Journal of Environmental Science and Engineering B 6 (2017) 1-16
doi:10.17265/2162-5263/2017.01.001

Performance of Kohonen Self-organising Map in Statistical Analysis of Hydrologic Data

Edward Naabil[1], Henry Man'tieebe Kpamma[2], Faustina Gyabaah[3] and Asaa Akunai Abunkudgu[1]

1. Department of Agricultural Engineering, Bolgatanga Polytechnic, Bolgatanga 767, Ghana

2. Department of Statistics, Bolgatanga Polytechnic, Bolgatanga 767, Ghana

3. Department of Civil Engineering, Bolgatanga Polytechnic, Bolgatanga 767, Ghana

Abstract: The statistical characteristics of a hydrological data for the purposes of decision making in water resource planning and management is only justifiable if the data has the right attributes. This requires that the data being analysed are consistent, free of trend and being part of a stochastic process whose random characteristics is described by an appropriate distribution hypothesis. The data available for statistical analysis had a lot of missing values which could not be ordinarily filled but required a more comprehensive approach to fill these missing values. The KSOM (Kohonen Self-organising Map) was used to fill the missing runoff data from the Jidere-Bonde, Lokoja and Makundi river sites in the Niger basin. Results from the studies have shown that KSOM is the best tool for filling hydrological data with high number of missing values. After the data had been processed, some statistical applications were used to establish the runoff time-series characteristics of the three river sites of the Niger River basin. The results showed good attributes for all three river sites, except that Jidere River's data exhibited inconsistency. The presence of trend was also established for all three river sites; Jidere River was modelled based on 3-parameter lognormal, the other two river sites were modelled based on normal distribution probability. The presence of trend and other attributes require that a more stochastic modelling process be carried out. However, the results established give reference for water resource planning and management.

Key words: Kohonen Self-organising Map, water, resources, planning, management.

1. Introduction

Scientific data are classified as either quantitative or qualitative. Quantitative data is described as data in its numerical form, whereas qualitative data describes the nature of objects or the environment. In the case of runoff data, it is analysed as quantitative data. This form of data can be grouped into two kinds: experimental data and historical data. The experimental data are measured through experiments and can often be obtained repeatedly by experiments. Historical data on the other hand are a collection of data from natural phenomenon which is observed to only occur once and will not repeat. Most hydrologic data as in the case of runoff are regarded as historical data which are observed from natural hydrologic process [1].

Corresponding author: Dr. Edward Naabil, Ph.D., research field: hydrology.

Statistical analysis of hydrological time series data, at the time scales, usually applied in water resource planning are mostly based on some assumptions. These are: the data series is homogenous, stationary, free from trend and non-periodic with no persistence [2]. Homogeneity means that the data series belong to one population and therefore have a time-invariant mean. Non-homogeneity arises due to changes in the method of data collection and or the nature of the environment in which it is carried out [3]. Stationarity implies that there is no change in the statistical parameters of the series computed from different samples except as a result of sampling variations. A change in series can happen suddenly (step change) or gradually (trend) or may take complex dimensions [4]. A significant correlation between observation and time indicates the presence of trend in a data set. Trends in hydrological data are normally introduced through natural and artificial changes. Persistence is

the tendency for the magnitude of an event to rely on the magnitude of previous event(s), that is to say the data has no randomness and usually quantified in terms of serial correlation coefficient [5].

Time series analysis has been extensively studied in various fields such as geology, ocean technology, and seismology and has been applied in many hydrological and climatological situations. The purpose of time series analysis in hydrological data is to develop mathematical models to generate synthetic hydrological data, to forecast hydrological events, to detect shifts and trends in hydrological data and to fill missing data and extend records [6].

This study assessed the performance of KSOM, unsupervised neural networks, for predicting the missing values and for replacing outliers of the runoff data of the Niger River Basin. The outcome of the KSOM prediction of missing runoff data was statistically analysed to establish the statistical characteristics of runoff time series data.

2. Data and Methods

2.1 Available Data Used for the Analysis

The data for this work was made available by the Niger Basin Authority. These were daily runoff data in m^3/s for three river sites: Jidere-Bode, Lokoja and Makurdi. The reference data period was from 1980 to 2008. All three river sites had a lot of missing values. For instance for Jidere-Bode River with 347 data numbers, had 26 missing values (about 7.5%) of the total data number, Makurdi River also with a total data number of 347 had 117 missing values (about 33.7%) and for Lokoja river, the number of missing values is 48 (about 13.8%) of the total data number. These missing values could be attributed to a number of factors, such as malfunction of equipment, the absence of engineer to read the results, etc.. Considering the extent of missing values for the various sites, it requires that the missing values are filled (estimated) before proceeding to analysing the data.

2.2 Data Processing

The runoff data were in its raw description, in terms of daily values in m^3/s. A historical runoff period of (1980-2008) observation was considered. This was chosen because it gives a good representation of the hydrological data available for the catchment under study. The runoff data provided represented three stations (Jidere-Bode, Makurdi and Lokoja) as shown in Fig. 1. They are located within Nigeria.

The daily runoff data were summed into monthly and annual observation to carry out the time-series analysis. This is to give good analysis and a picture of the monthly and annual statistical characteristics of the Niger River basin. Daily observations do not give a real picture of characteristics of a river.

2.2.1 Missing Data and its Estimation

Missing values for these stations could be attributed to a sensor that does not deliver a measurement value or a fault with the instrument used or by human error [7]. Depending on what instrument is used, missing values are usually represented as blanks, zeros and negative values. In the data provided for this work, the missing values were blank. For extensive analysis of this work, it is required that the missing values be estimated. Failure to estimate the missing values makes the sample incomplete and difficult to use. There are quite a number of methods of estimating missing values. Examples are: simple linear regression [8, 9], back propagation ANNs (Artificial Neural Networks) modelling [10].

However, the use of the above mentioned methods for estimating missing values in a long time-series is difficult and often cannot be dependable particularly for situations where the number of values to be in-filled is relatively high in comparison with the total length of record [11].

For this work, the method of in-filling the missing value used is KSOM (Kohonen Self-organising Map). The KSOM is a robust multivariate model of data analysis, providing good estimation of missing values taking into consideration its relationship or correlation with other pertinent variables in the data record.

Fig. 1 Locations of Jidere-Bode, Lokoja and Makurdi Rivers.

2.2.2 Training the KSOM

The input data with its multidimensional nature is first standardized by deducting the mean and then dividing the result by the standard deviation. A random selection of a standardized input vector is presented to each of the individual neurons for comparison with their code vectors in order to come out with the code vector most similar to the presented input vector. The Euclidian distance is used for identifying the code vector, which is defined as;

$$D_i = \sqrt{\sum_{j=1}^{n}\left(x_j - w_{ij}\right)^2} \; ; \; i = 1, 2, \ldots\ldots, M \qquad (1)$$

Where D_i = Euclidian distance between the input vector and the weight vector i; $x_j = j^{th}$ element of the current input vector; $w_{ij} = j^{th}$ element of the weight vector i; and $n =$ dimensional of the input vector. The neuron whose vector most closely matches the input data vector (i.e., for which D_i is minimum) is chosen as a winning node or the BMU (Best Matching Unit). The weight vectors of this winning node and those of its adjacent neurons are then adjusted to match the input data using Eq. (2), drawing the weight vectors further into agreement with the input vector [12]:

$$w_i(t+1) = w_i(t) + \alpha(t) hc_i(t)[x(t) - w_i(t)] \quad (2)$$

Where t denotes time; $\alpha(t) =$ learning rate at t; $hc_i(t) =$ neighbourhood function placed in the winner unit $c$at time t; the other variables remain the same as defined previously. This makes each node in the map internally, to develop the ability to recognise input vectors similar to it. This is referred to as self-organizing, because no other information is provided to lead to a classification [13]. The process of comparison and adjustment continues until the suitable number of iterations is reached or the specified errors are attained. Both the learning rate and the neighbourhood function need to be carefully chosen since it affects the learning effectiveness of the

KSOM. This is particularly the case for the learning rate; it decreases monotonically with increased number of iterations from Eq. (3) [12].

$$\alpha(t) = \alpha_0(0.005/\alpha_0)^{t/T} \quad (3)$$

Where $\alpha_0 =$ initial learning rate and $T =$ training length, thereby forcing the weight vector to converge. The neighbourhood function is normally chosen to be Gaussian, centred in the winner unit c, such that:

$$h_{ci}(t) = \exp\{-||r_c - r_i||^2 / [2\sigma^2(t)]\} \quad (4)$$

Where r_c and r_i are positions of nodes *c* and *i* on the KSOM grid and $\sigma(t)$ is the neighborhood radius. Like the learning rate $\alpha(t)$, $\sigma(t)$ also decreases monotonically as the number of iterations increases.

