RESEARCH ARTICLE OPEN ACCESS

Prediction of River Stage in an Ungauged Stream

*Onosakponome O. R, ** Mbachu V. C and ***Odenigbo C
*Department of Civil engineering, Madonna University, Akpugo campus.
**Department of Civil engineering, Madonna University, Akpugo campus
***Department of Civil engineering, Enugu State University of Science and Technology.

ABSTRACT
Hydrologic decisions are made to evaluate values of key variables and parameters needed for the design of any water resources system to make it perform adequately – in terms of safety and provision of the expected benefits. However, the degree of reliability of the hydrologic data used in the decision making process is of great significance. Unreliable data will seriously affect results. In Nigeria, as well as many parts of the less developed world, there are problems of data inadequacy, frequent gaps in the data available and non-existent data at development sites. This is the dilemma that confronts any designer of water resources systems in the less developed world. There is no doubt that sufficient and accurate hydrological data will lead to a sound engineering design of the water resources system. This study presents a solution for prediction of river stage in ungauged stream using the Principal Component Analysis (PCA). The research was illustrated by using the Imo River with a station at OBIGBO as a case study, upon which data was collected for analysis and possible development of a model. Twelve input variables were considered in the analysis; the most important contribution of PCA in this study was the identification of the key factors responsible for the changes in river stage. The amount of precipitation and the run-off discharge into the stream were the factors identified by the PCA, which can reasonably reflect the status of river stage in many streams. The developed model did not only predict the river stage of OBIGBO but also show great level of accuracy in predicting that of NEKEDE with an average correlation coefficient of 0.95. It can be concluded that the model has a great ability to predict river stage.

Keywords- Design, Hydrologic data, Prediction, River stage Water resources.

I. INTRODUCTION

River stage or flow rates are required for the design and evaluation of hydraulic structures. Most river reaches are ungauged and a methodology is needed to estimate the stages, or rates of flow, at specific location in streams where no measurements are available. Flood routing techniques are utilized to estimate the stages, or rates of flow, in order to predict flood wave propagation along river reaches. Models can be developed for gauged catchments and their parameters related to physical characteristics such as slope, reach width, reach length so that the approach can be applied to ungauged catchments in the region.

The design, planning, and operation of river systems depend largely on relevant information derived from the forecasting and estimation of extreme events. Reliable flood forecasts are particularly important for improving public safety and mitigating economic damages caused by inundations. During the past few decades, a great deal of research has been devoted to the modeling and forecasting of river flow dynamics. Such efforts have led to the formulation of a wide variety of approaches and the development of a large number of models. The existing models for river stage forecasting may broadly be grouped under two main categories namely, rainfall-runoff modeling or statistical techniques. Due to the realistic representation of watershed topography and ability to capture the surface and ground water interaction, the more reasonable method to predict a flood is the distributed and physically based model. However, extensive topographic, meteorological, and hydrologic data are required to describe the runoff process and time is also require to calibrate conceptual models (especially distributed models), which are important factors to be considered in their practical applications. Thus, the implementation and calibration of conceptual models can typically present various difficulties (Hu and Lam, 2001). In this context data-driven models, which can discover relationships from input-output data without having the complete physical understanding of the system, may be preferable. While such models do not provide any information on the physics of the hydrologic processes, they are in particular, very useful for river flood forecasting where the main concern is accurate prediction of a flood at specific watershed locations (Nayak, 2005). Flooding is a type of natural disaster that has been occurring for centuries, but can only be mitigated rather than completely solved. Prediction of

river stages becomes an important research topic in hydrologic engineering. An accurate water stage prediction allows the pertinent authority to issue a forewarning of the impending flood and to implement early evacuation measures when required.

II. THE ESTIMATION, PREDICTION AND FORECASTING OF RUNOFF

Ideally, all hydrological problems would be solved by the use of measured data, thus obviating the necessity for estimation, prediction, and forecasting. There are many circumstances, however, in which the use of these techniques becomes necessary. Thus, for example, there may be a deficiency of measured data for a particular area, but there may be the possibility of extrapolating future runoff trends either from existing runoff data relating to adjacent or nearby areas or from precipitation data. Alternatively, measured data may be collected too late to be of any use. Such is the case in areas where peaks of quickflow constitute a flood problem which must be viewed and solved in the light, not only of hydrological factors, but also of factors of settlement and communications, agriculture, and economics. Inevitably, the relevant measured data cannot become available until the flood peaks themselves have occurred and so, in these circumstances, the need is for techniques for accurately forecasting the volume and timing of quickflow peaks (Mesfin, 2008). Again, in areas where water supplies for agriculture, industry, or domestic uses are likely to be limited at times of low flow; the need is for accurate forecasts of the magnitude of dry-weather flows, and the time occurrence of minimum flow.[2-8]

