MAXIMUM LIKELIHOOD ESTIMATION OF THE WATER LEVELS IN THE TONO DAM, GHANA

Solomon SARPONG

University for Development Studies, Navrongo Campus

GHANA

ssarpong@uds.edu.gh

ABSTRACT

The quest of this research is to find the most appropriate probability distribution function that best approximates the water level in the Tono dam. The data used consist daily water level recordings from January 2010 to January, 2017. The sample size of the data is 2570. However, 2195 data points were used in the analysis whilst the remaining 375 was used for validation. After various probability distribution functions were fitted to the data, it was observed that the Weibull distribution best fits the data. From the Weibull distribution fitted, it can be observed that the level of the water in the Tono dam is dwindling with time.

Keywords: Weibull distribution, reservoir, probability distribution, validation

INTRODUCTION

Reservoirs and river basins are the most significant components of water resource management, providing effective multipurpose water storage that is employed for irrigation, water supply, hydropower, and flood drought control. To most effectively use this stored water, it is essential to optimally monitor the reservoir level to obtain the desired performance, Valizadeh *et al*., (2014).

Reservoir is one of the structural defense mechanism for flood. During heavy rainfall, reservoir hold excessive amount of water to reduce flood risk at downstream area. During less rainfall, reservoir maintains the water supply for major uses such as domestic and commercial usage. Early decision regarding the water release can be made if the future water level can be forecasted. Ishak *et al*., (2010) explored the potential of neural network model for forecasting the reservoir water level. Neural network model 24-15-3 was observed to be the best model.

It must be noted that, accurate prediction and monitoring of water level in reservoirs is an important task for the planning, designing, and construction of river-shore structures, and in taking decisions regarding irrigation management and domestic water supply. Using a historic data set of 22 years, Ghosh *et al.,* (2016) proposed a novel probabilistic nonlinear approach based on a hybrid Bayesian network model with exponential residual correction for prediction of reservoir water level on daily basis of Mayurakshi reservoir (Jharkhand, India).

Alam, Khan & Rahat (2014) observed that, dagum (4P) fit best for annual maximum water level whereas Cauchy was found best for annual maximum discharge in their quest to find the best model that approximates the annual maximum water level and annual maximum discharge. Izinyon and Ajumuka (2013), found that the best fit probability distribution models obtained for the different stations are log normal, log normal and log Pearson type III for the stations at River Katsina Ala at Serav, River Taraba at Garsol and River Mayokam at Mayokam respectively. Vivekanandam (2015) examined a number of probability distributions (Exponential, Gamma, Generalized Extreme Value, Generaliszed Pareto, Extreme Value Type 1 and Pearson Type 3). It was observed that, Gamma distribution was found to be better suited among the six distributions adopted in the estimation of MFD at Dedtalai and Pearson type-3 distribution for Ghala.

Preforming frequency analysis, Garba, Ismail, & Tsoho (2013) fitted probability distribution functions of Normal, Log Normal, Log Pearson type III and Gumbel to the discharge variability of Kaduna River at Kaduna South Water Works. From the measure of discrepancy, they observed that at selected level of significance of $\alpha = 1\%$, $\alpha = 5\%$, and $\alpha = 10\%$, all the four theoretical distribution functions were acceptable. Using 23 years of hydrological data collected from Sukhi reservoir project Gujarat State, Rani and Parekh (2014) sought to find the best model that will be appropriate predictor for real-time water level forecasting. They observed that, artificial neural network using forward back propagation is the most appropriate model for the prediction.

Some researchers have used time series, ARIMA, to model and forecast dam water levels. To the best of my knowledge, the use of time series has got some disadvantages. The use of ARIMA models assumes that there is the existence of linear relationship between the variables but in real-world, data are often nonlinear, Lin, Chiu & Lin (2012), Huang, Chuang, & Wu (2010), Ding, Li & Li (2009) and Gradojevic and Yang, (2006). Secondly, the ARIMA model selection procedure depends greatly on the competence and experience of the researchers to yield desired results. Unfortunately, the choice among competing models can be arbitrated by similar estimated correlation patterns and may frequently reach inappropriate forecasting results Ridhwan *et al.*, (2015).

Bessa (2011) observed that the best way to represent uncertainty is determined by end-user requirements and decision-making problems. In general, one cannot talk about better and worse uncertainty representations, only of more or less adequate representations. However, unlike ARIMA, a probability density function (pdf) gives the necessary flexibility for several decision-making problems.

Describing five alternative water supply scenarios, Alfarra (2012) reports on the implementation and calibration of the WEAP (Water Evaluation and Planning) model against dam operating rules, showing that it is possible to reproduce historical dam volumes accurately enough by analysis.

In the recent past, rainfall pattern has been changing. Due to the uncertainty in the quantity of rainfall to be expected during the raining season, there is the need to dam some rivers. The water in these dams can then be used for household use, irrigation and other purposes. As the dams are rain-fed, there is the need to monitor the level of water in the dam so as to mitigate and forestall any changes. It is in lieu of this that this research was undertaken.

This research seeks to find an appropriate ways to monitor the dam water level in the Tono dam.

The research paper is organized as follows: the study area and data; the methodology; results and discussion; conclusion.