The quality of the trained KSOM is measured by the total average quantization error and total topographic error. The quantization error is

$$q_e = \frac{1}{n}\sum_{i=1}^{N} ||X_i - W_c|| \quad (5)$$

Where $q_e =$ quantization error; $X_i = i^{th}$ data sample or vector; $W_c =$ prototype vector of the best matching unit for X_i; and denotesthe Euclidian distance. The topographic error is

$$t_e = \frac{1}{N}\sum_{i=1}^{N} u(X_i) \quad (6)$$

where $u_i =$ binary integer such that it is equal to 1 if the first and second best matching units for X_i are not adjacent units; otherwise it becomes zero.

The KSOM was described above and its approach was used to estimate or predict the missing values for the data provided. This analysis was carried out using a computer tool called SOM (Self-organising Map) part of the MATLAB 12 (Matrix Laboratory, version 12) software. The first step was to train the model using the available data set. The depleted vector set was presented to the KSOM to identify its BMU. The missing values were obtained as their corresponding values in the BMU (see Fig. 2).

2.2.3 Consistency Test

Consistency is defined as collected data belonging to the same statistical population. This was used to evaluate the runoff data. The double mass curve approach was used. This procedure is based on the comparison of cumulative values of two data sets in a diagram form, such that one of the data sets being consistent, while the other is taken as the suspect. The plot of the double mass diagram should show a linear relationship if the suspected data set is consistent; otherwise there will be a departure from linearity as shown in Fig. 4(a). The mean data for some sites is not likely to be affected by the changes at one of the component sites, it is therefore right to use the mean

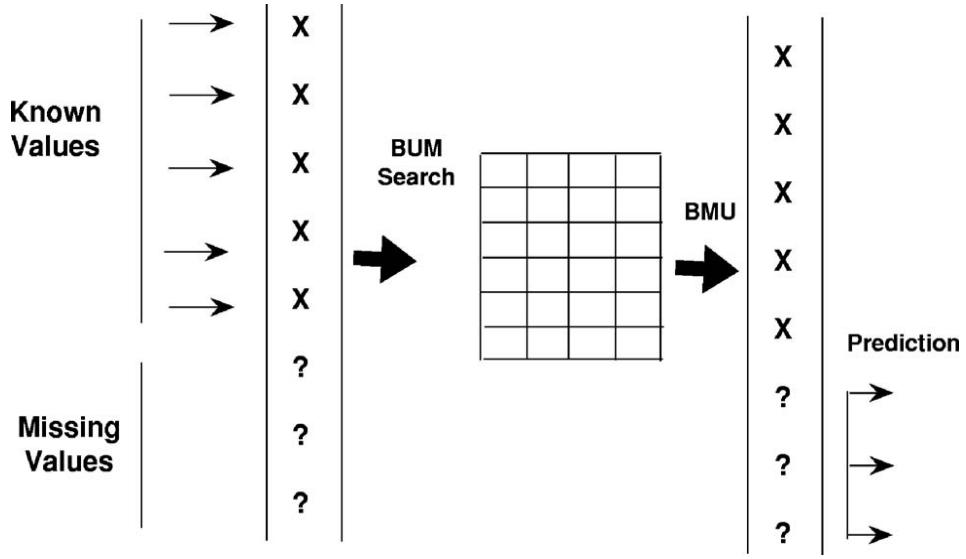

Fig. 2 Prediction of missing components of the input vector using the KSOM (BMU).

data as the reliable data or consistent data to check. Another approach is to use a long-term data station in an area which is not known to have human influences as the consistent set.

2.2.4 Trend Test

A data set with trend is one that has significant correlation (positive or negative) between the observations and time.

The common method of detecting trend is the linear regression between the data values and time of the form [14].

$$Q(t) = a + b(t) + \varepsilon(t) \quad (7)$$

Where $Q(t), t = 1, 2, \ldots, n$ is the data value at time t. Also a and b are the regression coefficients, and $\varepsilon(t)$ is a random error (white noise) with a mean of zero and variance of S_V^2. For a trend to exist in the dataset the regression estimate of the slope parameter will be statistically different from zero; if this is not the case, it will be difficult to justify the existence of trend. The shortfall of this approach is that it does not distinguish between trend and persistence [14].

The trend detection approached used to investigate the existence of or otherwise of long-term trend in this work was the SROC (Spearman Rank Order Correlation) nonparametric test. However there are other nonparametric trend detection tests available, e.g. the Mann-Kendall test. Among the mentioned trend detection approach, the SROC is the preferred approach recommended by the World Meteorological Organisation for trend detection in flow volumes [15]. The presentation on SROC [16] was adapted for this work and the methodology is illustrated:

Let data series $Q(t), t = 1, 2, \ldots, n$ be observed in time, t.

Ranks, R_{qt} are assigned to $Q(t)$, such that the largest $Q(t)$ has R_{qt}= 1 and the least $Q(t)$ has a rank, $R_{qt} = n$. Where there are ties in the $Q(t)$, each of the ties are assigned a rank equal to the mean of the ranks that would have been used had there been no ties. The difference $d_i = R_{qi} - i$ is computed. The coefficient of the trend r_s is computed as:

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2 - 1)} \quad (8)$$

The null hypothesis (H_0) that the time series has no trend, it can be shown that the test statistic:

$$t = r_s \sqrt{\frac{n-2}{1-r_s^2}} \quad (9)$$

Has a student's t-distribution with n-2 degrees of freedom. The critical values of the *t*-distribution are obtained for the chosen significance level, α and n-2 degrees of freedom. For a two-tailed test, these critical values are denoted by $\pm t_{\alpha/2,\, n-2}$.

The values of t obtained in equation…. are compared with the critical values. The null hypothesis, H_0 is rejected if $t > t_{\alpha/2,n-2}$ or $t < -t_{\alpha/2,\, n-2}$.

2.2.5 Probability Distribution Goodness-of-Fit Tests

The application of probability distribution goodness-of-fit test involves modelling univariate data with specific probability distribution. The steps to take into account in establishing the form of the probability distribution is determining the "best-fitting" distribution and estimation of the parameters (shape, location and scale parameters) for that distribution.

Quite a number of commonly used formal methodologies of testing the goodness of fit of time series data to describe theoretical probability distributions are available. Examples of these methods are chi-squared test, the Kolmogorov-Smirnov test, the PPCC (Probability Plot Correlation Coefficient) test and the moment (1-moments and p-moments) ratio diagrams test [17].

It is also possible to informally infer the form of the probability distribution by plotting the histogram of the data. The shape of the histogram could either be symmetrical, showing the presence of normal distribution or asymmetrical, indicating a nonzero skew, which is not possible to model with the normal distribution.

Apart from this approach, a simple statistical test can be carried out to determine whether or not the

sample skew coefficient estimated is statistically zero. These two simple approaches were used in this work. The sample skew was estimated and compared with the approximation 95% confidence limits for zero skew, i.e..

(-1.96S_{gy}, 1.96S_{gy}), where S_{gy} is the standard error of estimate of the sample skewness coefficient, g_y, given as:

$$g_y = n\sum_{i=1}^{n}\frac{(y_i-\mu_y)^3}{(n-1)(n-2)\sigma_y^3} \quad (10)$$

Where μ_y and σ_y are respectively the mean and the standard deviation of y, and S_{gy} is approximately equal to √(6/n), where n is the sample size. If the skewness coefficient is less within the 95% confidence limits, then the null hypothesis that the skewness is zero is not rejected; or else the data will be assumed to have a nonzero skewness coefficient.

