International Journal of Latest Trends in Engineering and Technology
Vol.(11)Issue(1), pp.030-039
DOI: http://dx.doi.org/10.21172/1.111.06
e-ISSN:2278-621X

AN ASSESSMENT OF PREDICTIVE ACCURACY FOR REGIONAL FLOOD FREQUENCY DISTRIBUTION ESTIMATION METHODS ON AWASH RIVER BASINS

Habtamu Ketsela Mengistu[1], Kamatchi Sivakumar[2]

Abstract- The most important face of flood from water resources improvement and management part is its returning interfering with interventions and actions made by people. The loss of life and damage can be pictured in terms of economic dead and risk to human life. The main concern is here to significantly analyze the occurrence and amount of the flooding intervention. The main objectives of this study includes identifying the best-fit statistical distributions to the data of each gauge and finding a suitable parameter estimation method for each station regions of Awash river basin. The l-moment and easy fit software was employed for selection of best-fit distributions and methods of parameters estimation for a station. Goodness-of-Fit tests such as Chi-square,Anderson-Darling and Kolmogorov–Smirnov are applied for checking the satisfactoriness of fitting of probability distributions to the recorded data. Kolmogorov–Smirnov test is used for the choice of a suitable distribution for estimation of maximum flood discharge. The performance of regional General Extreme value, General Pareto and Uniform distributions are found to be highly satisfactory and widely applied in this paper, however this paper reveals that the General Extreme Value distribution is better appropriate amongst seven distributions used in the estimation of maximum flood discharge at Awash River basins.

Keywords: Best-fit statistical distributions, Easy fit, L-moment, Parameter estimation methods, Kolmogorov-Smirnov test

1. INTRODUCTION

1.1. General Background

Stream discharges and flood flows have long been considered and used by engineers in the design of hydraulic structures and flood protection works, and in planning for flood plain use [7]. A flood frequency analysis is the basis for the engineering design of water resources ranging from small to mega scale projects and the economic analysis of flood control projects. A flood frequency analysis consists of a study of past records of flow discharge and an estimate of frequencies of future floods. If adequate records are available, the common methods give acceptably uniform results within the range of data. However, the location of gauging station seldom coincides with the station of interest, or the available records become too short to make consequential statistical inference [3].

1.2. Description of the study area

1.2.1. Physiographic characteristic

i. Drainage area

It covers a total land area of 110,000 km²of which 64,000 km² is in Western Catchment of the basin. This catchment drains to the Awash main river of its tributaries. The remaining 46,000 km², most of which comprises the so-called Eastern Catchment drains into a desert area and does not contribute to the awash main river course [5].

ii. Location

The Awash Basin is situated between latitudes 7°53'N and 12°N and longitudes of 37°57'E and 43°25'E in Ethiopia. The River Awash rises at an elevation of about 3,000m in the central Ethiopian highlands, west of Addis Ababa and flows through Koka Reservoir, to north-eastwards along the Rift Valley until eventually discharging into the wilderness of the Danakil Depression at Lake Abay 250meter above sea level (msl) at the border to Djibout.Topographical map of the river basins and location of the study area are located in Fig.1

[1] Department of Hydraulic & Water resources, Wollega University, Ethiopia
[2] Department of Hydraulic & Water resources, Wollega University, Ethiopia

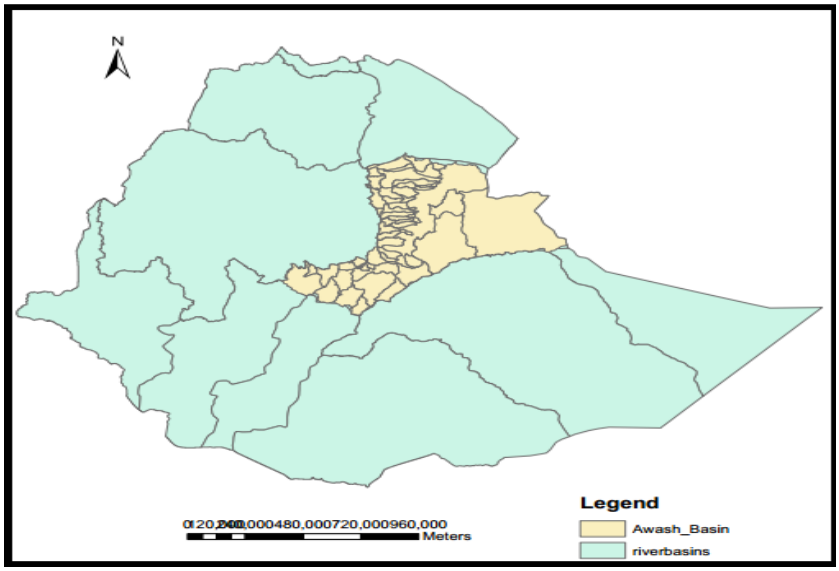


Fig.1Topographical map of the river basins and location of the study area

1.3. Objective of the research

1.3.1. General objective

The main objective of this study is assessing predictive accuracy for regional flood frequency distribution estimation methods.