The main requirements, therefore; are for techniques to forecast, for a given point within a drainage basin, both the total volume of runoff and the magnitude of the instantaneous peaks normally associated with sudden increases of quickflow and also to forecast the timing and magnitude of the minimum flows which are likely to be associated with decreasing volumes of baseflow, particularly groundwater flow. Most of the techniques currently in use were developed before the newer concepts of runoff formation. Interestingly, however, many of these methods yield reasonable results despite being conceptually weak or even erroneous. Such successes may be fortuitous but techniques are either highly empirical, and are often applicable only to restricted areas, or else are based upon factors which, although not directly cause-related to the patterns of runoff under consideration, are themselves directly affected by the real runoff-forming factors.

Although, in normal English usage, the terms forecasting and prediction are clearly synonymous, they are sometimes used in a more restricted sense by hydrologists. Thus, as Smith (1972) observed, prediction, in this context, refers to the application of statistical concepts to long periods of data, usually relating to extreme events, with a view to defining the statistical probability or return period of a given magnitude of flow. In other words, there is no indication of when this particular flow will occur. Forecasting, on the other hand, refers to specific runoff events, whether floods or low flows, and to the use of current hydro-meteorological data in order to provide a forecast of the magnitude of the runoff event and also, in many cases, of its timing. As far as possible, this distinction will be preserved in the ensuring discussions.

There are many techniques of runoff prediction and forecasting. Some of these are in widespread use, either because they work reasonably well over a wide range of conditions or else are easy to apply. The use of other techniques may be restricted to specific areas or to specific users, such as a particular Government agency. Most methods have little merit and yield poor results. It would clearly be impossible to deal with all methods or even a representative selection of the better ones. Indeed, in the present context this would not, in any case, be appropriate. Instead, the main lines of approach to the problem of river stage or runoff prediction and forecasting will be briefly reviewed in general terms and will be illustrated, where appropriate, by specific examples.

2.1 THEORY OF PRINCIPAL COMPONENT ANALYSIS

Principal component analysis (PCA) is a statistical technique for determining the key variables in a multidimensional data set that explains the differences in the observations and can be used to simplify the analysis and visualization of multidimensional data sets. In recent years, the method of principal component analysis has been widely used in many fields such as evaluation of irrigation water quality, evaluation of river water quality monitoring stations and comprehensive evaluation of the regional water resource carrying capacity.

The method of principal component analysis (PCA), using coefficients of linear correlation offers this possibility. Principal component analysis is also known as eigenvector analysis, eigenvector decomposition. Generally, the principal component of random vector X is obtained from the weight of X by special linear combination. Therefore, it is difficult to give explanation for the physical meaning of this linear combination when the dimensionless variables are different. In order to perform a PCA of the original data, random variables X have to be standardized. PCA seeks to establish combinations of variables capable of describing the principal

tendencies observed while studying a given matrix. In mathematical terms, PCA relies upon an eigenvector decomposition of the covariance or correlation matrix. The basic ideas of principal component analysis are to define F_n as a linear combination of weight X, find a linear combination of F_n for the weight X, and F_n reflects the changes of the weight X as far as possible. Here, F_1 is called the first principle component of X, and if it not yet fully reflects the changes of the weight X, we then find F_2, F_3,…..,F_r(r<n), till the information of the weight X was fully extracted and F can be given as:

$$F_n=A_1X_1+A_2X_2+...+A_nX_n \quad (1.1)$$

Given *X* observations on *n* variables, the goal of PCA is to reduce the dimensionality of the data matrix by finding *r* new variables, where r is less than *n*. Each principal component is a linear combination of the original variables, and so it is often possible to ascribe meaning to what the components represent.

It is used when there are a large number of different types of measurement for a given set of items. It aims at structuring the data by reducing the numerous variables to a smaller number of variables (components) which account for most of the variation in given data. It is a powerful technique which copes with the problems in both linear and non-linear least squares associated with statistical interrelations amongst the independent variables. It transforms the independent variables into new variables that are statistically unrelated.