The formal probability goodness-of-fit method used in this study is the PPCC test mentioned above. This method is easy to apply and has enough power to discriminate between different distribution hypotheses [17]. The test is based on the correlation coefficient between the ordered sample $q_1 \le q_2 \le q_3 \le \cdots \le q_n$ and their associated fitted quantiles, $w_i = G^{-1}(1-p_i)$ where $p_i, i = 1,2,3,\ldots,n$ is the exceedence probability of y_i and $G^{-1}(.)$ the inverse of the cumulative distribution function (cdf) of the distribution being considered. Cunnane provides the best estimate of p_i as [18]:

$$p_i = \frac{R_{yi}-0.4}{n+0.2} \quad (11)$$

Where R_{yi} is the rank of y_i for the ordered sample, i.e. $R_{y1}=1, R_{y2}=2, \ldots R_{yn}=n$. The estimated correlation coefficient between y_i and w_i is compared with the critical points of the PPCC for the particular distribution. The critical points for normal, lognormal and Gumbel probability distribution has been provided by Vogel, R. M. [19]. Vogel, R. M. and McMartin, D. E. [20] also provided critical values for gamma, Pearson type-3 (P3) and LP3 (Log-Pearson type-3) distributions. However, from the skew coefficients results for this study, the normal, lognormal, Pearson type-3 and 3-parametric lognormal distributions were applied for the PPCC test.

The parameter $G^{-1}(.)$ depends on the distribution hypothesis being tested. For the three-parameter lognormal distribution:

$$W_{iLN3} = G^{-1}(1-p_i) = v + exp^{(\mu_x + Z_{pi}\sigma_x)} \quad (12)$$

Where $log_e(y-v), v$ is the lower limit of y, μ_x and σ_x are respectively the mean and standard deviation of x and Z_{pi} is the standardised normal variate at p_i. The determination of Z_{pi} given p_i by Stedinger, J. R. [17] is given as:

$$z_{pi} = \frac{(1-p_i)^{0.135} - (p_i)^{0.135}}{0.1975} \quad (13)$$

Required algorithms for obtaining μ_x, σ_x, v are provided by Stedinger, J. R. [17]. The two- parameter lognormal distribution is a special case of the lognormal distribution in which the lower limit $v = 0$. Whereas the three-parameter lognormal distribution can be used to model any positive skewness, the same is not for the two-parameter, it is strictly limited to

$$g_y = CV^3 + 3CV \quad (14)$$

where $CV = {}^{\sigma_x}/_{\mu_x}$

Since the skew results for two stations were negative, the other parameters such as lognormal, and 3-parameter lognormal could not be applied as discussed above. Therefore, Pearson-type 3 parameter distribution was used for the analysis of those skew results.

Nguyen, T. V. and In-Na, N. [21] introduced a plotting formula which in a way provides unbiased flood estimates for the P3 distribution for systematic flood records. The formula used in his studies was also applied in this work. The formula is expresses as:

$$p_m = \frac{N-m+0.3\gamma+0.47}{N+0.3\gamma+0.05} \quad (15)$$

Where P_m represents the non-exceedence probability. The P3 formula can also be written in terms of the exceedence probability as:

$$P_m^{'} = 1 - P_m = \frac{m - 0.42}{N + 0.3\gamma + 0.05} \quad (16)$$

The formula above takes explicit account of the skewness coefficient of the parent distribution. Where skew coefficient is given as:

$$\gamma_m = \frac{g_{y(1-\rho_1^3)}}{(1-\rho_1^2)^{1.5}} \quad (17)$$

The plotting position for P3 distribution given by Nguyen, T. V. and In-Na, N. [21] is given as:

$$\rho_m = \begin{cases} \frac{m - 0.42}{k + 0.3\gamma + 0.05} \cdot \frac{k}{N}, m = 1, \dots\ k \\ \frac{k}{N} + \frac{N-k}{N} \cdot \frac{m-k-0.42}{s-e+0.3\gamma+0.05}, m = k+1, \dots\dots, g \end{cases} \quad (18)$$

3. Results and Discussion

3.1 KSOM and Missing Values

This section provides the performance of the KSOM in predicting the missing data as part of the pre-processing of the data. Figs. 3(a), (b) and (c) indicates the relationship between the KSOM predicted runoff with observe runoff for the rivers, Jidere-bode, Lokoja and Makurdi respectively. From these figures there is an indication of the ability of KSOM to predict missing hydrological data (in this case runoff) in respect to observe runoff data. The strength and reliability of this tool (KSOM) in filling in missing hydrological data was subjected to some statistical test which has been discussed in the next sections. However, in the face of the plots of KSOM predicted runoff with observe runoff; the results showed that KSOM is a good tool for estimating and predicting missing values, which have also been confirmed in other studies [7].

3.2 Consistency Test

The double mass curve for the Niger River basin, which constitutes three stations (Jidere-Bode, Lokoja and Makurdi) is shown in Fig. 4. From Fig. 4, the double mass curve shows a reasonable degree of consistency, with no clear change in slope for Lokoja and Makurdi stations except for Jidere-bode station which shows some inconsistency from the year 2002 to 2008. The inconsistency for Jidere-bode station can be adjusted by considering the deviation point

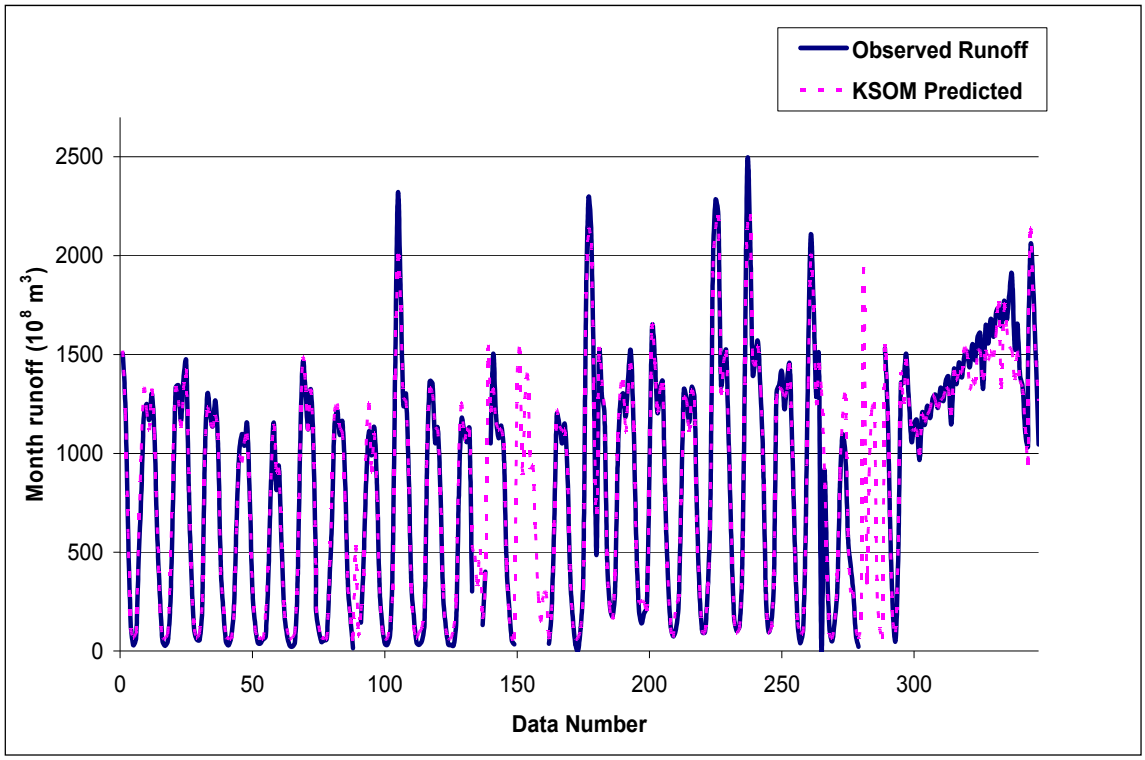

Fig. 3 (a) Runoff time-series for Jidere-bode.