1.3.2. Specific objectives

The specific objectives of this thesis are:

To identify the best-fit statistical distributions to the data of each gauges.

To develop a suitable parameter estimation method for each stations at the study area

2. SELECTION OF BEST-FIT DISTRIBUTION AND PARAMETER ESTIMATION

The choice of distribution to be used in flood frequency analysis has been a matter of importance for a long time [1]. Commonly, the chosen distribution should be widely accepted Simple and convenient to apply Consistent, flexible or strong (low sensibility to outliers).

There are many distributions that have been suggested for AM series models. Some of them are used for this particular study. When a theoretical distribution has been assumed, the validity of the assumed distribution may be verified or disproved statistically by goodness of test [2]. The results of the goodness of fit tests are used to select a distribution for frequency analysis of stations. The maximum likelihood (ML) is used for parameter estimation with the help of Easy fit software.

2.1. Kolmogorov-Smirnov Test

The test statistic in the Kolmogorov-Smirnov test is extremely simple; it is now the maximum vertical distance among the empirical cumulative distribution functions of the two samples. The empirical cumulative distribution of a sample is the proportion of the sample values that are a lesser amount or equal to a known value.

Kolmogorov-Smirnov (KS) test is a different and commonly used goodness-of-fit moreover Chi-square test. A statistic based on the deviations of the sample distribution function FN ((X) is use in this test. The test statistic DN is defined in Equation 2.1.

$$DN = \max_{1 \le i \le n} |F_n(x_i) - F_0(x_i)| \quad 2.1$$

The values of FN (x) are predictable as Nj/N where Nj is the cumulative number of sample events in class j. F0(x) is then 1/K, 2/k…..etc., Similar to the chi-square test. The value of DN must be less than a tabulated value of DN at the specified confidence level for the distribution to be received[4].

Hypothesis Testing

The Kolmogorov-Smirnov test is a hypothesis test method for formative if two samples of data are from the similar distribution. The test is non-parametric and completely nonbeliever to what this distribution really is. The truth that by no means have to know the distribution the samples come from is extremely helpful, particularly in software and operations where the distributions are durable to convey and complex to compute through.

2.2 Chi-Squared Test

Chi-Square goodness of fit test is a non-parametric test that is used to get exposed how the observed value of a particular phenomenon is considerably unlike from the estimated value. In Chi-Square goodness of fit test, the word goodness of fit is used to contrast the observed sample distribution with the estimated probability distribution. Chi-Square goodness of fit test determines how fine theoretical distribution (such as normal, binomial, or Poisson) fits the experimental distribution. In Chi-Square goodness of fit test, sample data is separated into intervals. Then the numbers of points that drop into the interval are compared, with the predictable numbers of points in every interval.

2.3. Anderson-Darling test

The Anderson-Darling test is used to test if a sample of data came from a population with a definite distribution. It is a revision of the Kolmogorov-Smirnov test and gives further influence to the tails than does the Kolmogorov-Smirnov test. The Kolmogorov-Smirnov test is distribution free in the logic that the critical values do not depend on the definite distribution being tested (note that this is right just for a completely specified distribution, i.e. the parameters are known).

The Anderson-Darling test makes utilize of the definite distribution in manipulative critical values. This has the benefit of allowing an additional perceptive test and the drawback that critical values should be intended for each distribution.

The critical values for the Anderson-Darling test are dependent on the specific distribution that is being tested.

The probability-probability (P-P) plot

P-P Plots is the variable's cumulative magnitude in opposition to the cumulative magnitude of any of a number of trial distributions. Probability plots are commonly used to determine whether the distribution of a variable matches a given distribution. If the selected variable matches the test distribution, the points come together approximately a straight line.

The probability-probability plot of analysis and observed discharge with the distribution were developed as the following steps [4].

Step 1: Order the items from smallest to largest.

Step 2: For each of the N data points, compute an empirical (observed) cumulative probability as: $F_n(x_i) = \frac{i}{n}$ for i= 1, 2... n.

Step 3: Then compute the cumulative distribution function

Step 4: Plot the graph of empirical cumulative probability and cumulative distribution function.

The quantile-quantile (Q-Q) plot

Quantile-Quantile (Q-Q) plots are plots of two quantiles against each other. A quantile is a small part where certain values fall below that quantile. The purpose of Q-Q plots is to get out if two sets of data come from the same distribution.

It is the graph of the input observed and analysis data values plotted against their theoretical (fitted) distribution quintiles. The quantile-quantile graphs are produced by plotting the data value $x_i (i = 1, \dots, n)$ against the X-axis, and the following values against the Y- axis [4].

$$F^{-1}(F_n(x_i)\frac{0.5}{n}) \qquad 2.2.$$

Where, F-1 = is the inverse cumulative distribution function

Fn = Empirical cumulative distribution function and

n = sample size

The quantile-quantile plot of analysis and observed discharge with the distribution were developed as the following steps:

Step 1: Order the items from smallest to largest.

