4 for the various distributions (see figure 1) and the resulting samples' distributions were noted. These were then considered as the candidate distribution for the region. Sometimes the sample coordinate lies between two or more distribution curves. To make a decision for the true distribution representing sample, standard error was calculated as follows:

Take one of the possible distributions and estimate its parameters using the sample data. Generate data using the parameters already estimated so that it is has similar statistical characteristics to the sample data. Calculate the standard error and repeat the procedure with the next distribution and each time the standard error is noted. The bona fide distribution is the one with lowest standard error.

Advantages of using regional data in flood frequency analysis has been outlined in chapter two. Basically one of the advantage is that it gives more accurate quantile estimates where the record length are not long enough. Several methods of delineating a hydrologically homogeneous region have been discussed in chapter two, and there advantages and disadvantages outlined. A brief review of these methods is again explored below.

2.4 Regionalization of the study area

The region of influence method (Tasker and Slade 1994), in which a unique regression model is developed for each ungauged site, has been used successfully in Arkansas. This method was proposed by Acreman and Wiltshire, (1987). The unique subset of gauging stations that comprise the region of influence for each of the (ungauged) site is made up of the N nearest neighbours. The nearness of a neighbour is measured in terms of the Euclidean distance D_{ij}, (similarity between watershed characteristics) rather than the physical distance.

$$D_{ij} = \left(\sum_{i=1}^{n}\left(\frac{x_{ik} - x_{jk}}{S_d X_k}\right)^2\right)^{1/2} \qquad \textbf{16}$$

where D_{ij} is the distance between sites i and j in the cartesian product space of the water shed characteristics, p is the number of water shed characteristics needed to calculate d_{ij} and x_{k}, represents the kth water shed characteristics. S_dX_k is the weight attached to the kth characteristics in this case the standard deviation for the x_k and x_{ik} is the value of x_k at the ith site. In this work several combinations of water shed characteristics and sample statistical characteristics were considered. They include: areal rainfall (MAR), channel slope (SL), coefficient of variation (Cv),L-coefficient of variation (LCv), L-coefficient of skewness (LCs), L-coefficient of kurtosis (LCk).

These characteristics were used in different combinations and the resulting regions were noted. The method of region of influence is in a way similar to the traditional mapping method only that in this case the mapping is based on mathematical facts other than eyes judgement and thus not subjective. The method ensures that only those sites with similar physical (and statistical) characteristics are grouped together. The procedure for grouping the sites together is as follows;

1. choose the catchment characteristics to be used in calculating the distances
2. assume each one of the N sites is a nucleus of a region
3. by applying equation 40 calculate the euclidean distance from that site to the other N-1 sites
4. take the next site and assume it is a nucleus of another region and calculate the distances to the other N-1 sites
5. repeat that procedure for all the sites This results to N(N-1)/2 distances separating each of the station to the other.
6. sort out stations whose euclidean distance are less than a given threshold value.

In some cases, the results are unrealistic and therefore the attributes (or the combinations) are changed.

2.5 Regional homogeneity test

This was performed using the test based on the coefficient of variation of the samples coefficient of variation (CC).

The procedure is as follows:

1) Start with a few sites depicting some cluster (ie with minimum euclidean distance).
2) Compute the CC and note the value
3) Add another station and compute the updated CC
4) If CC is greater than a given value (in this case 0.3) drop that site and take another, updating the CC ensuring that it does not exceed the maximum allowable value (0.3).

2.6 Regional flood frequency analysis

This involves the joint use of at site and regional data. If only a small sample of AM data are available at site, one could not hope to estimate the entire Q-T relationship from it. In fact no more than two parameters of the AM distribution would be estimated from the sample while the form of the distribution to be fitted would have to be chosen in the light of the regional experience. Indeed of the very large variation that occurs in the Cv estimated from small samples drawn from a parent population, the estimation of the second parameter from that small sample is of doubtful validity.