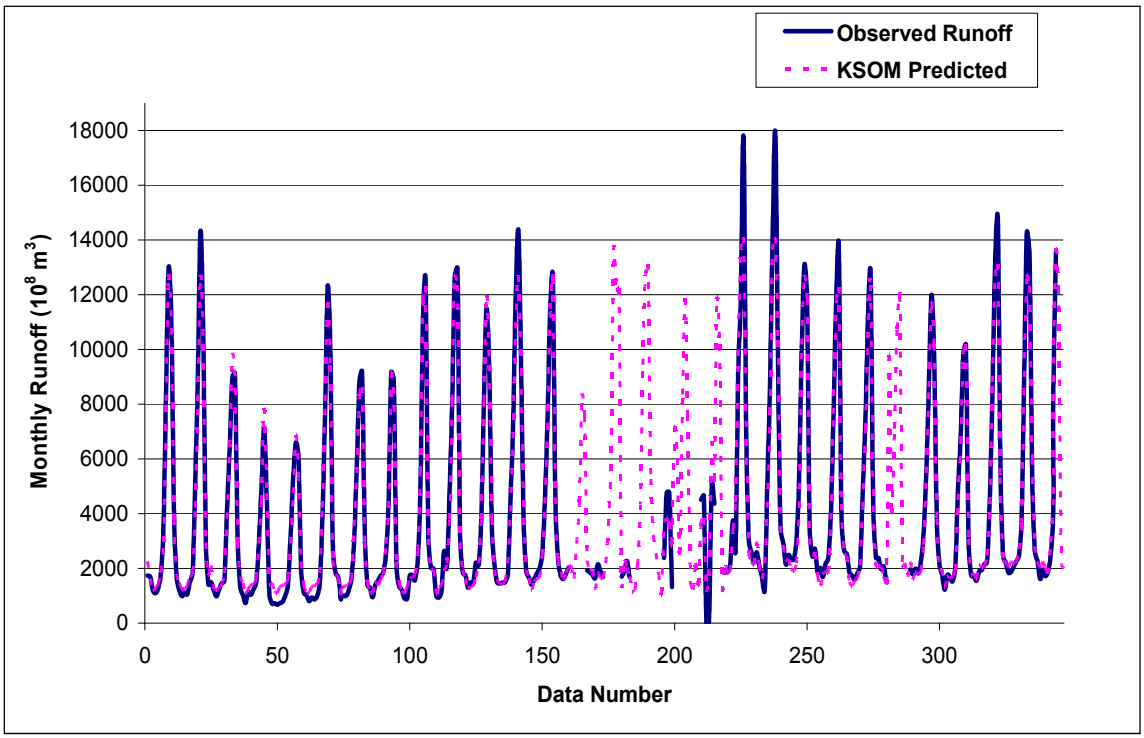

Fig. 3 (b) Runoff time-series for Lokoja.

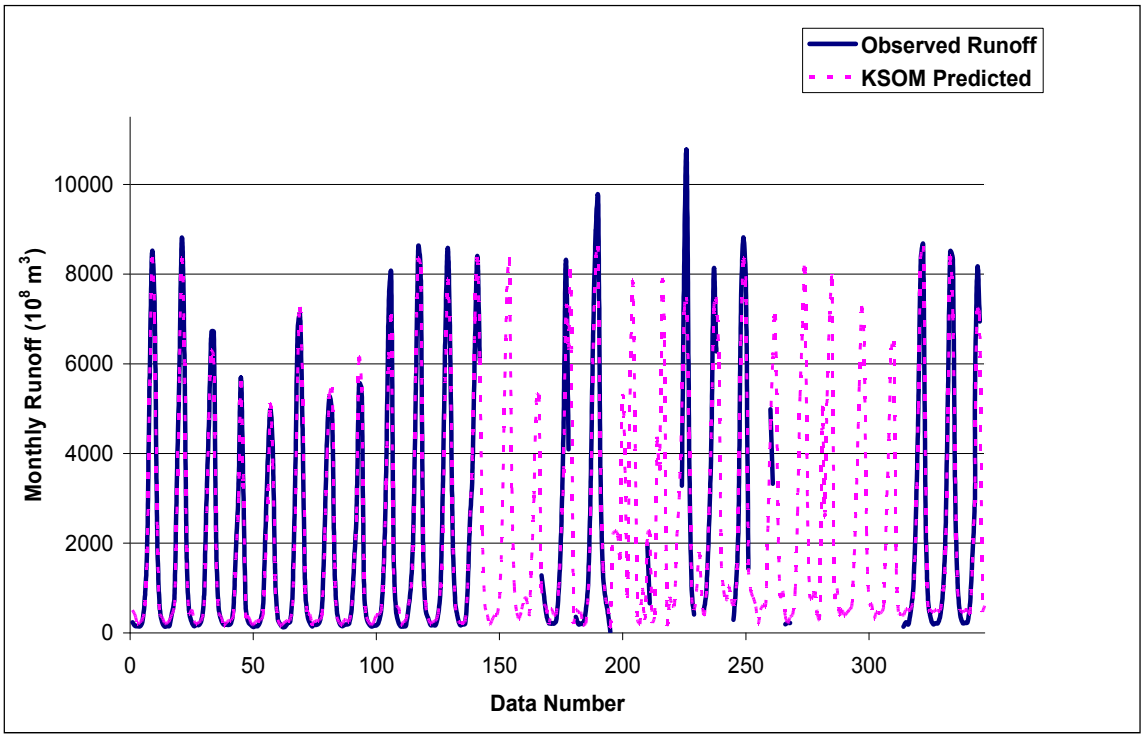

Fig. 3 (c) Runoff time-series for Makurdi.

(2002-2008) as a steeper slope (S_1). The adjustment depends on which section of the data is thought to be reliable. From Fig. 4, the pre-T record is more reliable. In that case the post-T records can be corrected for the deviation, S_1 by multiplying (S_1/S_0), where S_0 is the slope if no inconsistency.

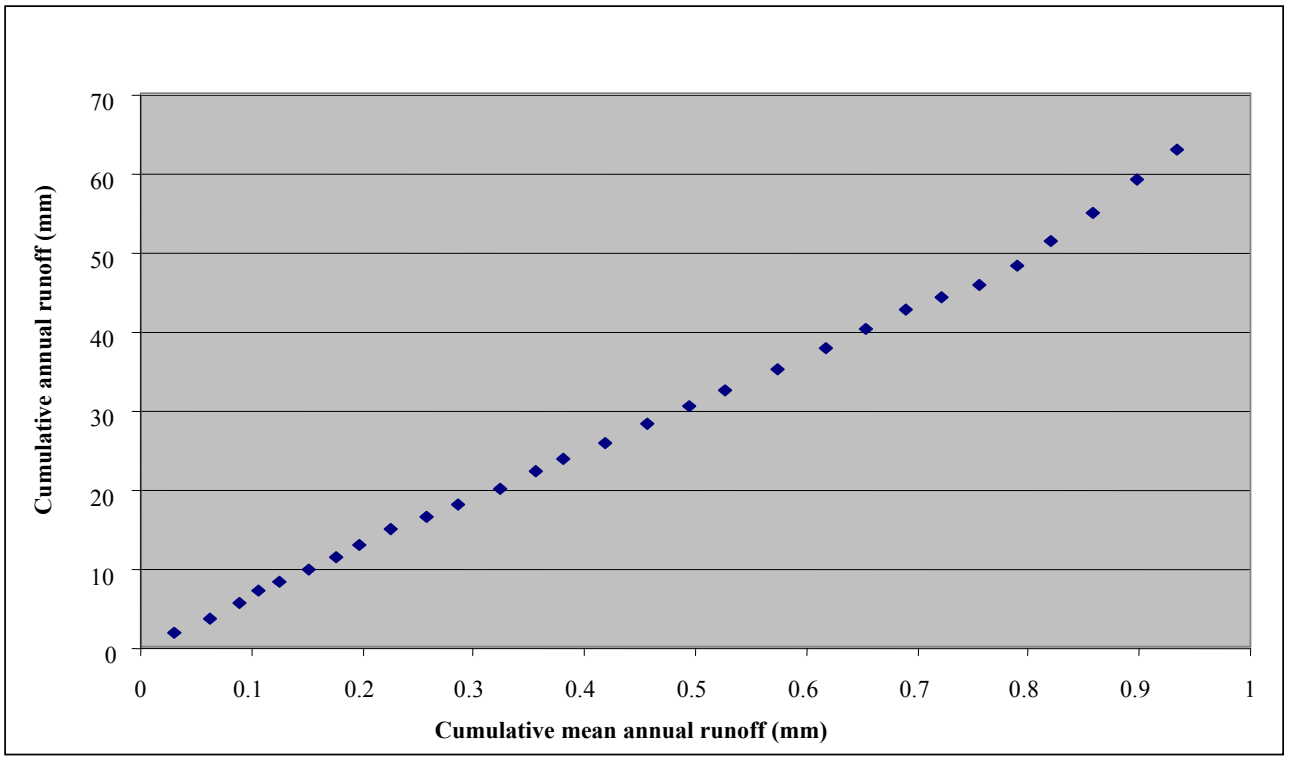

Fig. 4 (a) Consistency test of annual runoff of Jidere bode.