STUDY AREA AND DATA

Study Area

The area of the study is in the Upper East of Ghana, West Africa. The Upper East region is located in the north-eastern corner of Ghana. It is the second smallest of 10 administrative regions in Ghana, occupying a total land surface of 8,842 square kilometers or 2.7 per cent of the total land area of Ghana. It lies between longitude 0° and 1° west, and latitudes 10° 30'N and 11°N. The region shares boundaries with Burkina Faso to the north, Togo to the east, Upper West Region to the west, and the Northern Region to the south. The regional capital is Bolgatanga (Population and Housing Census, 2010). The rainy season in this area is from May/June to September/October with mean annual rainfall between 800 mm and 1100 mm. There is a long spell of dry season from November to mid-February.

Data Collection

The data used for this research was the daily water level recorded at the Tono dam. The dam is in the Upper East region of Ghana. This dam is used basically for irrigation whilst the locals fish in it. The water was dammed to provide water for irrigation for the farmlands as there is not enough rainfall in the Upper East region of Ghana.

METHODOLOGY

The data used in this research was obtained from the Tono irrigation dam in the Upper East Region of Ghana. The data consist daily water level recordings from January 2010 to January, 2017. The sample size of the data is 2570. However, 2195 data points were used in the analysis whilst the remaining 375 was used for validation. The research sought to find the best distribution that can adequately be used to forecast the water level in the Tono dam. The nature of skewness observed was used to eliminate some of the distributions obtained. The negative skewness of the data led to the elimination of all positively skewed distributions. For each of the remaining negative distribution, the goodness-of-fit was performed using Chi-square, Kolmogorov-Smirnov and Anderson Darling tests. The distribution with the best test of goodness-of-fit statistics adequately fits the data.

Akaike and Bayesian Information Criteria

The general form of Akaike information criterion, AIC, is given by: $AIC = -2lnL + 2 * K$ where *ln* is the natural logarithm; *L* is the value of the likelihood; *K* is the number of parameters in the model. AIC can also be calculated using residual sums of squares from regression: $AIC = n * ln\left({}^{RSS}/_{n}\right) + 2 * K$ where *n* is the number of data points; RSS is the residual sums of squares; AIC requires a bias-adjustment small sample sizes (if ratio of ${}^{n}/_{K} < 40$), then the bias adjustment can be used; $AICc = -2 * L + 2 * K + \left(2 * K * (K + 1)\right)/(n - K - 1)$. AIC takes into account both the statistical goodness-of-fit and the number of parameters that have to be estimated to achieve this particular degree of fit, by imposing a penalty for increasing the number of parameters. Lower values of the index indicate the preferred model, that is, the one with the fewest parameters that still provides an adequate fit to the data, Akaike (1977).

Bayesian Information Criteria, $BIC = -2lnL + kln(n)$ but under the assumption that the model errors or disturbances are independent and identically distributed, $BIC = -2ln(\hat{\sigma}^2) + kln(n)$; where $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$. Under the assumption of unbiasedness of the error variance, $BIC = -2ln\left(\hat{\hat{\sigma}}^2\right) + kln(n)$, where $\hat{\hat{\sigma}}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$. With the assumption of normality, $BIC = \chi^2 + kln(n)$. Where: *n* is the number of data points in x; *K* is the number of parameters in the model; *L* is the value of the likelihood; $\bar{x}$ is the mean of the observations in the data, x, Schwarz (1978).

Even though both Akaike and Bayesian information criteria helps in obtaining the best model, each of these try to balance model fit and parsimony of variables and penalizes differently for the number of parameters. These information criteria were used to compare the distribution that best fits the data. In order to get the best model fit, both statistics were be used.

Anderson Darling Test

The Anderson-Darling test is a statistical test of whether a given sample of data is drawn from a given probability distribution. In its basic form, the test assumes that there are no parameters to be estimated in the distribution being tested, hence distribution-free. The Anderson-Darling test assessed if a given data $\{Y_1 < Y_2 < \cdots Y_N\}$ comes from a cumulative distribution function, *F*. The test statistic is given by; $A^2 = -N - S$ where $S = \sum_{i=1}^{N}\frac{(2i-1)}{N}\left[lnF(Y_i) + ln\left(1 - F(Y_{N+1-i})\right)\right]$; *F* is the cumulative distribution function of the specified distribution; Y_i is the ordered data; *N* is the sample size of the data.

Kolmogorov-Smirnov Test

In addition to the Kolmogorov-Smirnov test being an exact test, it does not depend on the underlying cumulative distribution function being tested. Unlike the Anderson-Darling test, the distribution must be specified. The Kolmogorov-Smirnov test statistic is given by $D =$

$\underset{1\leq i\leq N}{max}\left\|F(Y_i)-\frac{i}{N}\right\|$where F is the theoretical cumulative distribution of the distribution being tested (the distribution must be continuous) and must also be fully specified. The null hypothesis for the Kolmogorov-Smirnov test is the data follows a specified distribution.

Chi-squared Test

The chi-squared test is given by $X^2 = \sum_{i=1}^{N}\frac{(O_i-E_i)}{E_i}$ where O_i is the observed frequency for *i;* E_i is the expected frequency for *i* and it given by $E_i = N\big(F(Y_u) - F(Y_l)\big)$; where F is the cumulative distribution function Y_u and Y_l are the upper and lower limits for class *i*. The test statistic is approximately chi-squared distributed with $(k-c)$ degrees of freedom where k is the number of empty cells and c is the number of estimated parameters for the distribution plus one. With the null hypothesis for the chi-squared test given as the data follows a specified distribution, the actual distribution a data follows can be known.

