LAUTECH JOURNAL OF ENGINEERING AND TECHNOLOGY 4(2) 2007: 6 – 12

PROBABILITY DISTRIBUTION FUNCTIONS FOR THE PREDICTION OF MEAN RESERVOIR INFLOW AT HYDROPOWER DAMS IN NIGERIA

S.A. Adebara[*,a], A.W. Salami[b], and D.O. Olukanni[c]

[a]Department of Civil Engineering, University of Ilorin, Ilorin P.M.B 1515, Ilorin.
[b]Department of Civil Engineering, Kwara State Polytechnic, P.M.B, 1375, Ilorin.
[c]Department of Civil Engineering, Covenant University, Ota P.M.B 1023, Ota.

***Correspondence author**

ABSTRACT

Four methods of probability distribution analysis were evaluated for the prediction of mean reservoir inflow at hydropower dams in Nigeria; the hydropower dams include Kainji, Shiroro and Jebba dams. The reservoir inflow data were subjected to probability distribution analysis such as Gumbel (EVI), Normal (N), Log-Pearson type III (LP_3) and Log-Normal (LN), the selection of the appropriate probability distribution model for each hydropower dam was based on the result from the goodness of fit tests performed on the various probability distribution models under consideration. The goodness of fit tests considered include: Chi-square, Correlation coefficient, Coefficient of determination and Standard error of estimate. However, for both Kainji and Shiroro hydropower dams the model evaluated to be the best fit was Log-Pearson type III, this is an indication that the reservoir inflow at both dams are skewed. While at the Jebba hydropower dam the best fit model was evaluated to be Log-Normal.

Keywords: Reservoir inflow, Probability distribution models, Probability curve Fittings, Return period and Goodness of fit tests.

INTRODUCTION

The probability distribution is a hydrological tool most widely used in flood estimation and prediction. The importance of reservoir inflow analysis at any hydropower dam to our daily life makes it imperative that appropriate probability distribution model be established to determine the discharge into the reservoir. Murray and Larry (2000) stated that the choice of the probability distribution model is almost arbitrary as no physical basis is available to rationalize the use of any particular function and the search for the proper distribution function has been the subject of several studies. Salami (2004) studied the flow along Asa River and established probability distribution models for the prediction of annual flow regime. For the low flows Log-Pearson type III was recommended, while for the peak flow Gumbel extreme value type 1 was recommended. Onozo and Bayazit (1994) work on the probability distribution of largest available flood sample with the aim to determine the distribution that best fit the observed flood. Also, in one of the studies on the search for the probability distribution of floods (Benson, 1968) stated the conclusion of a work established by the Water Resources Council of the USA with the objective of developing a uniform technique of determining flood frequency, the work applied the available methods to flood records at 10 stations in various parts of the USA. Record length varied and five methods were used, Gamma, Gumbel (EVI), Log-Gumbel, Log-Normal (LN) and Log-Pearson type III (LP_3) distributions. However, no statistical test was applied to determine the goodness of fit, instead flood discharge for various return periods (2 – 50 years) were obtained from the probability plot and compared with the corresponding values from the five hypothesized distributions. Among these, the LP_3 distribution was preferred in common use, and for being capable of fitting skewed data. Cicioni et al (1973) considered Log-Normal (LN), Log-Pearson type III (LP_3) and Gumbel (EVI) distributions for the flood data from 108 stations in Italy. Statistical tests such as chi-square (χ^2), kolmogrov smirnov (ks) and probability plot correlation coefficient (ppcc) were applied and the best fitting distribution was found to be LN by the chi-square test while EVI and LP_3 by the other tests. Beard (1974) estimated the 1000 years floods at 300 stations in the USA with four different models (LN, Gamma, Log-Gumbel and LP_3) LN and LP_3 came close to reproducing the expected exceedences and were concluded to be the best. Vogel et al (1993) explored the suitability of various models to the flood flow data at 38 sites in the South-West USA, the probability distribution models adopted include Normal (N), Log-Normal (LN), Gumbel (EVI type 1) and Log-Pearson type III, which were compared graphically with the observed data. The predicted models that compare favourably well with the observed values are considered as the best distribution models. This study focuses on the evaluation of four methods of probability distribution analysis for the prediction of mean reservoir inflow at the three hydropower dams in Nigeria.