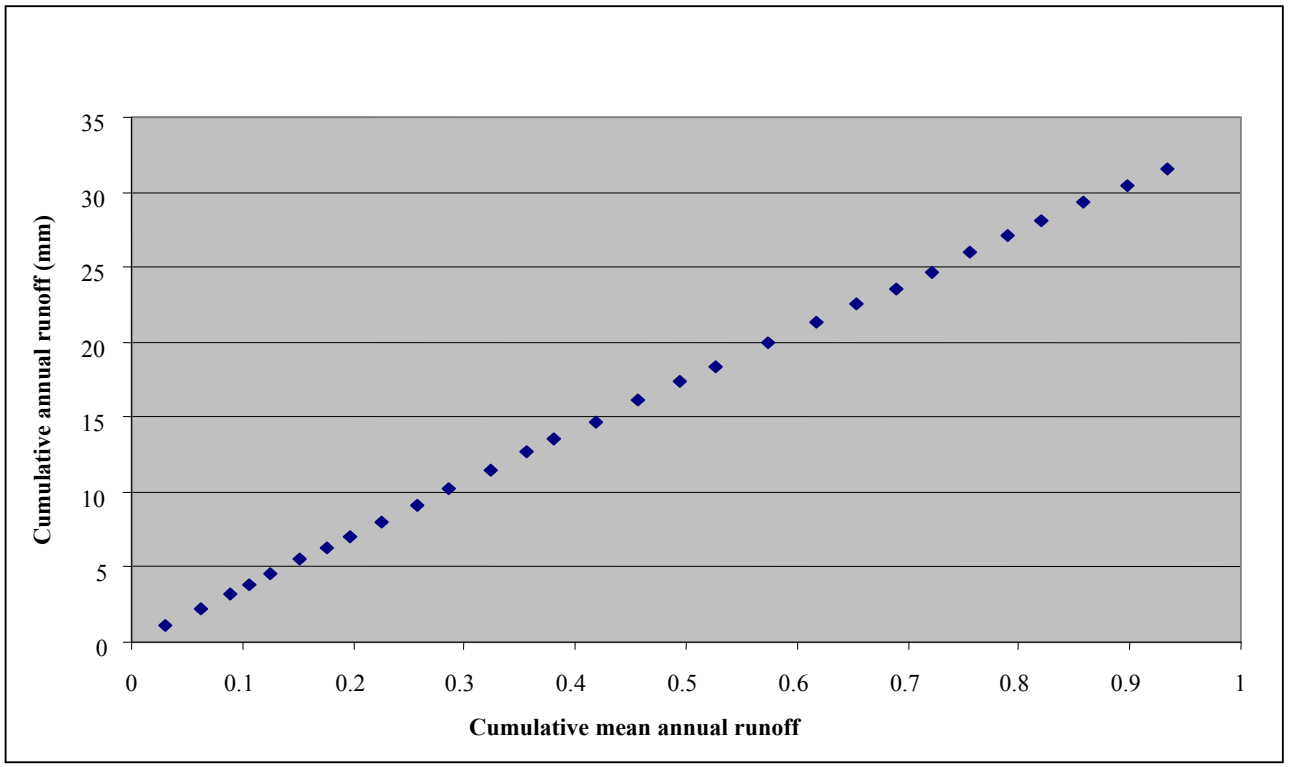

Fig. 4 (b) Consistency test of annual runoff of Makurdi.

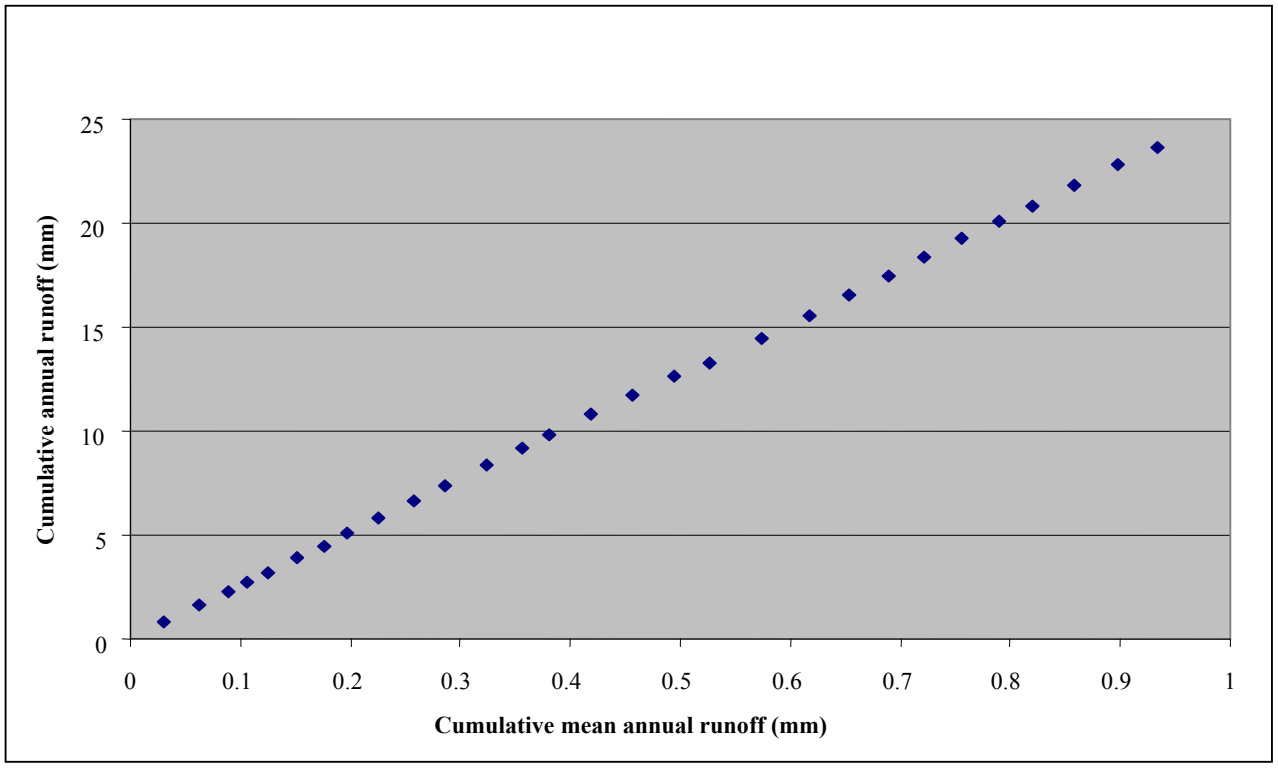

Fig. 4 (c) Consistency test of annual runoff of Lokoja.

3.3 Trend

The results of the SROC test statistic for the annual data at each of the stations are compared with the critical points of the *t*-distribution at 5% significance level in Table 1. From these results, the null hypothesis that the time series has no trend is rejected. This indicates the presence of trend and will require either a stochastic modelling of the series or much simpler approach would be to remove the trend from the data series. This is carried out with the linear equation, $Y_i = a + b.i + v_i$. A more comprehensive approach of trend analysis where serial dependence is also significant is presented by Hameed, T., Marino, M. A., DeVries, J. J. and Tracy, J. C. [14].

From Fig. 5, the results indicate the presence of trend which increases with time. This confirms the results from the trend analysis. It also shows the pattern of runoff distributions of the catchments. Some years have high runoffs and others low runoffs which indicate periods of high floods and drought of the river sites.