RESULTS AND DISCUSSIONS

Descriptive Statistics

The data was the daily recording of the water level of the Tono. The Tono dam is an irrigation dam in the Kassena-Nankana district in the Upper East region of Ghana. The data was the dam water level recordings from January, 2010 to January, 2017. However, the dam water level recordings from January 2010 to September 2016 was used as the training data. The remainder of the data (October, 2016 – January, 2017) was used to validate the model.

From the analysis, it was observed that the maximum water level recording was 179.91, whilst the minimum was 171.8. The mean water level recording for the period was 176.48 with standard deviation 1.98. The kurtosis was 2.41. As shown in figure 1, it can be observed that the graph tails off to the left hence, it is negatively skewed with value -0.39.

The quest of this research is to find the model that best fits the dam water levels and forecast adequately. In order to do this, a plot of the empirical density and the cumulative distributions were made. These PDFs help to explain the likely distribution that will best fit the data.

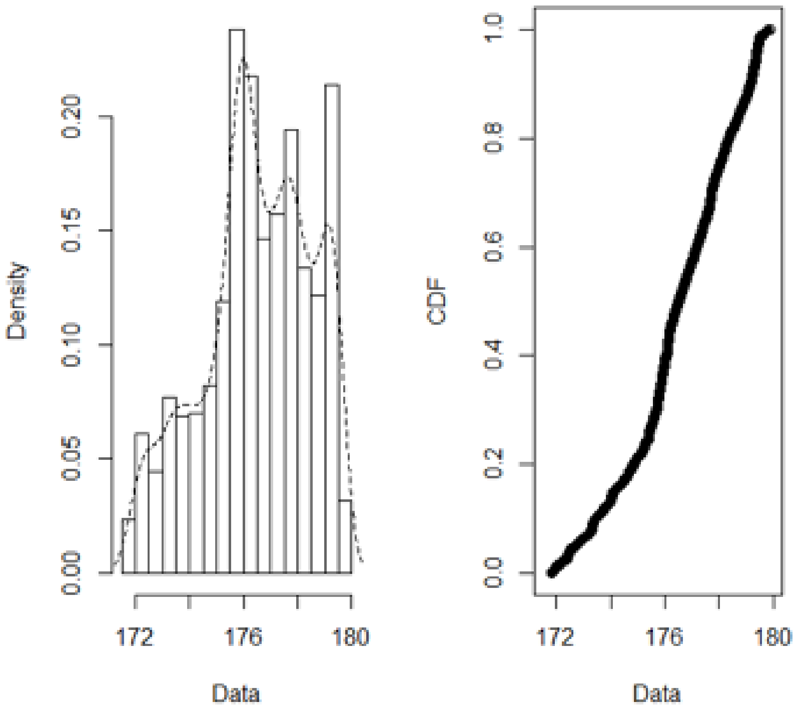

Figure 1: The empirical density (left) and Cummulative distribution function (right) plots for the dam level recording.

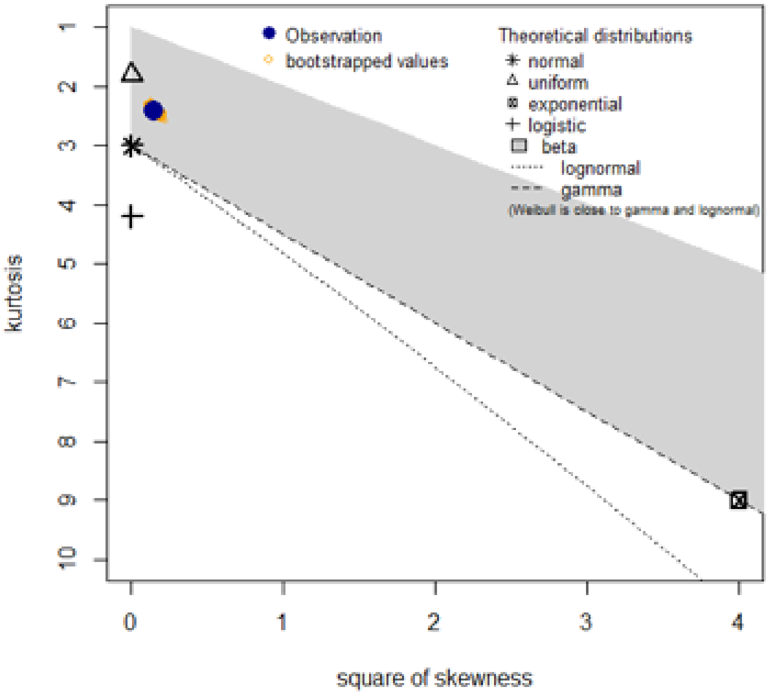

Figure 2: Skewness and Kurtosis plot for the dam water level recording.

In figure 2 shows several models that are possible fit for the data. The accuracy of the possible models present in the data was increased by bootstrapping the data. As shown in figure 2, the theoretical distributions present in the data include; Normal, Uniform, Beta, Exponential, Logistics, Lognormal, Gamma, and Weibull. Among the possible models in the data, models that have negative skewness will be investigated for a possible fit to the data. Figure 3 shows the graphical plots of the probability density functions identified in the data. From figure 3, it was observed that Logistics, Lognormal, Weibull, and Gamma distributions have negative skewness.