Theory on Probability Distributions and Goodness of Fit Tests

The probability distribution methods include; Gumbel (EVI type I), Normal, Log-Normal, and Log-Pearson type III. The theory on each model is briefly outlined.

Gumbel (EVI Type1) Distribution Models

It has been documented by Gumbel's extreme value distribution, that type 1 asymptotic distribution has been used successfully to represent the distribution of the yearly maximum of daily water discharge for a particular river at a specific measuring point (Viessman *et al*, 1989; Mustapha and Yusuf, 1999). The Gumbel distribution model is based on the fact that the cumulative probability that any of the events would equal or exceed a particular value having return period (T_r) are given below.

$$P = 1 - e^{-e^{-Y_T}} \tag{1}$$

$$Y_T = -\ln(-\ln(1-P)) \tag{2}$$

An event Q, having returned period T_r (years) is described by Gumbel model with a general equation of the form;

$$Q_{T_r} = Q_{av} + \sigma(0.78Y_T - 0.45) \tag{3}$$

Where Q_{av} is the average of all values, σ is the standard deviation of the series and Y_T is the reduced variants.

Normal and Log-Normal Distribution Models

The normal or Gaussian distribution is another best known statistical model and the most frequently used. It provides a reasonable approximation in the central, but is inadequate at one or both tail of the distribution (Warren *et al*, (1972) ; Mustapha and Yusuf, (1999)). The normal probability distribution model is for random variable with parameter mean (μ) and standard deviation (σ or s). The normal model has a general formula of the form given in equation (4).

$$Q_{T_r} = Q_{av} + k\sigma \tag{4}$$

For Log-Normal the observed data are transformed to its logarithmic value and the same procedure is followed as in Normal distribution method. The log-normal model has a general formula of the form given in equation (5).

$$LogQ_T = \overline{logQ} + k\sigma_{\log Q} \tag{5}$$

The parameter k in equations (4) and (5) can be selected for a particular probability or returned period from a table in a standard hydrological text books.

Log-Pearson Type III Distribution Models

The Log - Pearson type III is a probability distribution model which shows that annual flow series are rarely normally distributed, a histogram of such series is usually skewed in that the mean value does not coincide with the mode. He developed a family of curves to describe the skewness. The model is similar to the normal distribution model in estimating future events, with the additional complication of using a skew coefficient (G) given in equation (6) and the event of future year is estimated by using equation (7). The observed data are transformed to its logarithmic value and the mean , standard deviation and skewness coefficient are estimation for model development.

$$G = \frac{n^2\left(\sum \log Q\right)^3 - 3n\left(\sum \log Q\right)\left(\sum (\log Q)^2\right) + 2\left(\sum \log Q\right)^3}{n(n-1)(n-2)\left(\sigma_{\log Q}\right)^3} \tag{6}$$

$$LogQ_T = \overline{logQ} + k''\sigma_{\log Q} \tag{7}$$

Where k" can be selected for a particular probability or returned period and skew coefficient G from a table in a standard hydrological text books.

In order to ascertain whether the data obtained deviates significantly from the theoretical distribution, goodness of fit tests are required. The tests employed in this study include; chi-square test (x^2), correlation coefficient (r), coefficient of determination (R^2) and standard error of estimate (Se). These statistical tests are those that can test the independence of two criterion of classification and test whether two variables x and y are independent. This type of tests is called non-parametric tests; it requires no assumption and can be easily applied.