3.4 PPCC Goodness-of-Fit Probability Distribution Test

The Table 2 presents the PPCC (Probability Plot Correlation Coefficients) for the annual runoff at Jidere, Lokoja and Makurdi. All the correlation coefficients are positive which is expected, since one expects the observed quantiles to increase as the probability-distribution-predicted quantiles increase. From the results in the Table 2 the highest annual correlation coefficients is with the three (3)-parametric lognormal distribution. This affirms the results from the skewness coefficient and the histogram plot whereas the highest correlation coefficient for the other two stations (Lokoja and Makurdi) is with the normal distribution. This is not surprising based on the results from the histogram plots for both stations. This supports the results from the normality test and also from the skew test result.

For the monthly runoffs in Table 2, the lowest correlation coefficient is associated with lognormal distribution which confirms the results of low monthly skewness estimates [22]. Tables 3 and 4 have their

Table 1 Results of the SROC and run tests for trend and randomness respectively for three river sites.

Station	Trend: Test statistic	Critical value at 5% level (*t*-distribution)	Randomness: Test statistic	Critical value at 5% level (normal distribution)
Jidere-Bode	4.22	2.052	-1.93	-1.96
Lokoja	4.10	2.052	-3.10	-1.96
Makurdi	2.77	2.052	-1.54	-1.96

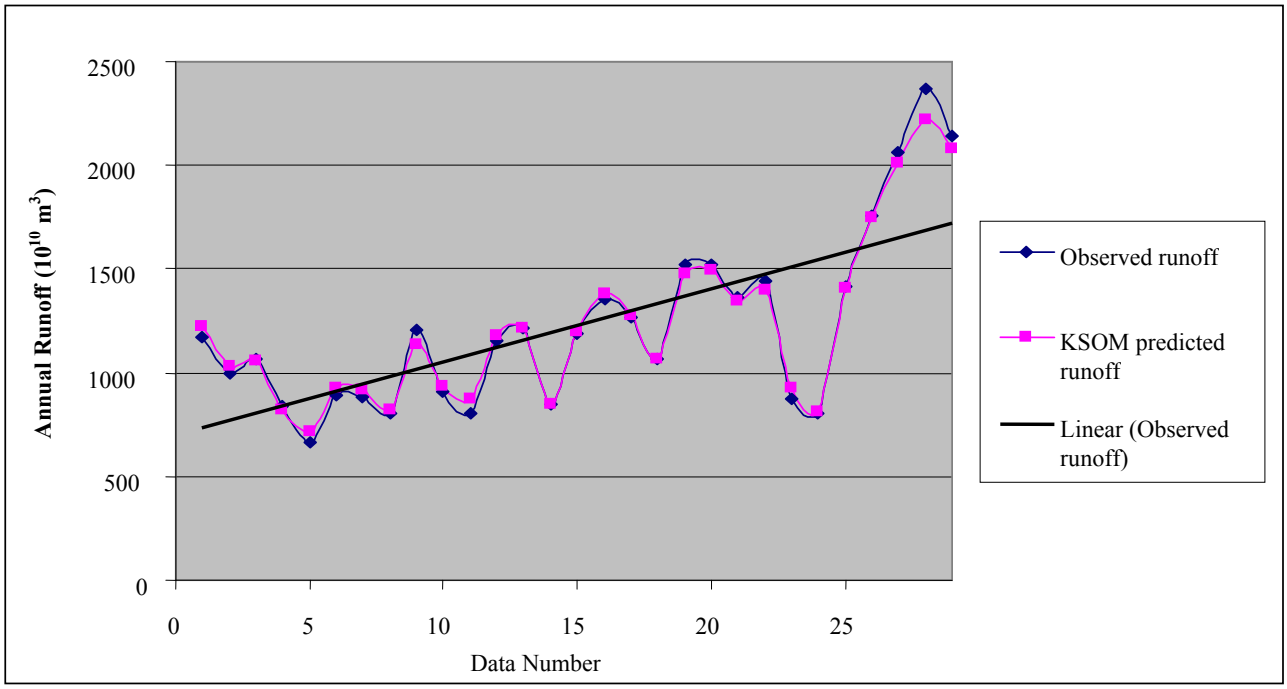

Fig. 5 (a) Annual Runoff time series also showing the KSOM predicted runoff for Jidere-Bode.

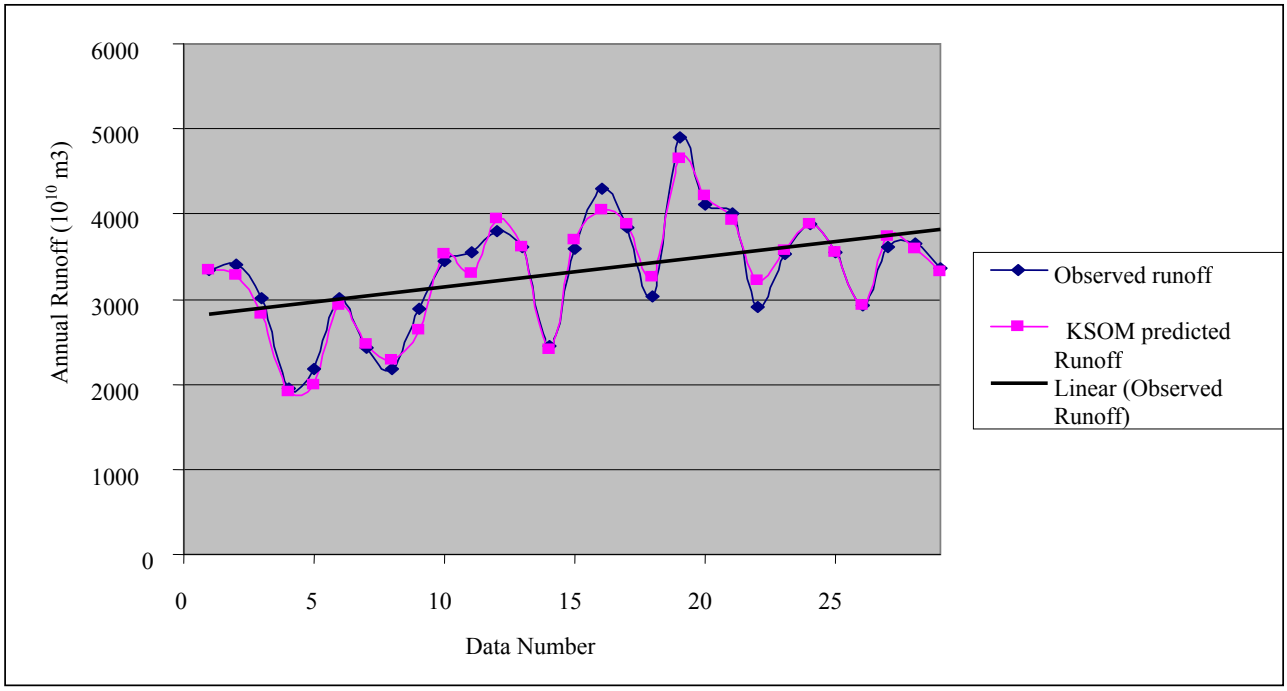

Fig. 5 (b) Annual Runoff time series also showing the KSOM predicted runoff for Makurdi.

lowest correlation coefficients associated to the normal distribution, which is not surprising given the high monthly skew estimates [22]. This also indicates that different months have their own "best" distribution on the basis of the maximum correlation coefficient of the PPCC test, which implies that, any choice of a stochastic modelling of the monthly runoffs would require a complicated mixing of distributions. To carry out a particular model it is important to device a criterion by which a single probability function could be chosen for use across all twelve months runoffs, as this would turn to simplify considerably the subsequent stochastic modelling of the runoffs. The approach adopted was the highest score out of the total number of occasions in which a given distribution performed better than the others. The total score for each of the probability distribution functions has been shown in Tables 2, 3 and 4. LN3 (Log Normal 3) produced the highest score for Jidere-Bode, whereas Normal produced the highest score for Lokoja and Makurdi sites. Therefore, Lokoja and Makurdi sites where modelled using the normal distribution, whereas Jidere-bode was consequently modelled with 3-parameter lognormal distribution.