Step 2: Draw a normal distribution curve.Divide the curve into n+ 1 segment.

Step 3: Find the z-value (cut-off point) for each segmentin Step 3. These segments are areas, so refer to z-table (or use software) to get a z-value for each segment.

Step 4: Plot your data set values (Step 1) against your normal distribution cut-off points (Step 3). The (almost) straight line on this q-q plot indicates the data is approximately normal.

3. EXPERIMENTAL MATERIALS AND METHODOLOGY

General

For this research, identifying clear and efficient methodology used is crucial for the effectiveness of the study not only from time plan point of view, but also from the quality of the research result.

Generally, the study involves the following procedure:

Collection of relevant data

Arranging andtransposing daily hydrological data in proper order

Filling of Missing data

Data Quality Assessment

Selection of the best fitted Distribution

Parameter Estimation

3.1. Source and availability of data

Awash River Basin has better hydro meteorological stations as compared to other Ethiopian's basins with regard to both density of gauging stations and record length of data. According to the data available on number of stations and density per kilometer square, Awash River Basin has better density of stations next to Rift Valley Lakes Basin and Blue Nile RiverBasin [6].

3.2 Hydrological Data

In Awash River basin there are about 72 gauging station, out of these 24 stations are used in the study area with different record length of years. The stations areas range in size from 166.4 km2 to 66308 km2. The data are

Daily stream flow,

Latitude, longitude and area of each stations, thus used to find the nearby station.

The data have been collected as soft copies, hard copies and maps, Hydrological data and digitized map of the sub-basin were collected from the Ministry of Water and Irrigation, from the department of Hydrology and GIS.