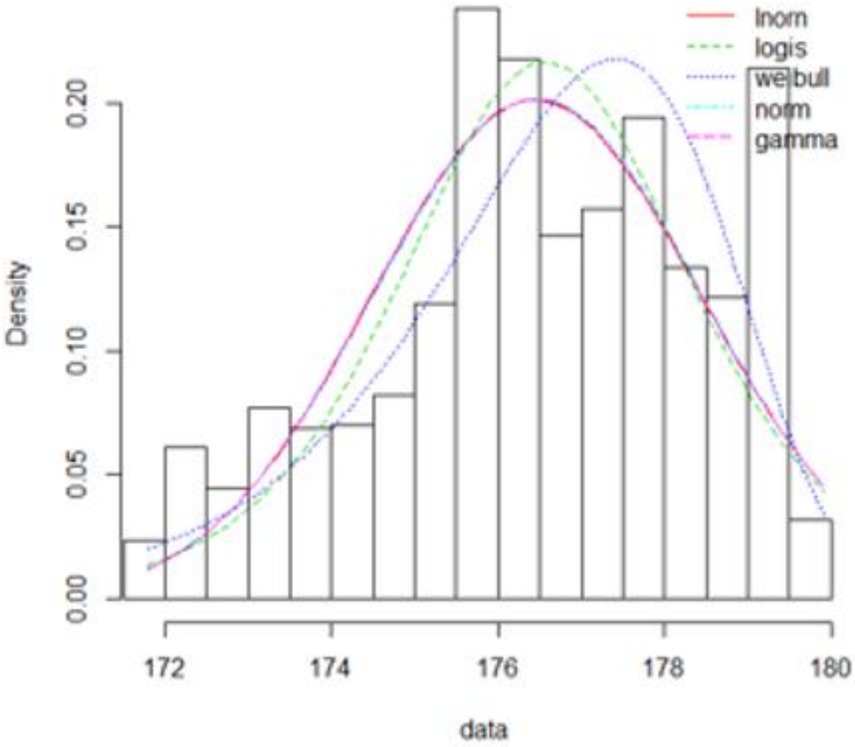

Figure 3: A plot of the histogram and theoretical densities the dam water level recording.

The Lognormal distribution was investigated to find out if it was a possible fit to the data. There was a log of mean value of 5.17 and log of standard deviation of 0.01. The Akaike and Bayesian Information Criteria were 10388.62 and 10400.24 respectively. Figure 4 shows the empirical and theoretical density and cumulative density plots; as well as the Q-Q and P-P plots. These plots show how the lognormal distribution approximates the data.

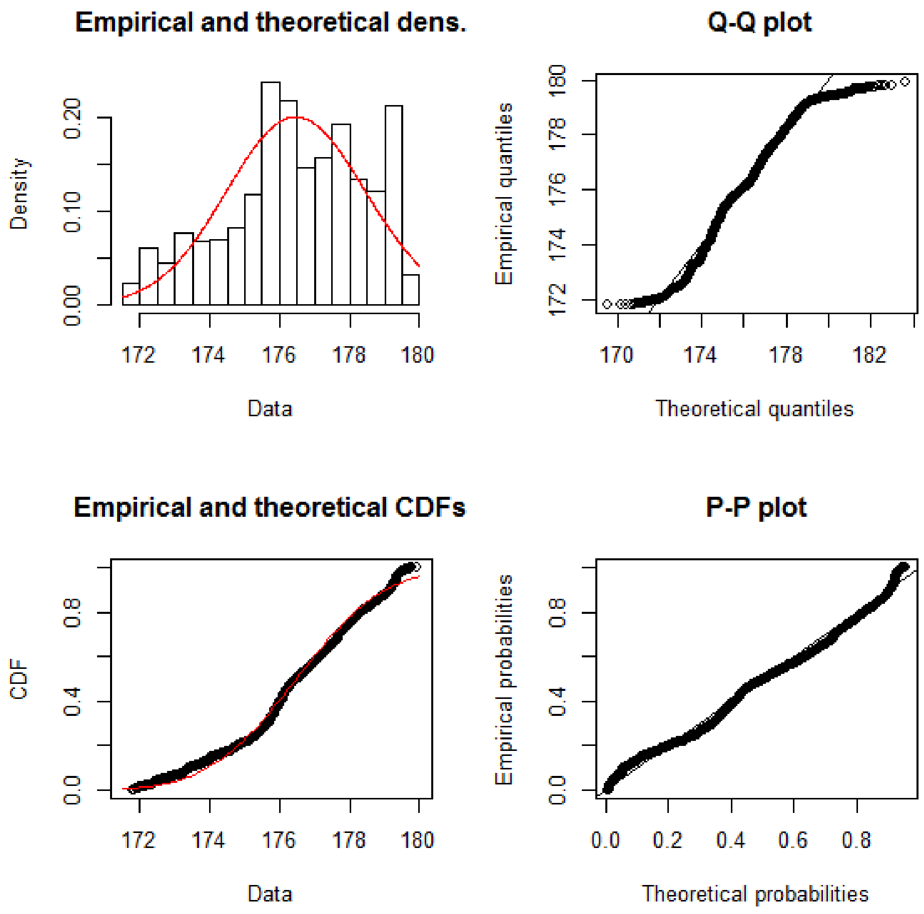

Figure 4: Graph of lognormal distribution representation for the data.