In addition, if a three parameter distribution is adopted, the third parameter, skewness or a function of it, cannot be estimated from the small sample because of the very high sampling variance involved. Hence an average regional value must be adopted. Index flood method whereby a series X is derived by dividing the AM series with their respective averages, X=Q/Q(bar), has been used successfully. The quantile Q_T(est) is estimated as Q_T(est)= Q(bar)*X_T. In this work the record length weighted average of the L-statistics were employed for each of the identified regions (Vogel, 1992). To reduce to the effect of the sampling error, the stations with more than 10 years of record were used. The regional L-statistics were computed as follows (equations 17-19). The values obtained for L-skewness and L-kurtosis are the regional estimates which when plotted against theoretical L-moment diagrams reveals the underlying parent distribution. It is from these parent distributions that the regional parameters were estimated.

$$L\text{-}Cv = \frac{\sum_{i=1}^{K} L\text{-}Cv(i) * N(i)}{\sum_{i=1}^{K} N(i)} \qquad \mathbf{17}$$

$$L\text{-}Cs = \frac{\sum_{i=1}^{K} L\text{-}Cs(i) * N(i)}{\sum_{i=1}^{K} N(i)} \qquad \mathbf{18}$$

$$L\text{-}Ck = \frac{\sum_{i=1}^{K} L\text{-}Ck(i) * N(i)}{\sum_{i=1}^{K} N(i)} \qquad \mathbf{14}$$

2.7 Regional regression models

Sometimes we require to estimate the flood quantiles for some sites with whose flow data is not available. When other catchment characteristics are available for example rainfall, slope and area, we resort to linear regression models to estimate the relationship between them and mean flow. To achieve this we define the dependent and independent variables and proceed to estimate the parameters of the regression model.

A simple linear regression model is of the form;

$Q(bar)=CA^aS^sR^r$

which can be transformed logarithmically to:

$Log\ Q(bar)=Log\ C + aLog\ A+ sLog\ S+ rLog\ R.$

By defining the probability of exceedance of a flood of magnitude Qt as :

$PE=1-F(Qt)$

where F(Qt) is the probability of non-exceedance of a flood corresponding to a return period of T years, we can obtain a simple linear regression model of the form:

$y_i=u+ax_i+e_i$

by setting

$Log\ Q = y_i$,

$Log\ C= u$ and

$Log\ A = x_i$.

Thus y_i is the probability or the independent variable, **u** is the location parameter and **a** is the scale, while x_i is the dependent variable and **e** is the residual value. **u** and **a** are the parameters to be estimated. The estimated regression equation takes the form:

$y_i(est)=u+ax_i$

the residual values are calculated as:

$e_i=y_i-(u+ax_i)$.

We then minimize the sum of the squares of e_i by Least Squares Method.

2.8 Multiple linear regression and correlation

The equation is of the form,

$$y_i = b_0 + b_1 x_i + b_2 x_{i,2} + ... + b_n x_{i,n} + e \qquad \textbf{20}$$

The estimated regression equation is of the form;

$$\hat{Y} = U + XB \qquad \textbf{21}$$

and the sum of squares of the errors is given by,

$$SSE = \sum e_i^{\,2} = (Y - \hat{Y})^T (Y - \hat{Y}) \qquad \textbf{22}$$

Using these procedures linear regression models were developed for both site and regional data.

3 Results and discussion

3.1 The underlying site distributions

By comparison of the sample moments t_3 and t_4 to those of the theoretical moments,τ_3 and τ_4, the candidate distributions were found to be; Generalized Pareto,(GPA), Wakeby (5-parameter) (WA5), Three Parameter Lognormal (LN3), Pearson Type 3 (P3), Generalized Extreme Value (GEV) and Log Logistics (LLG).

Table 1. **Some gauging stations and l-moments of annual instantaneous peak flow data**

S No.	Index	Latitude	Longitude	Area	Years	Cv	LCv	LCs	LCk	MEAN
1	1B1B	4 31	38 23	200	17	.64	.37	.19	-.05	10.67
2	1B4A	4 31	38 53	7138	24	.58	.34	.04	-.06	102.46
3	1C1	5 1	38 48	650.28	84	.46	.29	.10	.11	98
4	1D14	5 1	38 48	650.31	46	.26	.15	.06	.09	114.45
5	1D16	5 11	38 32	32680	22	.42	.25	-.08	-.03	130.99
6	1D17	5 18	38 38	32918	20	.57	.32	.17	.01	176.06
7	1D18	4 10	37 32	9970	14	.43	.23	.39	.25	69.69
8	1DA1	5 9	38 34	1278	34	.35	.20	.14	.13	17.62
9	1DA3A	4 58	38 23	277	18	.55	.31	.08	.17	27.49
10	1DB2A	4 28	38 4	194	28	.40	.55	.61	.46	39.80