The principal component analysis transforms the linear model

$$\overline{Q} = c_1 x_1 + \ldots + c_p x_p \quad (1.2)$$

To $$\overline{Q} = \beta_1\varepsilon_1 + \ldots + \beta_p\varepsilon_p \quad (1.3)$$

Where $\sum_1$ is a component or eigenvector such that

$$\text{Cov}\{\varepsilon_\alpha, \varepsilon_\beta\} = \sum \varepsilon_\alpha, \varepsilon_\beta \quad (1.4)$$

(i.e. the new variable are statistically independent)

The problem is made simpler by removing the scale effects of the original variables. Hence, the normalized original variables are defined.

$$Z_1 = \frac{x_1 - x}{s_1} \quad (1.5)$$

Letting the mean be zero and the standard deviation be unity. The first two moments of the normalized variants are

$$\sum_{j=1}^{n} Z_u = 0 \quad \text{(Mean)} \quad (1.6)$$

And

$$\frac{1}{n}\sum_{j=1}^{n} Z^2 u = 1 \text{ (Standard deviation)} \quad (1.7)$$

The solution to our problem begins by "plotting" all p of the original variables in p-dimensional space and rotating the axis until the orthogonal system of components is found. An attempt to demonstrate this for three dimensions is shown in fig.1.1. The data points are plotted as referenced to all the axes of the three original variables and then the axes are rotated until the components are orthogonal or statistically independent.

Statistically, this feat is achieved by minimizing the variance or spread around the components subject to the constraint that orthogonality must be achieved.

The simple correlation coefficient r_{ik} is given as

$$r_{ik} = \sum_{j=1}^{n} Z_u Z_{kj} \quad (1.8)$$

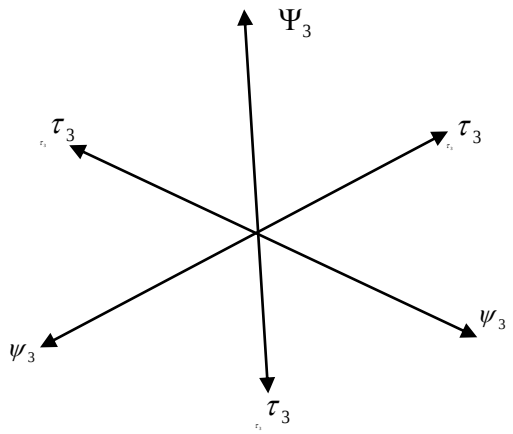

Fig.1.1: Location of components in three dimensions

The simple correlation matrix with λ as the eigenvalue in the diagonal is given as

$$\begin{bmatrix} (1-\lambda) & r_{12}\ldots r_{1p} \\ r_{21} & (1-\lambda).....r_{2p} \\ & \\ r_{p^2} & r_{p^2}.....1-\lambda \end{bmatrix} \begin{bmatrix} L_1 \\ L_2 \\ \\ L_p \end{bmatrix} = 0 \quad (1.9)$$

Where L_1------L_p represent the direction cosines i.e. the cosine of the angles of rotation of the axis.

Equation (1.9) can be written as

$$[r - \lambda 1]l = 0 \quad (2.0)$$

Since the directional cosine vector i.e. L = (L_i……..L_1) is nonzero

$$(r - \lambda 1) = 0 \quad (2.1)$$

Equation (2.1) is solved by expanding it as a determinant and obtaining an algebraic equation to the p[th] power.

$$C_1\lambda^P + C_2\lambda^{P-1} + ... + Cp + 1 = 0 \quad (2.2)$$

Equation (2.2) has p-roots, the eigenvalues. With each of these values, the eigenvectors are found by substituting back each of the eigenvalues in equation.

The principal components are found by noting the following; the variance of Z_a is

$$V(Z_a) = V\left\{\sum l_{i\beta}\varepsilon_\alpha\right\} \quad (2.3)$$

Where,

$$V(Z_\alpha) = \frac{1}{n}\left\{L_{11}\varepsilon_1 + L_{21}\varepsilon_2 + ... + L_{p1}\varepsilon_p\right\}^2 \quad (2.4)$$

This result in

$$V(Z_\alpha) = \left\{L_{11}^2 + L_{21}^2 + ... + L_{pl}^2\right\} \quad (2.5)$$

Or $$V(Z_\alpha) = \sum_{i=1}^{p} L^2 i\alpha \quad (2.6)$$

Hence, $L_{i\alpha}$ is a type of correlation coefficient representing the correlation of Z_p with ε_α. The directional cosine squared is a variance, which represents the fraction of Z_β explained by ε_α.