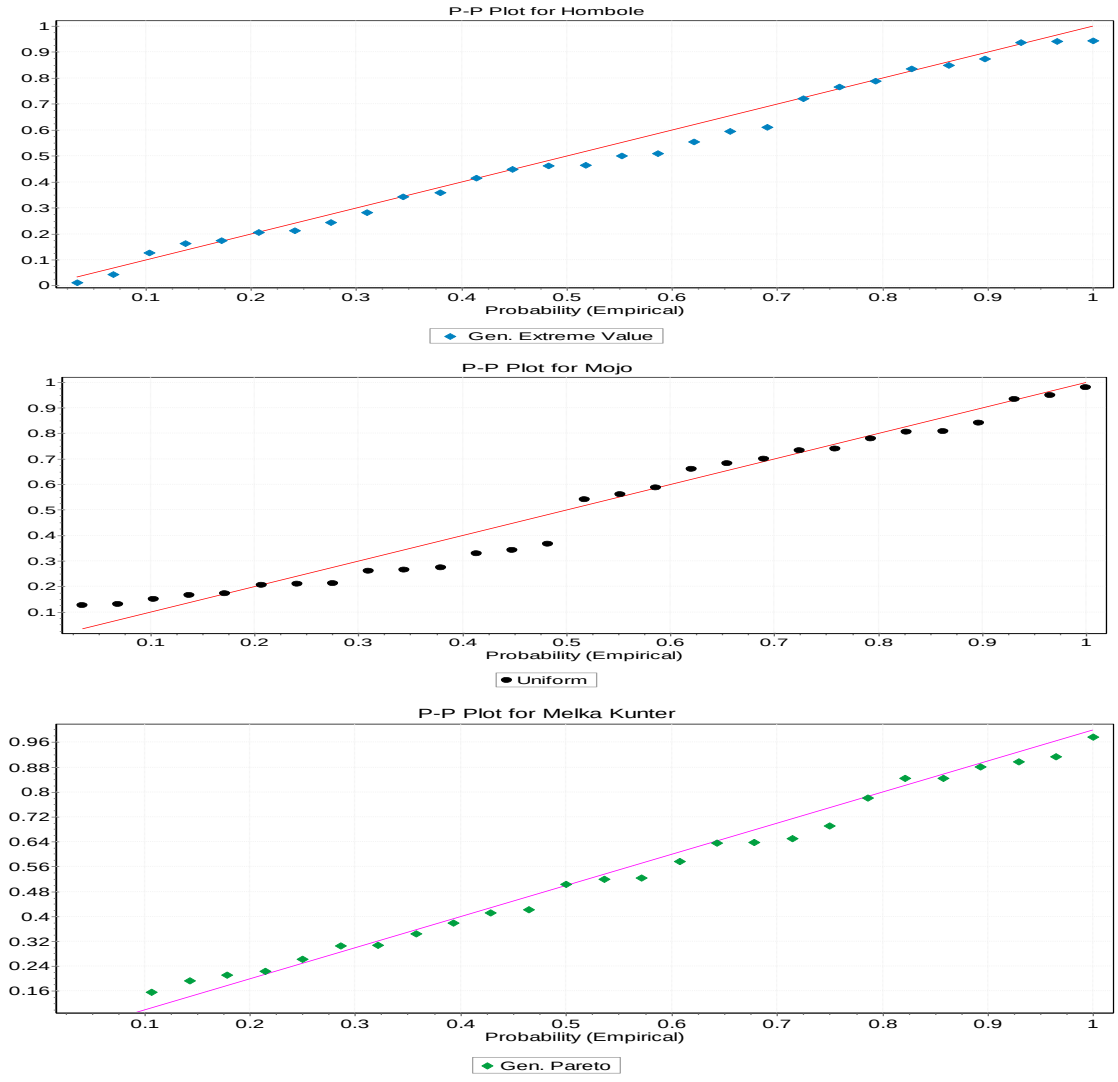
4. RESULTS AND DISCUSSION

4.1. Selections of best fitted statistical distribution and parameter estimation

The plan of this study is not to fit a particular data set but to attain quantile estimates of the distribution from which future data values will arise. When several distributions fit the data sufficiently, any of them is a sensible choice for use in the final analysis, and the best choice among them will be the distribution that is strongest. For this particular thesis MML methods are used for best fit statistical distribution and parameter estimation with help of Easy Fit Software.

The Probability-Probability (P-P) plot

The probability-probability (P-P) plot is a graph of the empirical values plotted against the theoretical values. It is used to determine how well a specific distribution fits to the observed data. The P-P plot for selected stations is shown in Fig. 4.