Logistic distribution was also investigated to verify if it is a possible fit to the data. From the analysis it was observed that the location and scale parameters were 176.57 and 1.15 respectively. Also, the Aikake and Bayesian Information Criteria were 10491.51 and 10503.13 respectively. Figure 5 shows the empirical and theoretical density and cumulative density plots; as well as the Q-Q and P-P plots. These plots show how the logistic distribution approximates the data.

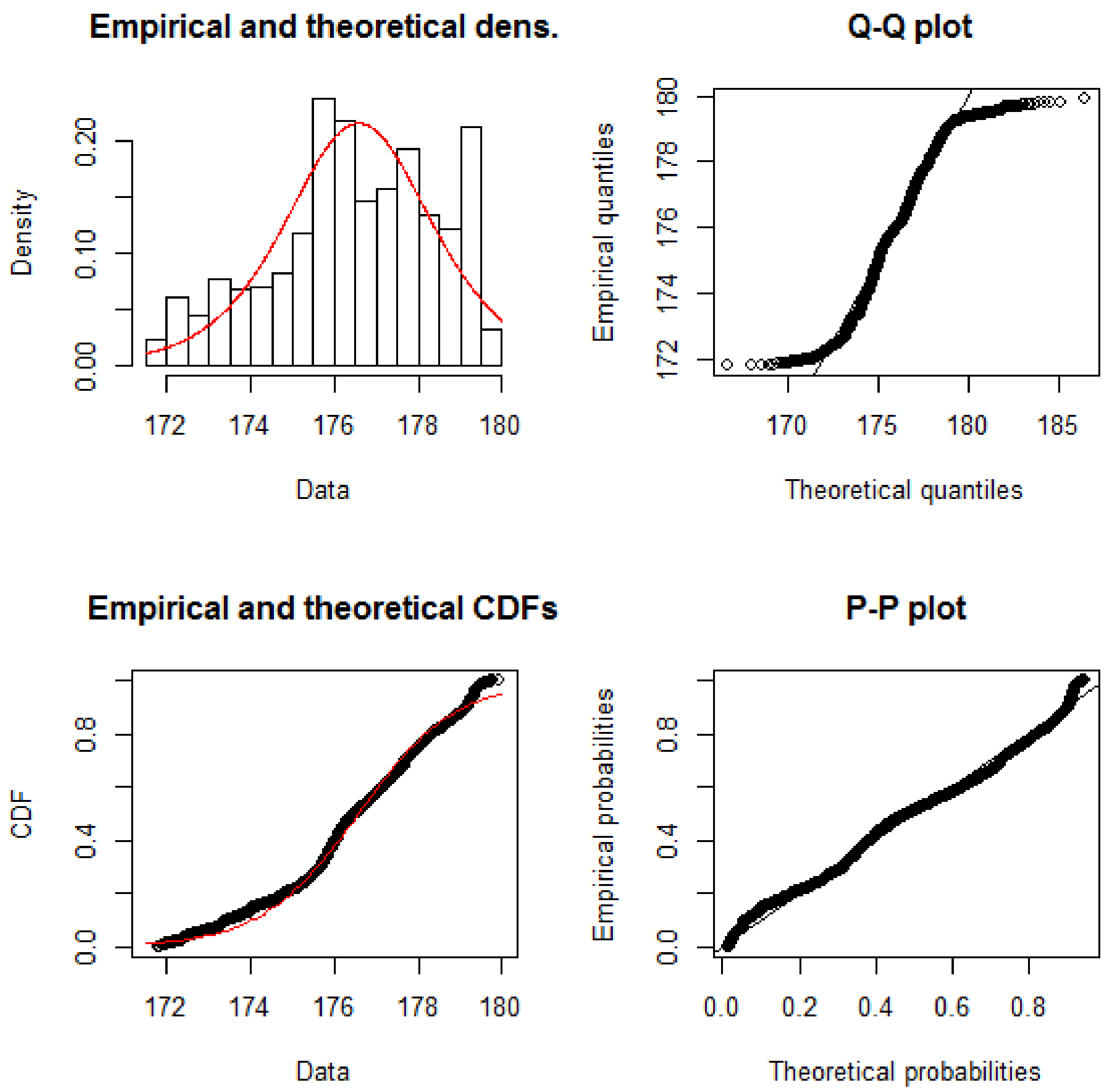

Figure 5: Graph of logistic distribution representation for the data.

Also, the Weibull distribution was investigated to verify if it is a possible fit to the data. From the analysis it was observed that the shape and scale parameters were 105.13 and 177.42 respectively. Also, the Aikake and Bayesian Information Criteria were 10293.03 and 10293.03 respectively. Figure 6 shows the empirical and theoretical density and cumulative density plots; as well as the Q-Q and P-P plots. These plots show how the Weibull distribution approximates the data.