Chi-square Test (x^2)

This is a statistical way of testing or measuring the differences between the observed and expected frequencies in a contingency table. The chi-square is used to determine how well the theoretical distribution fit the empirical distribution. This test was based on the sum of the squares of difference between the frequencies. The expression for the analysis of chi-square is presented in equation (8).

$$\chi^2 = \sum_{j=1}^{N} \frac{(o_j - e_j)^2}{e_j} \tag{8}$$

Where o = observed flow; e = predicted flow and N = total frequency

Murray and Larry, (2000) stated that if the computed value of chi-square is greater than some critical value (such as $\chi^2_{0.95}$ or $\chi^2_{0.99}$ which is the critical values of the 0.05 and 0.01 significance level respectively), it could be concluded that the observed frequencies differ significantly from the expected frequencies and it would be rejected, otherwise it would be accepted. Hence, if the χ^2-

value calculated from equation (8) is less than critical value from statistical table, the model can be concluded to be strong or the fit of the data is good. Another way by which the conclusion can be made is that if the value of the ratio of calculated chi-square to the table chi-square ($\chi^2_{cal} / \chi^2_{tab}$) is less than one the model is strong. The model that gives value very close to 1 is the best probability distribution model for that hydropower dam. The limitation of χ^2 test is that it is highly sensitive to location of the data near the class limit and small errors in the parameter of the distribution or observed data values may have significant effect on the test result (Murray and Larry, 2000).

Correlation Coefficient (r)

A correlation coefficient is a number -1 and 1 which measures the degree to which two variables are linearly related. If there is perfect linear relationship with positive slope between the two variables, we have a correlation coefficient of 1, if there is a perfect linear relationship with negative slope between the two variables, the correlation coefficient is -1. A correlation coefficient of zero implies that there is no linear relationship between the variables.

Coefficient of correlation r is the most commonly used statistical parameter for measuring the degree of association of two linearly dependent variables. It is defined as

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} \qquad (9)$$

$-1 \leq r \leq +1$

where $\bar{x}$ and $\bar{y}$ are the mean value of x and y variables respectively.

Coefficient of Determination (R^2)

The coefficient of determination (R^2) represents the proportion of variation in the dependent variable that has been explained or accounted for by the regression line. Thus R^2 is a relative measure of the goodness of fit of the observed data points to the regression line or is a measure of the strength of relationship between the predictor and response variable. According to Dibike and Solomatine (1999), the coefficient of determination in the regression theory is defined as given in equation (10)

$$R^2 = \frac{E_o - E}{E_o} \qquad (10)$$

where

$$E_o = \sum_{l=1}^{N}\left(Q_{i(obs)} - Q_{i(mean)}\right)^2 \quad and \quad E = \sum_{i}^{N}\left(Q_{i(obs)} - Q_{i(est)}\right)^2 \qquad (11)$$

The model is said to be strong, if R^2 is very close to one.

The value of the coefficient of determination may vary from zero to one ($0 \leq R^2 \leq 1$). A coefficient of determination of zero indicates that none of the variation in y is explained by the regression equation; where as a coefficient of determination of 1 indicates that 100 % of the variations of y have been explained by the regression equation. That is, the regression line perfectly fits all the observed data points. For example, if $R^2 = 0.4$ this means that 40% of the total variation in the observed values of y is explained by the observed values of x. Therefore, the better the fit, the closer will R^2 lie towards 1.

Standard Error of Estimate (Se)

This is a measure of the spread about the regression line of y versus x or y with respect to x and is given in equation (12).

$$Se = \sqrt{\frac{1 - r^2}{n - 2}} \qquad (12)$$

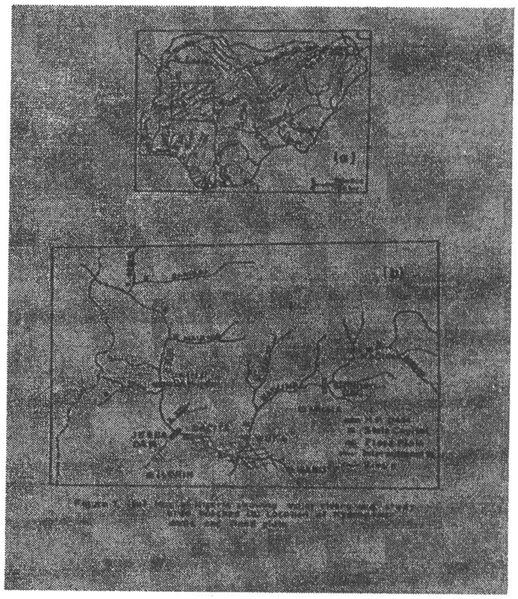
The closer the value of Se to zero implies that the distribution function best fit the data.