Although L-Moments are unbiased estimators and are less influenced by outliers, the fourth moment ratio is more sensitive to round off errors as compared to the third moment ratio. Therefore when selecting the underlying distribution from the comparison curves, we move parallel to the fourth moment ratio. The curves below shows the plot of the ratios for all the sites over the study region.

3.2 Spatial distribution of site statistical distributions

The spatial distribution of at site statistical distributions shows that they are scattered all over the study area without any cluster pattern. Thus it is not possible to form clusters using the site distributions. Worthwhile noting is that even the physical proximity of the neighbouring stations does not render them similar distributions whatsoever. This means that, especially where the record length is not long enough for the intended return period, it is not sufficient to estimate quantiles using statistical distributions obtained from site alone.

Table 2. S**ummary of the distributions and the record length weighted statistics**

No	Distribn.	Stn*Yrs	Avg Yrs	L-Cv	Cv
1	LN3	249	22.6	0.204	0.46
2	GPA	556	22.2	0.295	0.604
3	GEV	421	23.8	0.327	0.716
4	LLG	318	20.5	0.342	0.741
5	P3	301	23.2	0.351	0.759
6	WA5	318	21.2	0.394	0.830

3.3 Regional flood frequency models

By comparing the sample L-statistics with the theoretical relationship between L-kurtosis and L-skewness, most of the regions were found to be described by LN3 and P3 (Table 3). Five candidate distribution are found to be describing the 12 regions in Tanzania.

Table 3. **Regional estimates of l-cv ,l-skewness and l-kurtosis**

REGION (STN*YRS)	LCV	LCS	LCK	DISTRIBUTION
A (13)	0.157	-0.181	0.225	LLG
B (83)	0.288	-0.024	0.114	GEV
C (498)	0.299	0.208	0.145	P3
D (530)	0.257	0.137	0.126	P3
E (189)	0.462	0.353	0.324	LLG
F (102)	0.244	0.054	0.074	GPA
G (70)	0.360	0.301	0.210	LN3
H (249)	0.225	0.116	0.129	LN3
I (126)	0.272	0.113	0.137	LN3
J (95)	0.166	-0.177	0.125	LN3
K (355)	0.406	0.308	0.173	P3
TZ (2199)	0.309	0.174	0.152	P3

3.6 Estimation of parameters

Parameters were estimated using the method of PWM as it suffers less bias and variance. Lognormal distribution. PWM also eliminates the negative effects of outliers in the sample. Table 4 to 8 shows the results.

Table 4. LLG/ PWM

REGION	A	B	C
A	**	**	**
E	-0.043	0.747	0.436

** Too few stations to commit any distribution

Table 5. P3/PWM

REGION	QMEAN	QSTDV	QSKEW
C	1.000	0.53	1.28
K	1.000	0.594	1.486
TZ	1.000	0.58	1.51

Table 6 LN3/MoM

REGION	XBARL	SDL
D	-0.17	0.57
G	-0.39	0.83
H	-0.16	0.59
I	-0.11	0.51

Table 7 GPA/PWM

REGION	K	ALPHA	ZETA
B	0.867	1.547	0.172
F	0.794	1.226	0.317

3.7 Regional flood frequency curves

These curves relate the magnitude of flood (or index flood) to there respective return period. Normally EV1 reduced variate is used to ease comparison for different regions (or sites) and statistical distributions. The curves below reveal that AM flood series in Tanzania are mainly positively skewed (which is expected for extreme flow data anyway), except for two regions A and J. These two regions have very short record length besides being represented by very few stations and therefore the conclusion about their skewness is not exhaustive.