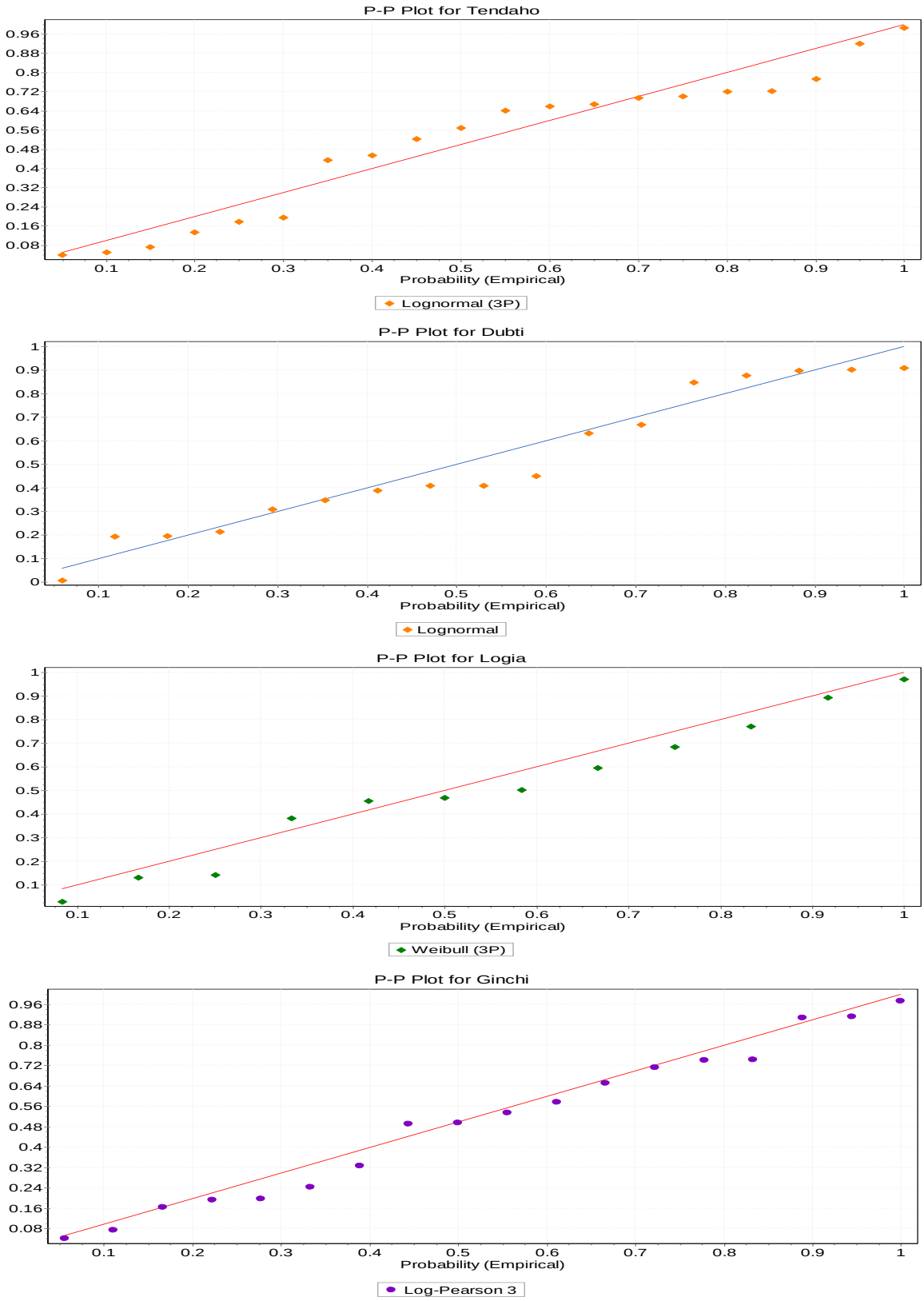
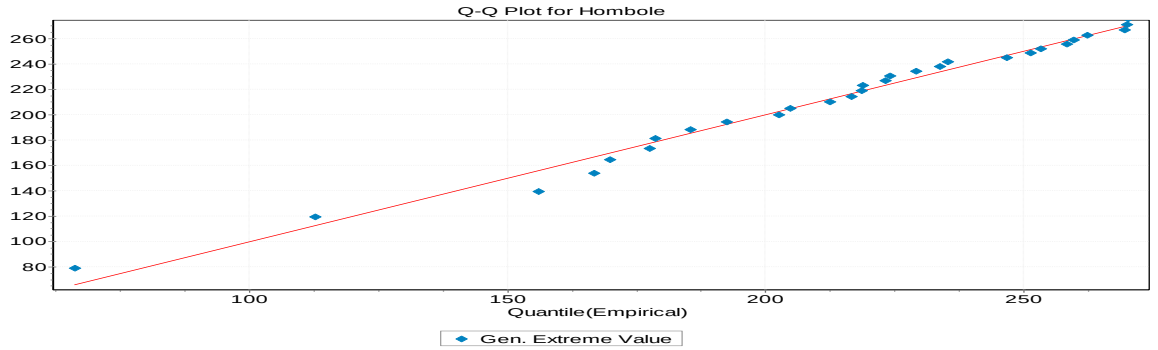
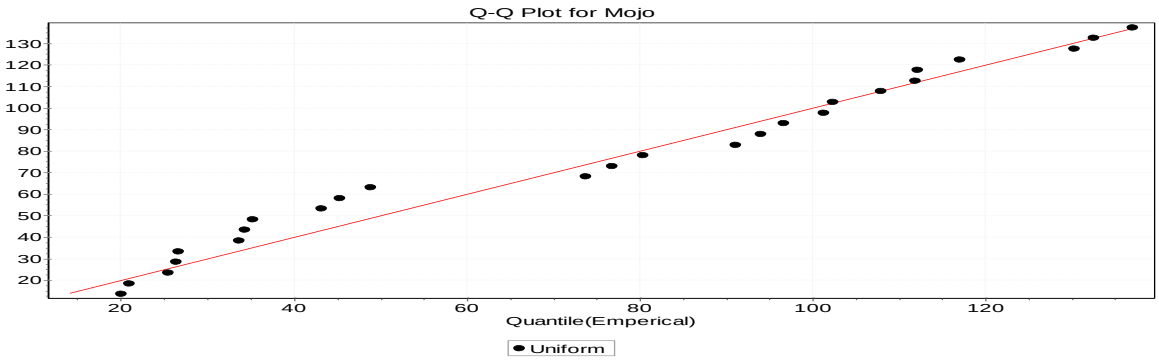
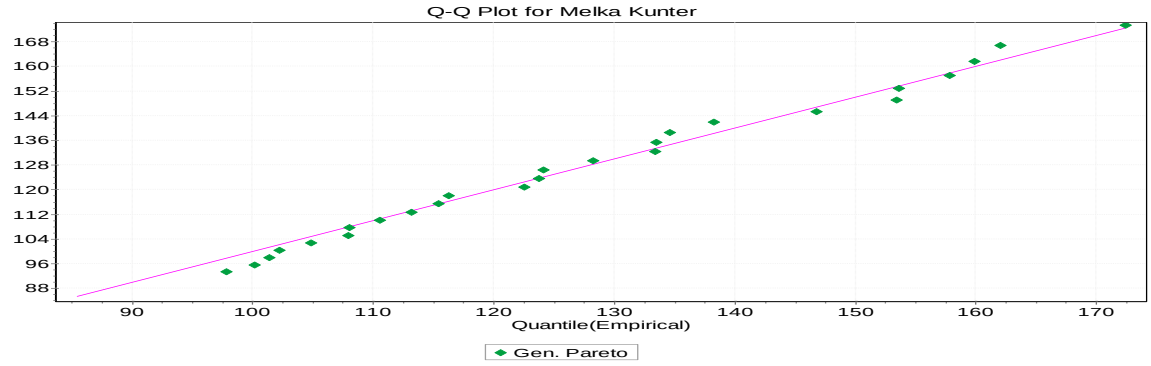
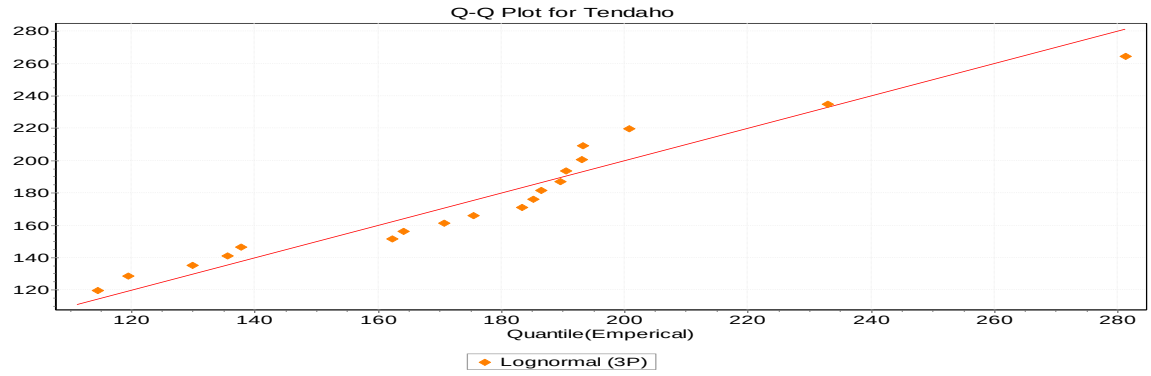