where r = correlation coefficient and n – 2 = degree of freedom

METHODOLOGY

Data Collection

The reservoir inflow data were collected from the three hydropower stations in Nigeria; these are Kainji, Shiroro and Jebba hydropower stations. The locations of the hydropower stations are shown in Figure 1. The data were available with the hydrological unit of each hydropower stations. A total of 35 years inflow data (1970 - 2004) were collected from Kainji, while a total of 15 years inflow data (1990 - 2004) were collected from Shiroro and a total of 21 years inflow data (1984 - 2004) were collected from Jebba.



Data Analysis

The mean reservoir inflow was estimated from the collected data for each year and ranked according to Welbull's plotting position. The probability of event p ($X \geq x$) was estimated using equation (13), while the return period (T_r) was estimated as the reciprocal of the probability, equation (14), Mustapha and Yusuf (1999).

$$P = \frac{m}{n+1} \qquad (13)$$

$$T_r = \frac{1}{P} \qquad (14)$$

Where m is the series of events ranking 1 for highest value and so on in descending order and n is the number of event in the series. The mean reservoir inflow data were evaluated with six methods of probability distribution function to determine the best – fit model for each of the hydropower stations.

Evaluation of Probability Distribution Models

The probability distribution analysis was carried out in accordance to standard procedure (Wilson, 1969; Viessman et al, 1989; Mustapha and Yusuf, 1999). The mathematical expressions obtained for various probability distributions functions are presented in Table 1.

Testing of the Probability Distribution Models

The suitability of the developed probability models were tested by using four statistical tests (goodness of fit tests) presented in sub-section 2.2. The statistical tests include chi-square (χ^2), probability plot coefficient of correlation (r), coefficient of determination (R^2) and Standard error of estimate (Se). The statistical tests were carried out in accordance with standard procedure (Chowdhury and Stedinger (1991); Adegboye and Ipinyomi (1995); Murray and Larry (2000)). The results obtained for chi-square, ppcc (r), R^2 and Se tests were presented in Table 2 along side with the ranking of the distribution models. The observed and predicted mean reservoir inflows were plotted against cumulative probability in order to compare the observed mean flow with the predicted values based on the probability distribution models established. The graphical comparisons are presented in Fig. 1 – 12.

Table 1: Model Equation for the probability distributions

S/N	Hydropower dams	Probability Distributions	Developed equations
1.	Kainji	Gumbel (EVI)	$Q_T = 759.050 + 131.150Y_T$
		Normal	$Q_T = 838.180 + 175.840\ K_T$
		Log – Normal	$\text{Log}\ Q_T = 2.915 + 0.089\ K_T$
		Log – Pearson	$\text{Log}\ Q_T = 2.915 + 0.089\ K'_T$
2.	Shiroro	Gumbel (EVI)	$Q_T = 265.940 + 22.210Y_T$
		Normal	$Q_T = 278.760 + 28.480\ K_T$
		Log – Normal	$\text{Log}\ Q_T = 2.443 + 0.049\ K_T$
		Log – Pearson	$\text{Log}\ Q_T = 2.443 + 0.049\ K'_T$
3.	Jebba	Gumbel (EVI)	$Q_T = 845.520 + 487.080Y_T$
		Normal	$Q_T = 1126.520 + 624.460\ K_T$
		Log – Normal	$\text{Log}\ Q_T = 2.948 + 0.115\ K_T$
		Log – Pearson	$\text{Log}\ Q_T = 2.948 + 0.115\ K'_T$