Table 2 Correlation coefficients obtained from the PPCC test for historical runoff data in Jidere-Bode site (maximum PPCC is underlined).

Month	Normal	LN	LN3	P3
January	0.979	0.918	0.978	0.842
February	0.969	0.960	0.974	0.774
March	0.900	0.978	0.985	0.709
April	0.796	0.975	0.983	0.609
May	0.778	0.951	0.952	0.643
June	0.814	0.960	0.986	0.676
July	0.905	0.979	0.980	0.742
August	0.988	0.985	0.993	0.797
September	0.941	0.971	0.977	0.980
October	0.950	0.853	0.950	0.946
November	0.919	0.718	0.918	0.875
December	0.920	0.719	0.918	0.910
Total Score	3	0	7	1
Annual	0.946	0.982	0.992	0.980

Table 3 Correlation coefficients obtained from the PPCC test for historical runoff data in Lokoja site (maximum PPCC is underlined).

Month	Normal	P3
January	0.730	0.748
February	0.989	0.636
March	0.983	0.602
April	0.989	0.669
May	0.692	0.717
June	0.929	0.776
July	0.982	0.843
August	0.963	0.910
September	0.914	0.979
October	0.980	0.945
November	0.904	0.879
December	0.698	0.814
Total Score	8	4
Annual	0.986	0.979

Table 4 Correlation coefficients obtained from the PPCC test for historical runoff data in Makurdi site (maximum PPCC is underlined).

Month	Normal	P3
January	0.634	0.815
February	0.811	0.610
March	0.752	0.647
April	0.773	0.680
May	0.635	0.783
June	0.914	0.844
July	0.959	0.741
August	0.965	0.910
September	0.911	0.980
October	0.977	0.945
November	0.915	0.710
December	0.640	0.881
Total Score	8	4
Annual	0.984	0.980

The probability plots of the annual runoff data for the three stations are shown in Fig. 6. The probability axis in both plots is based on normal distribution. The upper and lower boundary limits of the confidence interval at 5% level of significance are also included in the plots. The normal distribution for Figs. 6(b) and (c) cast doubts on the PPCC results, since the normal distribution plot will usually lie approximately straight on the line, however it is not the case here; the normal distribution plots shows some degree of curvature which is expected in some cases [23].

4. Conclusions

In pre-processing the data, it is imperative that missing values were filled to give good and justifiable results. There are quite a number of pre-processing

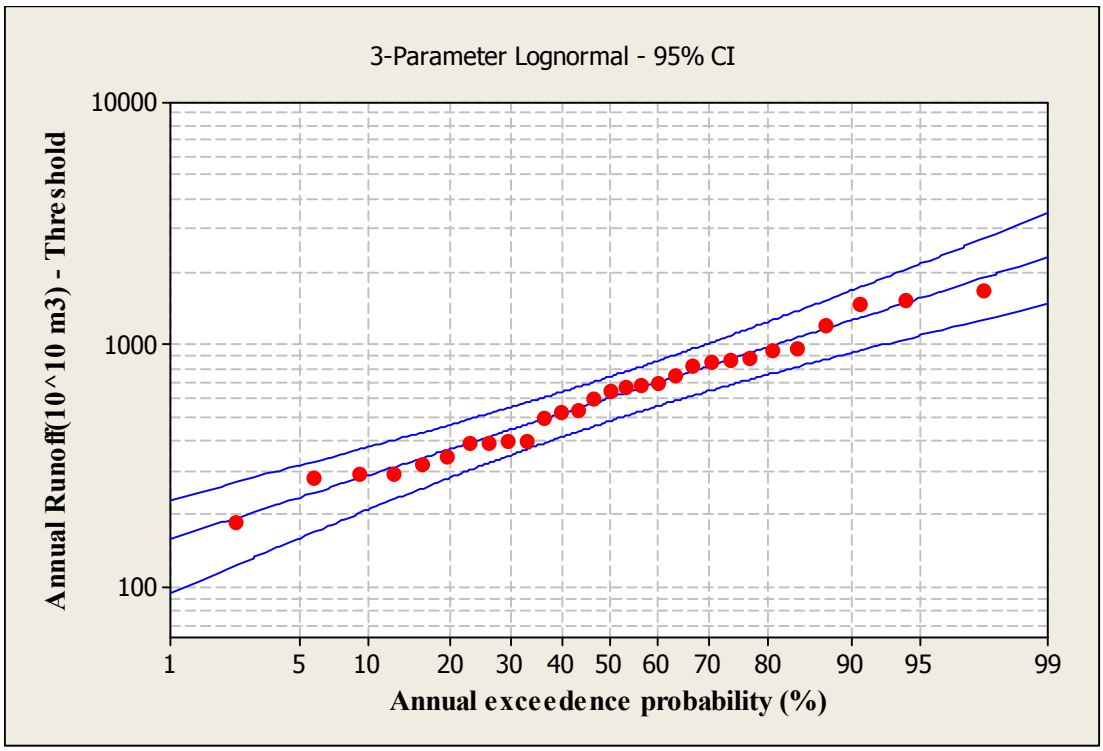

Fig. 6 (a) 3-Paramter lognormal probability plot of annual runoff data for Jidere-Bode (CI = Confidence interval).

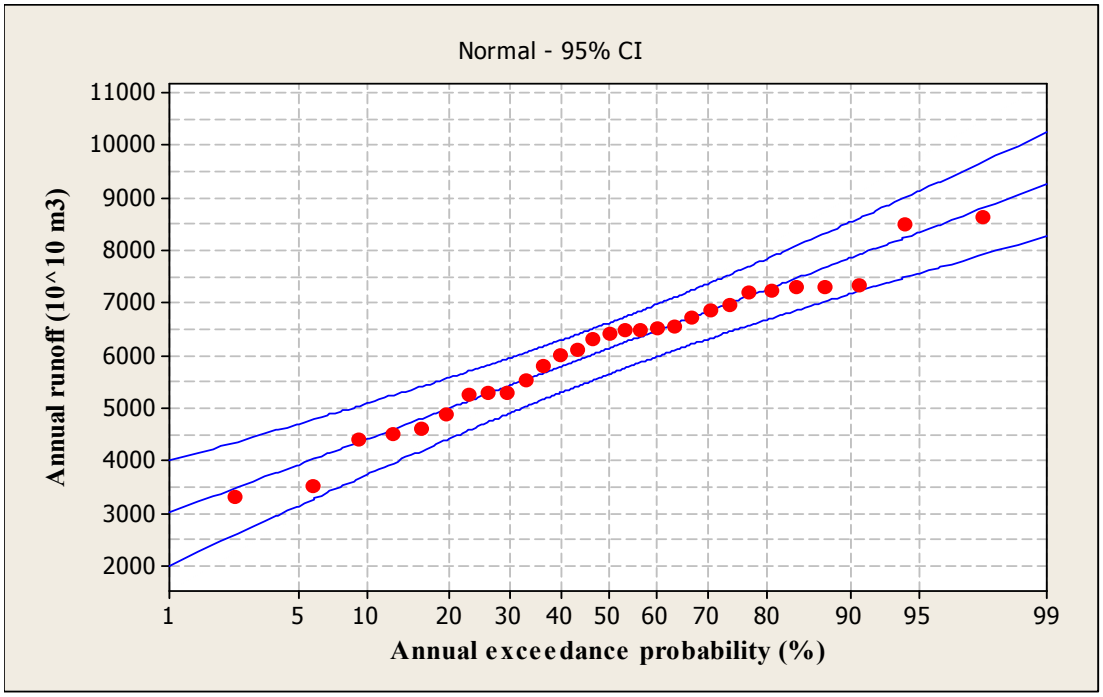

Fig. 6 (b) Normal probability plots of annual runoff data for Lokoja (CI = Confidence interval).