3.8 Multiple linear regression results

This was done using the catchment characteristics; area(AR), mean annual rainfall(MAF) and channel slope(SL). Region J although based on a few sites shows the highest prediction efficiency of 0.99, while the overall performance for the whole country shows very poor approximation of mean flood with the three variables. The general form of the regional regression model is;

Qm(est)= a Log AREA + p Log MAR + s Log SL + Log c

where Qm(est) is the estimated logarithm of the mean of the AM series, and

a,p and s are the coefficients of area, mean annual rainfall and channel slope respectively.

All the regions except region F shows that mean flow increases with area. For region J, mean flow is inversely proportional to mean annual rainfall (**p = -0.55).**

While slope is expected to be inversely proportional to mean flow, regions D, H, J, K and the whole country (TZ) have their slope being directly proportional to mean flow. Slope plays the lowest role in estimation of mean flow for most of the regions except I, J and K. In fact it is the dominant factor in region I. Precipitation and area play complimentary roles in most of the regions, both are the dominant factors for estimating mean flows in regions E and F.

Regression for the whole country with area only shows that estimation of mean flood with area is not possible since the standard error of estimate is higher than the model efficiency itself. The model is of the form;

Q(bar)=0.50 Log AREA + Log C

R(sq)=0.44, while standard error of estimate is as high as 0.47

For regions E and I area plays some insignificant role in estimation of mean flow since standard error of estimate of the coefficients are higher than the coefficients themselves. Similarly for region D, role of precipitation is minimum on the model while slope is not significant for regions B, E and K. Principle of parsimony requires that a compromise has to be made between model efficiency and it complexity in terms of the number of parameters incorporated. Therefore it is recommended that those variables whose parameters are of low significance be left out of the model. In this case we can attain desirable results if we use area and precipitation for region B and K, area and slope for region D and H, precipitation only for region E, and slope only for region I.

Table 9. **Multiple linear regression results based on catchment rainfall, area and channel slope**

REGION	R(sq)	a	p	s	c
B (7)	0.96	0.77	2.20	-0.31	2.85
Se		0.18	1.15	0.48	0.25
C (20)	0.75	0.56	0.94	-0.25	-2.82
Se		0.13	0.75	0.24	0.41
D (17)	0.60	0.58	0.60	0.19	-2.46
Se		0.15	0.63	0.17	0.34
E (18)	0.42	0.22	5.57	-0.21	-14.94
Se		0.46	3.41	0.62	0.46
F (5)	0.99	-3.38	17.12	-3.71	-38.07
Se		0.09	0.59	0.09	0.014
H (12)	0.91	1.03	0.46	0.42	-3.07
Se		0.14	0.57	0.21	0.21
I (8)	0.72	0.13	.16	-0.24	-5.60
Se		0.20	0.45	0.17	0.10
J (5)	0.94	1.16	-0.55	1.071	0
Se		0.31	0.33	0.96	0.26
K (9)	0.53	0.84	0.12	0.38	-1.16
Se		0.44	0.45	0.64	0.25
TZ (100)	0.48	0.59	0.88	0.14	-2.84
Se		0.08	0.30	0.11	0.42

NB 1. (*) Indicates the number of sites in the region
2. Data for regions A and G was not sufficient for this regression analysis.

4 Conclusion

In view of the analysis carried out, the following conclusions can be made;

1. P3, LN3, LLG and GPA were identified as the best fit distributions for modelling instantaneous annual maximum flows in Tanzania based on 139 station with average record length of 21.6 years.
2. At site flow data alone is not sufficient for estimation of flood quantiles mean for large design purposes since they show great variation in space.
3. The regions (C,D,K) with moderate rainfall and terrain and fall in transition zone between two and one rainfall season, are described by P3. Regions (A,E) which receives very high rainfall in one season and are located on the wind ward side near mountainous areas are represented by LLG. Regions (G,H,I,J) with one rainy season and whose soils are mainly sandy are represented by LN3.
4. The region of influence method of regionalization can be employed to delineate hydrologically homogeneous regions in Tanzania especially when coupled with a statistical tool for determining regional homogeneity.
5. Where the record length of the sample data is long enough, the method of L-moment diagrams proves to be more efficient and robust in distinguishing the underlying parent distribution.
6. The study revealed that, instantaneous annual maximum flows in Tanzania may be estimated by P3. The parameters of P3 for the whole country usinf standardised flow are; QSTDV=0.58 and QSKEW=1.51. Linear regression model for the country takes the form:

   Log Q(bar)=0.59 Log A + 0.88 Log R + 0.14 Log S + Log C

   Where Log C= -2.84 with R squared = 0.48 and standard error of Q(bar) estimate = 0.42. This model is a poor estimator of Q(bar), however regional multiple linear regression models have higher efficiency as shown above.