Table 2: Results of statistical test for the selection of the best probability distribution model and ranking of the models

Statistical Tests	(Chi-square)	Correlation coefficient	Coefficient of determination	Standard error of estimate	Best fit model rank
Stations	$\chi^2_{cal} / \chi^2_{0.95}$	R	R^2	Se	
Kainji	EVI (0.9500)	EVI (0.8500)	EVI (0.9600)	EVI (0.0156)	2nd
	N (1.2300)	N (0.8900)	N (0.9400)	N (0.0187)	4th
	LN (0.8700)	LN (0.8600)	LN (0.9600)	LN (0.0141)	3rd
	LP_3 (0.8900)	LP_3 (0.9200)	LP_3 (0.9600)	LP_3 (0.0111)	1st
Jebba	EVI (45.3815)	EVI (1.9700)	EVI (0.0100)	EVI (0.0670)	3rd
	N (95.5738)	N (2.1100)	N (-0.3300)	N (0.0820)	4th
	LN (2.6200)	LN (0.8800)	LN (0.9600)	LN (0.0140)	1st
	LP_3 (8.8958)	LP_3 (1.0300)	LP_3 (0.8800)	LP_3 (0.0310)	2nd
Shiroro	EVI (0.4791)	EVI (0.9900)	EVI (0.9900)	EVI (0.0034)	2nd
	N (0.3682)	N (1.0100)	N (0.9900)	N (0.0025)	3rd
	LN (0.3953)	LN (1.0100)	LN (0.9900)	LN (0.0027	4th
	LP_3 (0.3540)	LP_3 (1.0000)	LP_3 (0.9900)	LP_3 (0.0025)	1st

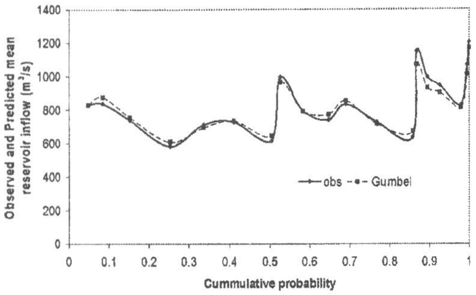


Fig. 1 Comparison of observed data to Gumbel predicted data for Kainji H.P dam

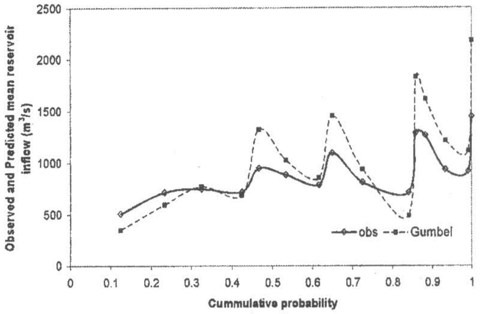


Fig. 5 Comparison of observed data to Gumbel predicted data for Jebba H.P dam

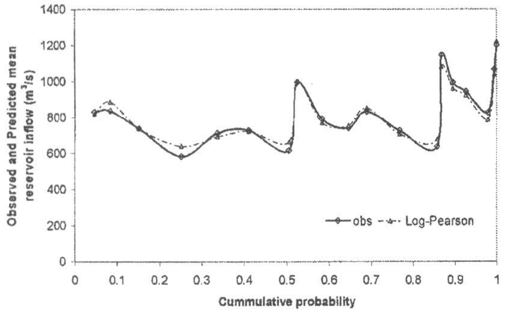


Fig. 2 Comparison of observed data to Log-Pearson predicted data for Kainji H.P dam