Fig. 4. 2: Q-Q Plot of selected station.


Q-Q Plot for Hombole
Quantile(Empirical)
Gen. Extreme Value


Q-Q Plot for Mojo
Quantile(Emperical)
Uniform


Q-Q Plot for Melka Kunter
Quantile(Empirical)
Gen. Pareto


Q-Q Plot for Tendaho
Quantile(Emperical)
Lognormal (3P)

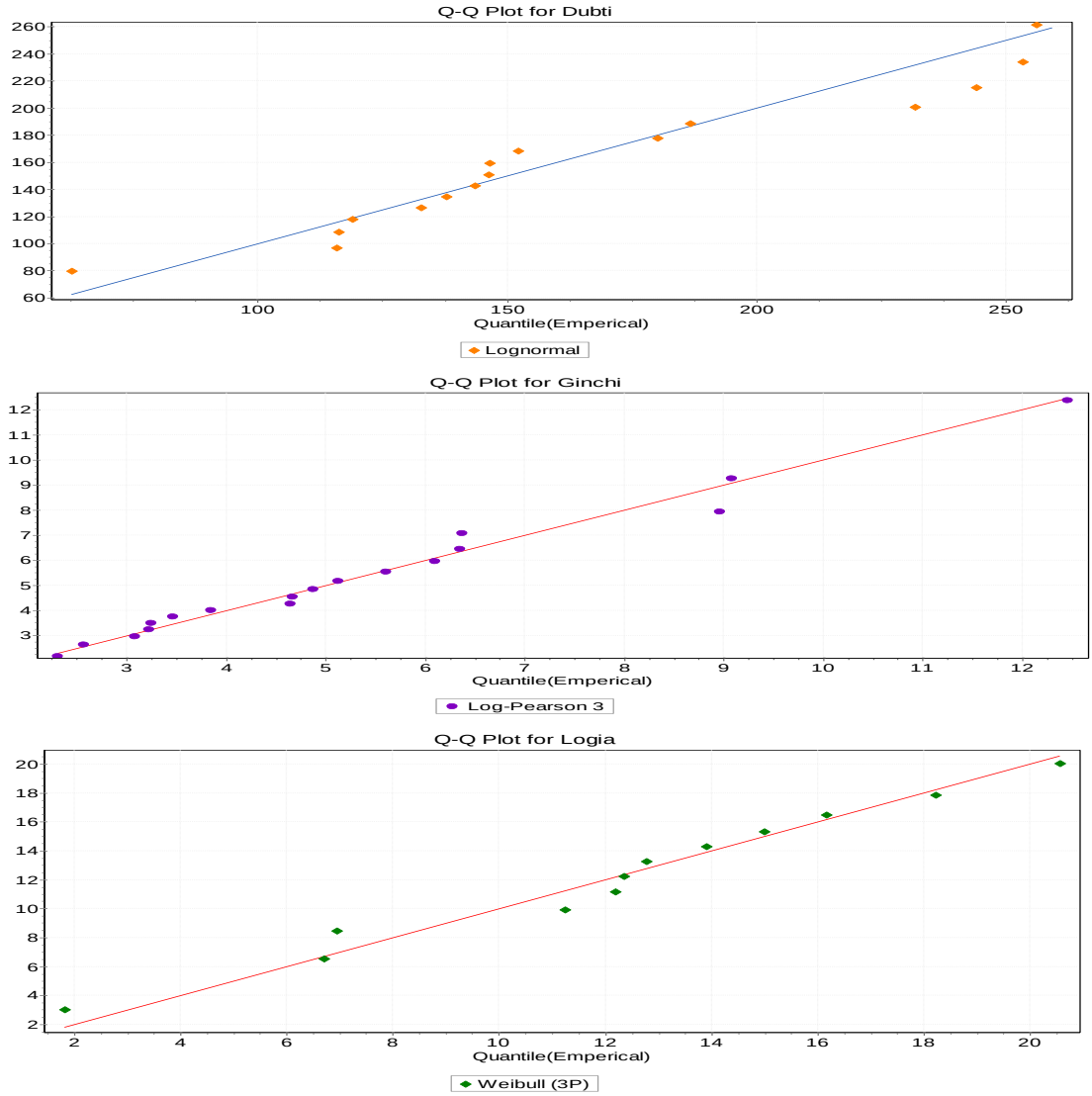


Fig 4 P-P Plot of selected station

The graphical assessment in Fig.4 is the competence of fitted distribution is generally performed by plotting the observations so that they would fall approximately on a straight line if a postulated distribution were the true distribution from which the observations were drawn. Let (Xi) denote the observed values and X (i), the ith largest value in the sample, so that X (n) $\leq$ X (n-1) $\leq$ …..$\leq$X.