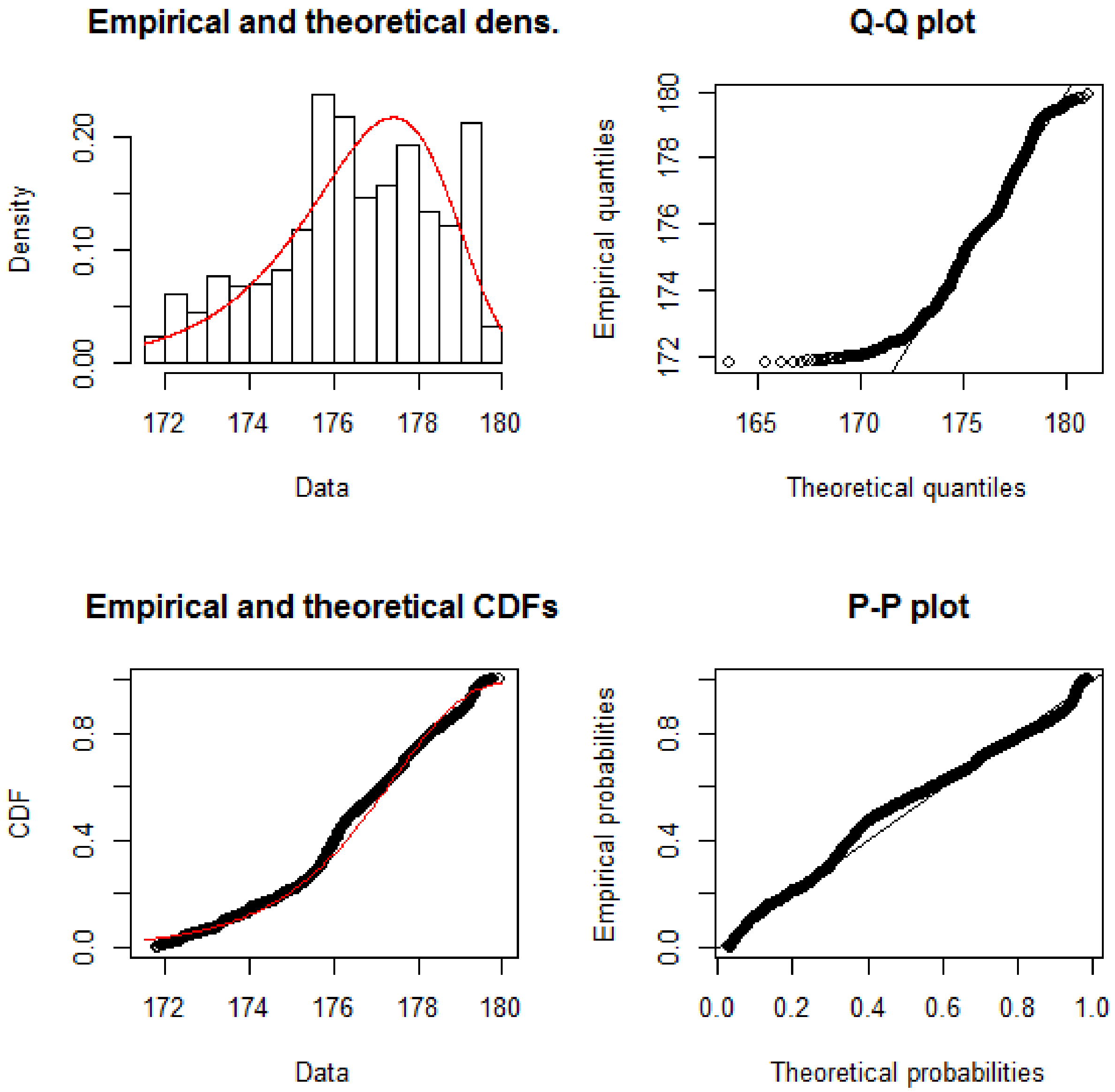

Figure 6: Graph of Weibull distribution representation for the data.

The gamma distribution was investigated to find out if it was a possible fit to the data. The shape and rate parameters were 7941.27 and 445.00 respectively. The Akaike and Bayesian Information Criteria were also 10384.89 and 10396.51 respectively. Figure 8 shows the empirical and theoretical density and cumulative density plots; as well as the Q-Q and P-P plots. These plots show how the gamma distribution approximates the data.