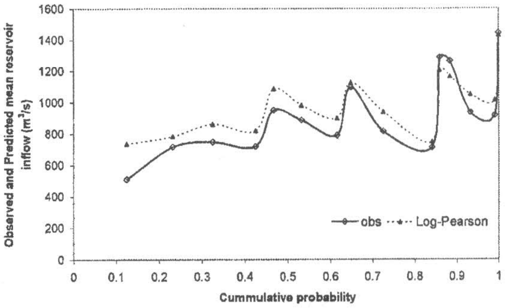


Fig. 6 Comparison of observed data to Log-Pearson predicted data for Jebba

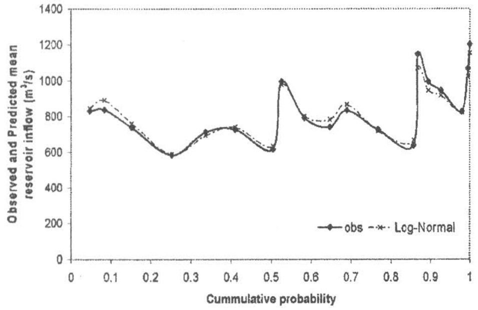


Fig. 3 Comparison of observed data to Log-Normal predicted data for Kainji H.P dam

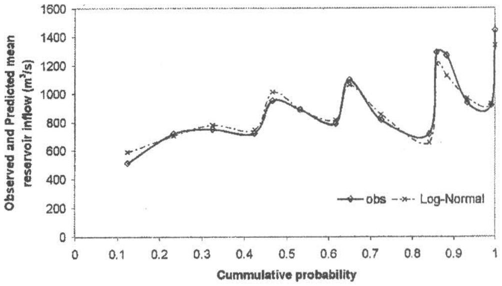


Fig. 7 Comparison of observed data to Log-Normal predicted data for Jebba H.P dam

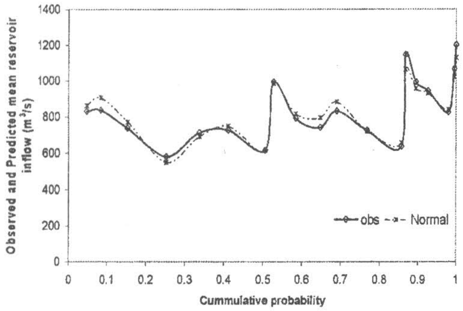


Fig. 4 Comparison of observed data to Normal predicted data for Kainji H.P dam

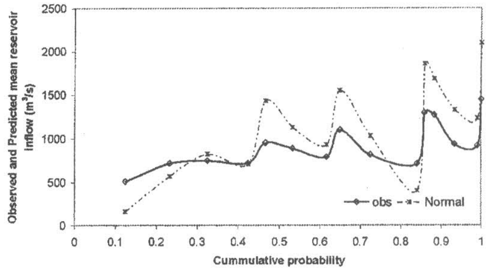


Fig. 8 Comparison of observed data to Normal predicted data for Jebba H.P dam

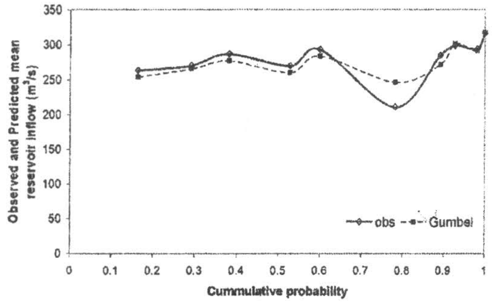


Fig. 9 Comparison of observed data to Gumbel predicted data for Shiroro H.P dam

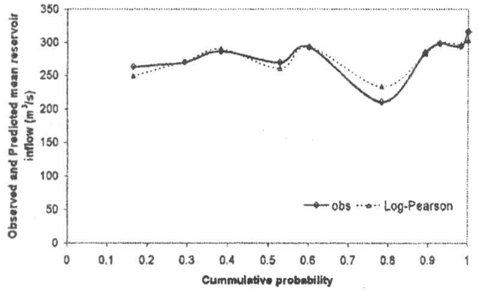


Fig. 10 Comparison of observed data to Log-Pearson predicted data for Shiroro H.P dam