5 Recommendations

More research needs to be carried out in this field of regional flood frequency analysis, using smaller units of hydrologically homogenous regions in Tanzania especially using the mathematically based methods of defining a homogeneous region. More data concerning pertaining to the catchment need to be collected. This is especially, mean catchment elevation, the catchment soil type, land use and vegetation coverage as well as slope and catchment mean annual rainfall.

Some areas especially the ones with dense gauging stations network and complex topography coupled with high mean annual rainfall should be studied separately and in details. This would avoid masking of useful information from the sites with peculiar characteristics.

References

Acreman M.C and Wiltshire S.E, 1987, Identification of regions for regional flood frequency analysis. *EOS 68(44) 1262 (Abstract)*

Burn D. H, 1990b, Evaluation of regional flood frequency analysis with a region of influence approach. *Water Resources Research 26(10) 2257-2265*

Chowdhury J.U, Stendinger J.R and Lu L-H, 1991, Goodness of fit test for regional generalized extreme value flood distributions. *Water Resources Research 27(7) 1765-1776*

Cunane C, 1989 , Statistical distributions for frequency analysis

Denis Gingras, Kaz Adamswoki and Pilon P.J, 1994, Regional flood equations for the province of, Ontario and Quebec. *Water Resources Bulletin 30(1) 55-67*

Gary D. Tasker, 1982, Comparing methods of hydrological regionalization. *Water Resources Bulletin 18(6) 965-970*

Hebson C.S and Cunane C, 1987, Assessment of use of at site and regional flood data for flow flood frequency analysis. *Hydrological Frequency Modelling 443-448*

Hosking J.R M, 1994, The four parameter Kappa Distribution. *IBM J. Res Develop. Vol 38 no. 3*

Hosking J.R.M, 1986, The theory of probability weighted moments. *RC 12210(#54860) 10/6/86*

Hosking J.R.M and Wallis J.R, 1987, Parameter and quantile estimation of the generalized Pareto distribution. *Technometrics 29(3) 339-349*

Kamau D.N, 1996, Regional flood frequency analysis for South Africa catchments Including Lesotho and Swaziland. *Msc Dissertation*

Landwehr J.M , Matalas N.C And Wallis J.R, 1979 , Probability weighted moments compared with some traditional techniques in estimating parameters and quantiles. *Water Resources Research 15(5) 1055-1064*

Parida B.P and Shretha D.B, Regional flood frequency analysis of Mahi-Sabarmatti basin (subzone 3-a) using index flood method with L-moments

Pearson C.P, 1991, New Zealand flood frequency analysis using L-moments. *Journal of Hydrology 30(2) 53-65*

Ribeiro-Correa J. et al, 1995, Identification of the hydrological neighbourhoods, using canonical correlation analysis. *Journal of Hydrology 17(.) 71-89*

Shuzheng Cong, Yuazheng Li and Vogel J. L, 1993, Identification of the underlying distribution form of precipitation by using regional data. *Water Resources Research 29(4) 1103-1111*

Tasker D.G, Scott A. H and Shane C.B, 1996, Region of influence regression for estimating the 50-year flood at ungauged catchments. *Water Resources Bulletin 32(1) 163-170*

Vogel R. M et al, 1992, Flood frequency model selection in Australia. *Journal of Hydrology Vol 146 421-449*

Burn D. Hm 1990a , An appraisal of the region of influence approach to flood frequency analysis., *Hydrological Science Journal 35(24) 149-165*

Vogel R.M, Fenessey N.M, 1993, L-moments should replace product moments diagrams. *Water Resources Research 29(6) 1745-1752*

Zrinji Z. and Burn D.H, 1994, Flood frequency analysis for ungauged sites using a region of influence approach. *Journal of Hydrology* 153(1994) 1-21