Once principal component analysis has been completed, the regression can be performed whereby $\sum_\alpha$ is considered to be the independent variables. Then, the model coefficient C_i in equation (1.2) can be derived from the regression coefficients in equation (1.3).

Equating the two models results in

$$\begin{bmatrix} L_{11} & L_{21}..L_{p1} \\ L_{12} & L_{22}..L_{p2} \\ \vdots & \vdots \\ L_{1p} & L_{2p[}..L_{pp} \end{bmatrix} \begin{bmatrix} C_1 \\ C_2 \\ \vdots \\ C_p \end{bmatrix} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} \quad (2.7)$$

And the solution of C_i is

$$C = L^{-1}\beta \quad (2.8)$$

When component analysis is combined with nonlinear least squares, the original variables are $\frac{\partial \hat{Q}}{\partial \alpha j}$ and the coefficients are h_i.

III. DATA AND METHOD

In this study, the data were mainly collected from the Anambra-Imo River Basin Development Authority; statistical year book 1999-2009. In the development of a model for predicting the water stage in River Imo with station at OBIGBO, 12 variables were considered believed to be contributing significantly to the water level fluctuation using the Principal Component Analysis (PCA).

The 12 variables were;

(1) Evaporation rate (X_1)
(2) Stream Discharge (X_2)
(3) Stream width (X_3)
(4) Average velocity of flow (X_4)
(5) Channel slope (X_5)
(6) Infiltration rate (X_6)
(7) Runoff discharge into the stream (X_7)
(8) Population size (X_8)
(9) Efficiency of drainage network (X_9)
(10) Catchment area (X_{10})
(11) Precipitation amount (X_{11})
(12) Length of main stream (X_{12})

The simple correlation matrix of the 12 variables was obtained. There are several significant correlations of the variables (X_i) with the water stage(y) but at a glance it would be nearly impossible to decide which one to choose. Further, there are a number of significant interrelations amongst the independent variables, but since the entire correlation procedure is a matter of degree, it would be impossible to filter out objectively the variables at this point. In addition, the standardized values of factors were calculated from the original data by SPSS.

3.1 Model performance

The performance of the model developed in this study was assessed using various standard statistical performance evaluation criteria. The statistical measures considered were multiple correlation coefficients (MCC), standard error of estimate (SEE), coefficient of correlation (CORR), mean absolute percentage error (MAPE), and root mean square error (RMSE).

IV. DATA ANALYSIS

Table 1.0: Percent variance of the 12 variables

Variables	I	II	III	Total
X_1	82.17	12.05	4.03	98.25
X_2	87.59	5.56	3.97	97.12
X_3	36.64	18.98	37.12	92.74
X_4	2.99	42.59	0.08	45.66
X_5	4.60	67.16	5.90	77.66
X_6	0.30	83.01	0.64	83.95
X_7	1.28	1.03	84.05	86.36
X_8	42.11	26.70	29.01	97.82
X_9	0.07	31.20	54.95	86.22
X_{10}	87.97	4.91	6.93	99.81
X_{11}	92.49	4.99	0.56	98.04
X_{12}	70.02	9.97	6.04	86.03

Table 1.1: Eigenvalue contribution rates and accumulated contribution rates of the principal components.

Component	Eigenvalue	% of variance	Cumulative %
1	5.08	42.33	42.33
2	3.08	25.67	68.00
3	2.33	19.42	87.42

The principal component analysis (PCA) operates as a filter of redundant information and as mechanism for model building. The equation, $(A-\lambda I) = 0$ was solved, where A is the correlation matrix of variables. Because it led to a 12 x 12 determinant, it was solved by computer. The first three largest eigenvalues (i.e., the values of λ) were selected i.e., 5.08, 3.08 and 2.33 in descending order and the others rejected. With each of these values, the eigenvectors were obtained by substituting each at a time in the equation $[A-\lambda I] [L] =0$ (equation 1.9). The eigenvectors were normalized. We see that component 1 is highly correlated with X_1, X_2, X_{10}, X_{11} and X_{12} indicating that these five variables are highly interrelated since eigenvectors are made like correlation coefficients. In component 2, we see that it is highly correlated with X_5 and X_6, while component 3 is highly correlated with X_7 and X_9 indicating that these two variables are highly interrelated. The variance are explained by the components which are the eigenvectors squared and are referred to as loading.