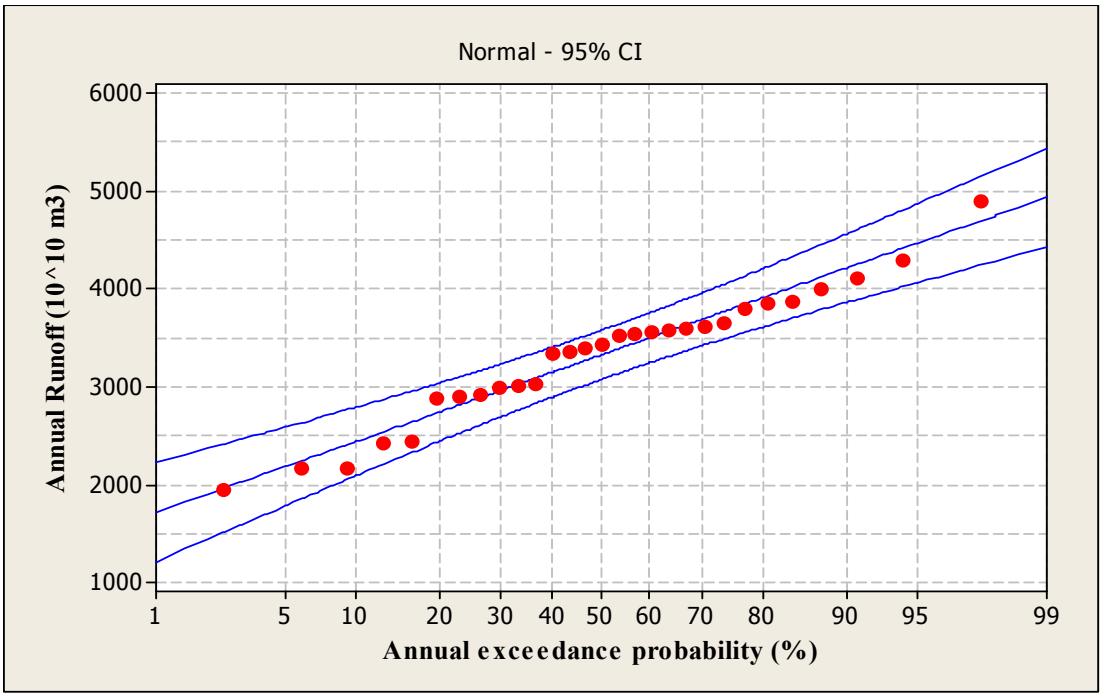

Fig. 6 (c) Normal probability plots of annual runoff data for Makurdi (CI = Confidence interval).

tools or techniques. The common one is data record augmentation and extension. The approached used in filling the missing values in this study is the KSOM tool. The results from KSOM has shown that it is the appropriate tool to fill in the missing values as there were no sharp statistical difference from that of the observations (historical data).

This gave credence to the runoff data for the three (3) sites of the Niger River to be statistically analysed to serve as the bases for decision to be taken for the

purposes of water resources planning and management. The statistical test carried out were the consistency of the data set, the presence of a long-term trend in the data set or not and the selection of the most appropriate probability distribution function for modelling the data. From the results, all three sites demonstrated the right attributes, except in the case of consistency test, in which Jidere-bode site exhibited some inconsistencies. Also in the case of normality test, Lokoja and Makurdi had their P-values greater than 0.10 which indicates normalness. However, the normality curve did not indicate the presence of normalness but this did not warrant the rejection of the hypothesis for the two sites [23].

The PPCC goodness-of-test also exhibited sufficient power to discriminate between individual probability distribution functions for modelling the data sets. However, based on the annual skew results, not all probability distribution functions could be used to model all the sites. For instance, Lokoja and Makurdi from the annual summary statistics have negative skewness, therefore some probability distributions like Lognormal, three-parameter lognormal and gamma cannot be modelled with these sites. However, the results from the studies give reference for a more stochastic analysis to be carried out.

Acknowledgement

This research would not have been made possible without the support of Niger Basin Authority. They provided the historical runoff data for the studies. Authors are deeply grateful to them. The contribution and critic of Dr. A. Adeloye towards this work cannot beover looked. It was actually through the effort of Dr. A. Adeloye we had access to the historical data. Authors do express their gratitude to him.

References

[1] Yevjevich, V. 1964. "Statistical and Probability Analysis of Hydrological Data." *Handbook of Applied Hydrology* Section 8-I: 9-50.

[2] Machiwal, D., and Madan K. J. 2008. "Comparative Evaluation of Statistical Tests for Time Series Analysis: Application to Hydrological Time Series." *Hydrological Sciences Journal* 53 (2): 353-66.

[3] Fernando, D. A. K., and Jayawarden, A. W. 1994. "Generation and Forecasting of Monsoon Rainfall Data." *Proc. of the 20th WEDC Conference. Colombo, Sri Lanka*: 310-3.

[4] Shahin, M., Van Oorschot, H. J. L., and De Lange, S. J. 1993. *Statistical Analysis in Water Resources Engineering*. AA Balkema.

[5] McMahon, T. A., and Mein, R. G. 1986. *River and Reservoir Yield.* Colorado: Water Resource Publication.

[6] Salas, J. D. 1993. "Analysis and Modelling of Hydrologic Time Series." In *Handbook of Applied Hydrology* 19: 1-72, edited by Maidment, D. R..

[7] Rabee, R., and Adebayo, J. A. 2007. "Replacing Outliers and Missing Values from Activated Sludge Data Using Kohonen Self-Organising Map." *Journal of Environmental Engineering* 133 (9): 909-16. doi:10.1061/(ASCE) 0733-9372.

[8] MacDonald, I. L., and Zucchini, W. 1997. *Hidden Markov and Other Models for Discrete Valued Time Series.* Vol. 110. CRC Press.

[9] Harvey, C. R. 1989. "Time-Varying Conditional Covariances in Tests of Asset Pricing Models." *Journal of Financial Economics* 24 (2): 289-317.

[10] Maier, H. R., and Dandy, G. C. 1996. "The Use of Artificial Neural Networks for Prediction of Water Quality Parameters." *Water Resour Res* 32 (4): 1013-22.

[11] Rosen, C., and Lennox, J. A. 2001. "Multivariate and Multi-scale Monitoring of Wastewater Treatment Operation." *Water Res* 35 (14): 3402-10.

[12] Vesanto, J., Himberge, J., Alhoniemi, E., and Parhankangas, J. 2000. "Self-Organizing Map (SOM) Toolbox for Matlab 5." *Helsinki Univ. of Technology, Finland*. Rep. No. A57.

[13] Penn, B. S. 2005. "Using Self-Organising Maps to Visualize Highdimensional Data." *Computer & Geoscience* 31 (5): 531-44.

[14] Hameed, T., Marino, M. A., DeVries, J. J., and Tracy, J. C. 1998. "Method for Trend Detection in Climatological Variables." *Journal Hydrologic Engineering* 2 (4): 154-60.

[15] WMO. 1988. "Analysing Long Time Series of Hydrological Data with respect to Climate Variability." WCAP-3 (World Climate Application Programme), WMO/TD No.224, World Meteorological Organisation, Geneva, Switzerland.

[16] McGhee, J. W. 1985. *Introduction Statistics.* New York: West Publishing Co..

[17] Stedinger, J. R. 1993. "Frequency Analysis of Extreme Events." *Handbook of Applied Hydrology* 18, edited by D.

R. Maidment. McGraw-Hill, New York, USA.

[18] Cunnane, C. 1978. "Unbaised Plotting Positions-a Review." *Journal of Hydrology* 37 (3-4): 205-22.

[19] Vogel, R. M. 1986. "The Probability Plot Correlation Coefficient Test for the Normal, Log-normal, and Gumbel Distribution Hypothesis." *Water Resources Research* 22 (4): 587-90.

[20] Vogel, R. M., and McMartin, D. E. 1991. "Probability Plot Goodness-of-Fit and Skewness Estimation Procedures for the Pearson Type 3 Distributions." *Water Resources Research* 27 (12): 3149-58.

[21] Nguyen, T. V., and In-Na, N. 1992. "Plotting Formula for Pearson Type III Distribution Considering Historical Information." Environmental Monitoring and Assessment 23 (1): 137-52.

[22] Montaseri, M. 1999. "Stochastic Investigation of the Planning Characteristics of within-year and over-year Reservoir Systems." Ph.D. Thesis, Heriot -Watt University, Edinburg, UK.

[23] Ryan, T. A. 1974. *Normal Probability Plots and Tests for Normality*. Statistics Department, The Pennsylvania State University.