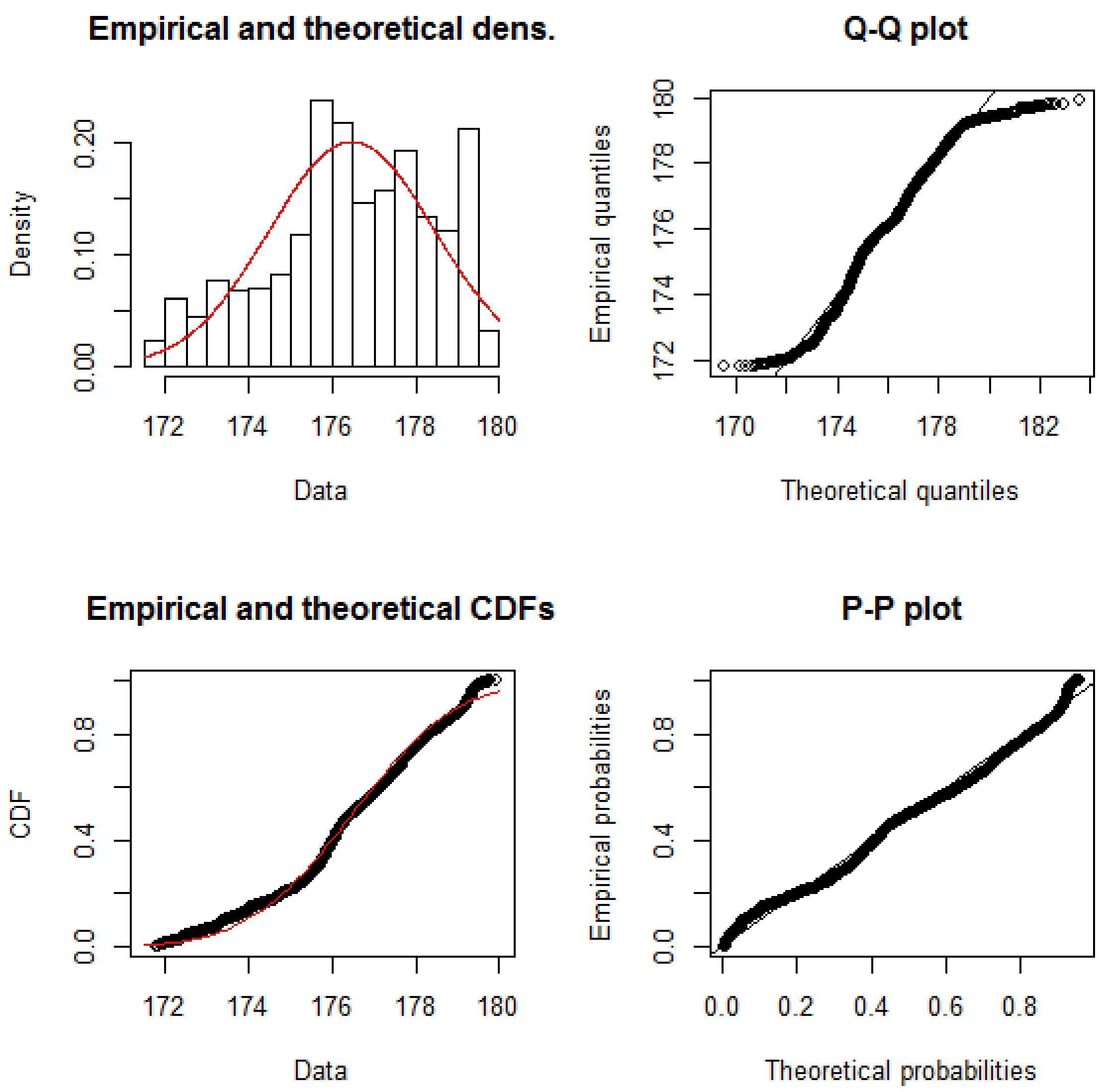

Figure 7: Graph of Gamma distribution representation for the data.

In order to ascertain the best model that fits the data, the goodness-of-fit criteria were used. The statistics that were considered were Bayesian and Akaike's information criteria. The best distribution is the one that has the lowest values for these statistics. From table 1, it can be observed that Weibull distribution has the lowest values for all these statistics. Hence, the data is likely to follow the Weibull distribution.

Table 1: The Goodness-of-fit Criteria for the dam water level recording.

Distribution	**Aikaike Information Criterion**	**Bayesian Information Criterion**
Log-normal	10388.62	10400.24
Gamma	10384.89	10396.51
Logistic	10491.51	10503.13
Weibull	10293.03	10304.66

The goodness-of-fit criteria was used to verify the adequacy of Weibull distribution fitting the data. In order to do this, the Anderson-Darling, Cramer-Von Misses, and Kolmogorov-Smirnov were used. The statistics from these goodness-of-fit tests confirm that the data on the dam water level follows the Weibull distribution. From the Wilcoxon signed rank test with continuity correction, it was observed that at 5% level of significance (*p*-value = 0.001) the simulated and the actual data are identically distributed.

Chang and Melick (1999) were of the view that, one way of judging the adequacy of a derived probability density function is from its empirical performance. That is its ability to predict accurately. Hence, the empirical performance of the identified distribution was verified. Data was simulated using the Weibull distribution. The simulated data from the Weibull distribution was compared with the remaining of the data (January, 2016 – January, 2017) using *t*-test. From the *t*-value = -41.189, *p*-value = 0.11. Hence, it can be observed that the two data have the same distribution. Three thousand data points were simulated from the Weibull distribution. The simulated data and the original data were plotted together as shown in Figure 8.

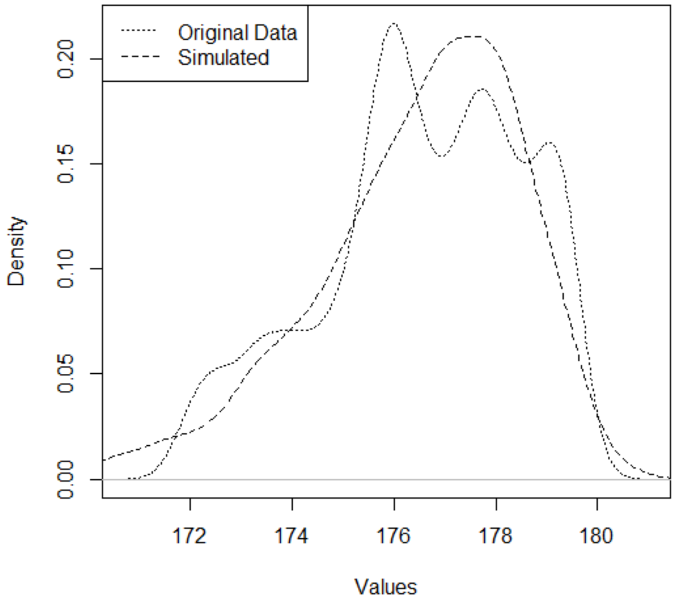

Figure 8: A graph of the distribution of real dam water levels and the simulated dam water levels.