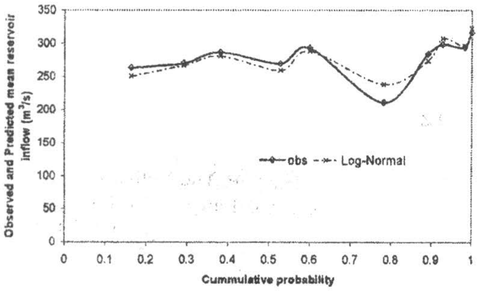


Fig. 11 Comparison of observed data to Log-Normal predicted data for Shiroro H.P dam

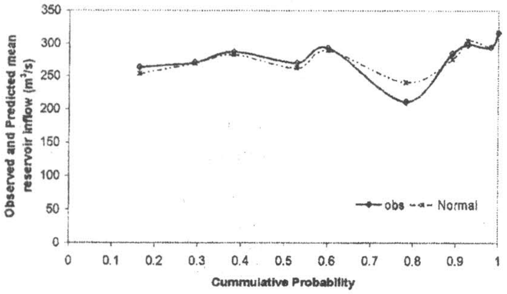


Fig. 12 Comparison of observed data to Normal predicted data for Shiroro H.P dam

RESULTS AND DISCUSSION

A total of 35, 21 and 15 years reservoir inflow data were obtained from Kainji, Jebba and Shiroro hydropower stations respectively. The reservoir inflows are evaluated by various probability distribution functions to determine the best fitting model, the mathematical representations of the evaluated probability functions are presented in Table 1. For the purpose of theoretical determination of best probability function, statistical tests were adopted. The statistical tests adopted include chi-square (χ^2), probability plot correlation coefficient (r), coefficient of determination (R^2) and the standard error of estimate Se. The results of the statistical tests are presented in Table 2 along side with the ranking of the distribution models. Also for the purpose of comparison and to select the best fit model, the observed and predicted inflow data were plotted against cumulative probability as presented in Figures 1 – 12.

The mean reservoir inflow at Kainji hydropower station has a value of (χ^2_{cal} / χ^2_{tab}), r, R^2 and Se as 0.9500, 0.8500, 0.9600 and 0.0156 respectively for Gumbel extreme value (EVI) type I distribution, while a value of (χ^2_{cal} / χ^2_{tab}), r, R^2 and Se as 0.8900, 0.9200, 0.9600 and 0.0111 respectively for Log-Pearson type III (LP_3) distribution. From this result chi-square test suggest EVI, while other tests suggest LP_3 as the best fit model for the mean reservoir inflow data. The indication of a higher value of correlation coefficient (r) for LP_3 shows that there is a close linearity between the observed and the predicted reservoir inflow. But based on the graphical comparison (Fig. 1 - 4) the Gumbel (EVI typeI) distribution model has its curve closer to that of the observed mean reservoir inflow better than that of other probability distribution models. Hence, statistical tests suggest Log-Pearson type III as the most appropriate model for the mean reservoir inflow at Kainji hydropower dam and graphical comparison suggest Gumbel (EVI typeI) distribution thus both are selected as the best fit models.

The mean reservoir inflow at Jebba hydropower station has a value of (χ^2_{cal} / χ^2_{tab}), r , R^2 and Se as 2.6200, 0.8800, 0.9600 and 0.0140 respectively for Log-Normal (LN) distribution, while a value of (χ^2_{cal} / χ^2_{tab}), r, R^2 as 8.9000, 1.0300, 0.8800 and 0.0310 respectively for Log-Pearson type III (LP_3) distribution. From this result chi-square test did not satisfied the condition for selection of any model, also the value of correlation coefficient (r) for LP_3 did not satisfied the condition for selection of model. However, the results of other tests suggest Log-Normal as the best fit model for the mean reservoir inflow data and based on the graphical comparison (Fig. 5 and 8) the Log-Normal distribution model has its curve closer to that of the observed mean reservoir inflow better than that of other probability distribution models. Hence, Log-

Normal is the most appropriate model for the mean reservoir inflow at Jebba hydropower dam and thus selected as the best fit model.