The random variable Ui defined as

Ui=1-Fx (X (i))

Consequently, to the exceedance probability related with the ith largest observation. If the original observations were independent it repeated sampling the Ui have a beta distribution with mean and variance knowing the distribution of the exceedance probabilities Ui, one can develop estimators' qi of their values which can be used to plot each X(i) a against the probability scale.

The test used the correlation r between the ordered observations X (i) and the corresponding fitted quantiles Wi=G-1(1-qi) determined by plotting position qi for each X (i).

The Quantile-Quantile (Q-Q) plot

It is a graph of the input (observed) data values plotted against the theoretical (fitted) distribution quintiles. Both axes of this graph are in units of the input data set. The Q-Q plot for selected stations is shown in Fig. 4.2.

Name of stations	Selected distributions	Parameters (MML)	Value	Name of stations	Selected distributions	Parameters (MML)	Value

Hombole	GEV	K	-0.712	Mojo	Uniform	a	-3.802
		□	52.806			b	139.83
		μ	206.10				
Wonji	Gen.Pareto	K	-0.512	Teji	Uniform	a	4.202
		□	58.148			b	22.758
		μ	35.862				
MelkaKunter	Gen.Pareto	K	-0.682	Metahara	GEV	K	-0.003
		□	62.446			□	20.45
		μ	87.682			μ	60.073
MelkaWerer	GEV	K	-0.463	Keleta	GEV	K	-0.14
		□	49.093			□	4.332
		μ	169.0			μ	10.0
Below Koka Dam	Gen.Pareto	K	-0.273	Awash 7 Kilo	Uniform	a	54.608
		□	49.384			b	225.66
		μ	34.706				
Tendaho	Lognormal(3P)	□	0.219	Mille	Uniform	a	8.075
		□	5.192			b	61.0
		□	-11.362				
Dubti	Lognormal	□	0.366	Jawoha	Uniform	a	3.165
		□	5.071			b	25.84
BolkenaKombolcha	Lognormal(3P)	□	0.78983	Jara	GEV	K	-0.187
		□	1.5002			□	2.895
		□	1.8246			μ	6.077
Robi	Gen.Pareto	K	-0.57282	Ataye	GEV	K	-0.545
		□	18.834			□	2.011
		μ	-1.5953			μ	3.366
MelkaSedi	Uniform	a	61.611	Belo	GEV	K	-0.673
		b	172.71			□	9.308
						μ	37.632
Kessem (A.Melka)	GEV	K	-0.42762	Logia	Weibull(3P)	□	6.785
		□	15.699			□	30.544
		μ	55.182			□	-16.176
Akaki	GEV	K	-0.5244	Ginchi	Log-Pearson 3	□	32.262
		□	24.278			□	0.080
		μ	57.958			□	-1.021

Table 4.1 : Estimated Parameters for selected distributions

The Q-Q plot will be more or less linear if the particular theoretical distribution is the correct model. Easy Fit displays the suggestion crossways line along which the graph points must fall.
Selected distributions for analysis discharge of each station were done by using easy-fit statistical computer software and the results are shown in the following Table 4.1.

S.No.	Name of stations	Kolmogorov-Smirnov (KS)	Anderson-Darling	Chi-Square ($\chi 2$)	Selected Distribution
1	Hombole	GEV	GEV	Lognormal(3P)	GEV
2	Wonji	Gen.Pareto	Gen.Pareto	Weibull(3P)	Gen.Pareto
3	Mojo	Uniform	Uniform	Weibull	Uniform
4	MelkaKunter	Gen.Pareto	GEV	Gamma(3P)	Gen.Pareto
5	Teji	Uniform	GEV	GEV	Uniform
6	MelkaWerer	GEV	GEV	Gamma	GEV
7	Metahara	GEV	GEV	Weibull(3P)	GEV
8	Below Koka Dam	Gen.Pareto	GEV	GEV	Gen.Pareto
9	Keleta	GEV	GEV	Normal	GEV