The Wilcoxon signed rank test with continuity correction was conducted to test if the two datasets have the same distribution or not. From the test, it was observed that there was a *p*-value of 0.4132. Hence with a significance level of 0.05, it can be observed that the two datasets are identical.

CONCLUSION

From the analysis, it was observed that the value of the shape parameter for the Weibull distribution is greater than one. From the characteristics of the Weibull distribution, a shape parameter greater than one indicates that the failure rate of a component increases with time or a component is more likely to fail over time, Lai, PraMurthy, & Xie (2006). This implies that, the water level in the dam has been reducing with time. Hence the dam may become dry with time.

REFERENCES

Akaike, H., (1977). On entropy maximization principle. In: Krishnaiah, P.R. (Editor). Applications of Statistics, North-Holland, Amsterdam, pp. 27–41.

Alam S., Khan S. M. & Rahat S., H., (2015). A study on selection of probability distributions of extreme hydrologic parameters for the peripheral river system of Dhaka City. 15th World International conference, Bangkok, Thailand, pp. 29 – 34.

Alfarra A., Kemp-Benedict E., Hötzl H., Sader N., & Sonneveld B. (2012). Modeling Water Supply and Demand for Effective Water Management Allocation in the Jordan Valley. *Journal of the American Statistical Association*, 1(1), pp. 1-7, World Academic Publishing.

Bessa R.J., Miranda V., Sumaili J., Botterud A., Zhou Z., & Wang J., (2011). Wind Power Forecasting with Probability Density Estimation: A Tool for the Business, Windpower 2011 Conference and Exhibition, Anaheim, CA – USA.

Ding Y., Li S. & Li L., 2009. An analysis on chaos behavior of currency exchange rate undulation. First international workshop on education technology and computer science, Wuhan, Hubei, vol., 2, pp. 599-602. [doi: 10.1109.ETCS.2009.394].

Garba H., Ismail A., & Tsoho U., (2013). Fitting probability distribution functions to discharge variability of Kaduna River. International journal of modern engineering research (IJMER), 3(5), pp. 2848 – 2852.

Ghana Statistical Service (2012). 2010 Population & housing census. Sakoa Press Limited.

Ghosh, S. K., Chowdary, V. M., Saikrishnaveni, A, & Sharma R. K. (2016). A Probabilistic Nonlinear Model for Forecasting Daily Water Level in Reservoir, Water Resources Management, 30(9), pp. 3107–3122.

Gradojevic, N. and Yang, J., (2006). Non-linear, non-parametric, non-fundamental exchange rate forecasting. J. Forecast. No. 25, 227–245.

Huang, S.C., Chuang, P. J., & Wu, C.F., (2010). Chaos-based support vector regressions for exchange rate forecasting. Expert Systems with Applications 37 (12), 8590 –8598.

Ishak W., Hussain, W., Ku-Ruhanan K-M & Norita M. N., (2010). *Reservoir water level forecasting model using neural network.* International Journal of Computational Intelligence, 6 (4). pp. 947-952. ISSN 0973-1873

Izinyon O. C. & Ajumuka N. H., (2013). Probability distribution models for flood predictions in Upper Benue river basin-part II. Civil and environmental research 3(2), pp. 62 – 74.

Lai C. D., PraMurthy D. N., & Xie M., (2006). Weibull distributions and its applications. In Springer handbook of engineering statistics; Pham, H., ed.; Springer-Verlag: London, U. K., pp. 63 – 78.

Lin C-S, Chiu S-H & Lin T-Y, (2012). Empirical mode decomposition–based least squares support vector regression for foreign exchange rate forecasting. Economic Modelling 29, 2583 – 2590.

Nwobi-Okoye C. C., & Igboanugo (2013). Predicting water levels at Kainji dam using artificial neural networks. Nigeria Journal of Technology (NIJOTECH), PP. 129 – 136.

Radhwan A., Kamel M., Dahab M. Y., & Hassanien A, (2015). Forecasting exchange rates: A chaos-based regression approach. International Journal of Rough Sets and Data Analysis, 2(1), 38 – 57.

Rani S. & Parekh F., (2014). Predicting reservoir water level using artificial neural network. International journal of innovative research in science, engineering and technology (IJIRSET), 3(7), pp. 14489 – 14496.

Schwarz, G. E. (1978). Estimating the dimension of a model. Annals of Statistics 6 (2): 461–464. doi:10.1214/aos/1176344136. MR468014.

Valizadeh N., El-Shafie A., Mirzaei M., Galavi H., Mukhlisin M., & Jaafar O., (2014). Accuracy Enhancement for Forecasting Water Levels of Reservoirs and River Streams Using a Multiple-Input-Pattern Fuzzification Approach. The Scientific World Journal, doi:10.1155/2014/432976.

Vivekanandam N., (2015). Estimation of maximum flood discharge using Gamma and extreme value family of probability distributions. International journal of world research (IJWR), 1(16), pp. 16 – 23.