The mean reservoir inflow at Shiroro hydropower station has a value of (χ^2_{cal} / χ^2_{tab}), r , R^2 and Se as 0.4791, 0.9900, 0.9900 and 0.0034 respectively for Gumbel extreme value (EVI) type I distribution, while a value of (χ^2_{cal} / χ^2_{tab}), r, R^2 and Se as 0.3540, 1.0000, 0.9900 and 0.025 respectively for Log-Pearson type III (LP_3) distribution. The statistical tests follow the same trend as in the case of Kainji hydropower. That is, the chi-square test suggests EVI, while other tests suggest LP_3 as the best fit model for the peak reservoir inflow data. The indication of a higher value of correlation coefficient (r) for LP_3 also shows that there is a close linearity between the observed and the predicted reservoir inflow. Also, based on the graphical comparison (Fig. 9 and 12) the Log-Pearson distribution model has its curve closer to that of the observed mean reservoir inflow better than that of other probability distribution models. Hence, Log-Pearson type III is the most appropriate model for the mean reservoir inflow at Shiroro hydropower dam and thus selected as the best fit model.

The best fit probability distribution model for the prediction of the peak reservoir inflow at each hydropower station is presented in Table 3

Table 3: Best - fit probability distribution models for mean reservoir inflow

S/N	Hydropower dams	Best – fit models
1.	Kainji	Log – Pearson and Gumbel (EVI type I)
2.	Jebba	Log - Normal
3.	Shiroro	Log - Pearson

CONCLUSION

Various probability distribution models were fitted to the peak reservoir inflow records to evaluate the model that is most appropriate for the prediction of peak reservoir inflow at the three hydropower stations in Nigeria. Various models were established for each hydropower station and the suitable model was selected based on the goodness of fit tests. The log-Pearson type III probability distribution model was found to be appropriate for both the Kainji and Shiroro hydropower dams, while Log-Normal was found to be appropriate for Jebba hydropower dam. The establishment of the best fit probability distribution model would be of useful guide in the prediction of the near future peak reservoir inflow at the three hydropower dams. Also, the Log-Pearson type III model that adequately fit the reservoir inflow at two of the hydropower dams indicates that the inflow data are skewed.

REFERENCES

Beard, L.R (1974). "Flood flow frequency techniques", Center for Research in Water Resources University of Texas.

Benson, M.A (1968). "Uniform flood estimating methods for Federal agencies" Journal of Water Resources. 4 (5); PP 891 – 908.

Cicioni, G, Guiliano, G and Spaziani, F.M (1973). "Best fitting of probabilityfunctions to a set of data for flood studies", Water Resources Publication, Fort Collini, Company,

Dibike, B.Y and Solomatine, D.P. (1999). "River flow forecasting using Artificial Neural Networks" Paper presented at European Geophysical Society (EGS) XXIV, General Assembly, The Haugue, The Netherlands, 1 – 11.

Murray, R.S and Larry, J.S (2000). "Theory and Problems of Statistics" Tata McGraw – Hill Publishing Company Limited, New Delhi, pp. 314 – 316, 3rd edition

Mustapha, S and Yusuf, M.I (1999). "A Textbook of Hydrology and Water Resources" Jenas Prints & Publishing Co., Abuja, Nigeria, pp.164 – 184, !st edition.

Salami, A.W (2004). "Determination of statistical distribution model for flow analysys at Asa gauging station, Ilorin, Nigeria. Nigerian Journal of Technological Development (NJTD). Vol. 4 : 49 – 53

Viessman,W, Krapp, J.W and Harbough, T.E (1989). "Introduction to Hydrology".Harper and Row Publishers Inc., New York. 675 – 695.

Vogel, R.M, Thomas, W.O and Mahon, T.A (1993). "Flood flow frequency model selection in South Western United States", Journal of Water Resources Planning Management. ASCE 119 (3): 353 – 366.

Warren, V, Terence, E.H and John, W.K (1972). "Introduction to hydrology" Published, Intext Educational Publishers, New York, 106 – 141.