10	Awash 7 Kilo	Uniform	Uniform	Uniform	Uniform
11	Mille	Uniform	Uniform	GEV	Uniform
12	Tendaho	Lognormal(3P)	Logistic	Weibull	Lognormal(3P)
13	Jawoha	Uniform	Uniform	Lognormal(3P)	Uniform
14	Jara	GEV	GEV	Normal	GEV
15	Dubti	Lognormal	Lognormal	Exponential	Lognormal
16	Ataye	GEV	Weibull(3P)	Lognormal	GEV
17	BolkenaKombolcha	Lognormal(3P)	Lognormal(3P)	Gen.Pareto	Lognormal(3P)
18	Robi	Gen.Pareto	Gen.Pareto	Uniform	Gen.Pareto
19	Belo	GEV	GEV	Lognormal	GEV
20	MelkaSedi	Uniform	GEV	Gamma(3P)	Uniform
21	Kessem(A.Melka)	GEV	GEV	Normal	GEV
22	Akaki	GEV	GEV	Log-Pearson 3	GEV
23	Logia	Weibull(3P)	Weibull(3P)	Weibull(3P)	Weibull(3P)
24	Ginchi	Log-Pearson 3	Log-Pearson 3	Weibull(3P)	Log-Pearson 3

Table 4.2: Selected distribution of each station

Probability plot are extremely helpful for visually instructive the character of a data set. Plots are a valuable way to see what the data look like and to determine it fitted distribution appears reliable with the data. Analytical goodness-to-fit criteria are constructive for in advance an approval for whether a particular removal of the data from model is statistically significant.
In the most case a number of distributions will provide statistically acceptable fits to the available data so that goodness-of-fit tests are incapable identify the accurate or best distribution to use. Such tests are important when they can express that some distributions appear contradictory with the data.
Some basic issues arise when selecting a distribution. One should differentiate between the following questions.
1. What is the true distribution from which the observations are drawn?
2. What distribution should be used to obtain sensibly perfect and strong estimations of design quantiles and hydrologic risk?
3. Is a planned distribution reliable with the available data for a site?
Depend on the above criteria true distribution from which the observations in both P-P and Q-Q plot are similar at Kolmogorov-Smirnov test, and also if you are compare a number of distributions, select the one that gives the largest p-value; this is the closest match to your data.
Based on the above reason the stations Tendaho,Teji, Ataye, MelkaKunter,MelkaSedi and Belo Kokadam distributions is selected based on P-P and Q-Q plot with Kolmogorov-Smirnov test rather than Chi-square and Anderson-Darling test. The advantage of the Kolmogorov-Smirnov tests over the Chi-square test it that it is not essential to divide the data into bins; hence the problems associated with the chi-square approximation for small number of intervals would not appear with the Kolmogorov-Smirnov test.

4.2. Parameters for selected distributions

Estimation by the maximum likelihood (ML) method involves the choice of parameter estimates that produce a maximum probability of occurrence of the observations. The parameter estimates that maximize the likelihood function are computed by partial differentiation with respect to each parameters and setting these partial derivatives equal to zero and finally solve the resulting set of equations simultaneously. The equations are usually complex as a result of this difficulty; the solution set may not properly found.Estimated parameters for selected distributions are calculated below with the help of easy fit software. Parameters for selected distribution is located in Table 4.2. The result from theTable 4.2 is parameters estimation of a distribution is used to determine quintile estimates which correspond to different return periods T may be computed.

5. CONCLUSION

The following conclusions are drawn from the study:
1. The study presents the selection of suitable distribution evaluated by Goodness of-fit-tests with the help of Chi-square ($\chi 2$), Kolmogorov-Smirnov (KS) and Anderson-Darling tests.
2. The Kolmogorov-Smirnov (KS) test results showed that the General extreme value, Uniform, General Pareto, Lognormal, Weibull, Log Pearson-3 and Lognormal (3P) distributions using ML are acceptable for estimation of maximum flood discharge at Awash River basins.
3. By considering the drift shape of the fitted curves using estimated maximum flood values, the study accessible that the General extreme value (GEV) distribution is better suited amongst seven distributions studied for estimation of maximum flood at Awash River basins.

6. REFERENCES

[1] Addis Hamed, K.H. and Rao,A.R.,(2000), Flood Frequency Analysis. Florida: CRC press LLC.Ababa, Ethiopia
[2] Ang and Tang,(1975), a). probability concepts in Engineering planning & design,Vol.I.
[3] Badreldin G and Fengo P, (2012), Regional rainfall frequency analysis for the luanhe basin using l-moment and cluster techniques, Procedia APCBEE, 126-135.
[4] Dessalegn, B, (2016), Evaluation of extreme flow quantiles estimated from global reanalysis runoff data,A case study of blue nile river basin, A Masters Thesis, Addis Ababa University Institute of Technology, Addis Ababa.
[5] Girma Taddese,(1998),The Water of the Awash River Basin a future challenge to Ethiopia, Addis Ababa.
[6] Kloos H. and Legess W,(2010),Water Resuorces Management in Ethiopia: Implication for the Nile Basin, Cambria Press, Innovative Publisher of Academic Research, Amsterdam, 444.
[7] Vivekananda N,(2015),Flood frequency analysis using method of moment and l-moment of probability distribution, J Cogent Engg, 2(1).