Our analysis indicates that we can summarize the data with just three components. Table 1.1 contains the three principal components and their corresponding eigenvalues. The results showed that of the first three components, the first component accounted for about 42.33%, the second component about 25.67% and the third component about 19.42% of the total variance in the data set. These three components together accounted for about 87.42% of the total variance and the rest of the components only accounted for about 12.58%. Therefore, our discussion focused only on the first three components.

The following decisions were made; only variable X_{11} was used since it explains 92.49% of the information contained in the component and is by far the easiest of the five interrelated variables to measure.

- Only variable X_7 was used since it explains 84.05% of the information contained in that component.
- The remaining components were deleted because the marginal variance that they explain was deemed insignificant.

Upon making the above decision, the formulated model was;

$$Y_m = ax_7^b x_{11}^c \quad (2.9)$$

Where Y_m is mean annual river stage and, a, b and c are constants.

The model was transformed to linear form by taking the natural log of equation (2.9) to get:

$$\text{In } Y_m = \ln a + b \ln X_7 + c \ln X_{11} \quad (3.0)$$

This reduces to the form;

$$Y = z + mQ + nP \quad (3.1)$$

Where P and Q are the precipitation amount (mm) and runoff discharge (m^3/s) into the stream respectively, z, m and n are regression constants.

Using the highest values of the data set (1999-2005), the constants were determined using the least square method. The model is given by;

$$Y=4.02+0.009P+0.014Q \quad (3.2)$$

The multiple correlation coefficient, MCC = 0.95 while the standard error of estimate, SEE = 0.2

4.1 Results and Discussion

The available data set was divided into two sets, from (1999 – 2005) and (2006 – 2009). The first data set was used to perform the principal component analysis and to calibrate the resulting model. The second data set was used for verification. During the verification phase, attempt was made for the validation of the model by its application to predict the river stage of OTAMIRI with its station in NEKEDE. It has a catchment area of 100 SQ KM and is located within the Imo river system. The following statistical parameters were used for the evaluation and the results presented in table 1.2 below. The parameters are; MAPE, it measures the absolute error as a percentage of the forecast, and RMSE evaluates the residual between observed and predicted river stage. CORR evaluates the linear correlation between the observed and predicted river stage.

Table 1.2; Performance of the model at different stations

	Station Performance			
	OBIGBO		**NEKEDE**	
Year	CORR	MAPE RMSE	CORR	MAPE RMSE
2006	0.95	2.15 7.49	0.94	2.34 8.00
2007	0.97	1.34 3.12	0.96	2.15 3.55
2008	0.98	1.15 2.65	0.96	1.73 5.02
2009	0.96	2.56 4.11	0.95	3.78 5.51

From the result in 2006, the model performance at OBIGBO in terms of CORR, MAPE and RMSE were 0.95, 2.15 and 7.49, respectively, which were better than those obtained at NEKEDE

(0.94, 2.34 and 8.00 respectively). The same result applies to the remaining years (2007, 2008 and 2009), the reason being that the model was developed using OBIGBO data set. However, the model proved valid being able to predict to high degree of accuracy, the river stage of OTAMIRI with station at NEKEDE. See graphical illustrations in fig.1.2-1.9 below.

Fig. 1.3: Comparison between measured and predicted River Stage in Obigbo, 2007

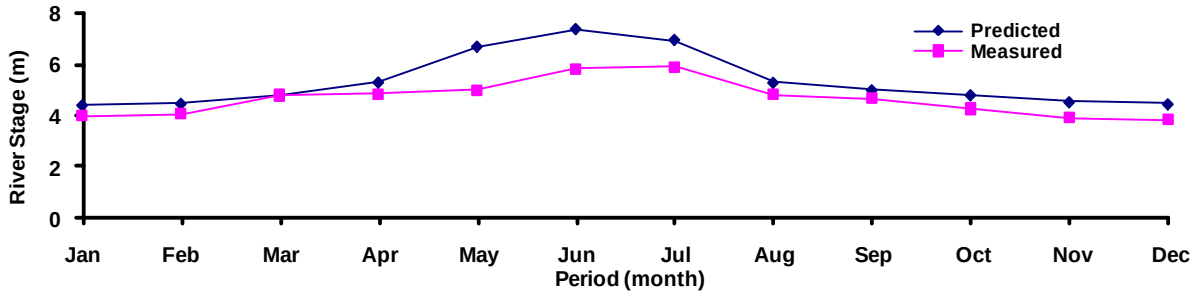

Fig 1.4: Comparison between measured and predicted River Stage in Obigbo, 2008.

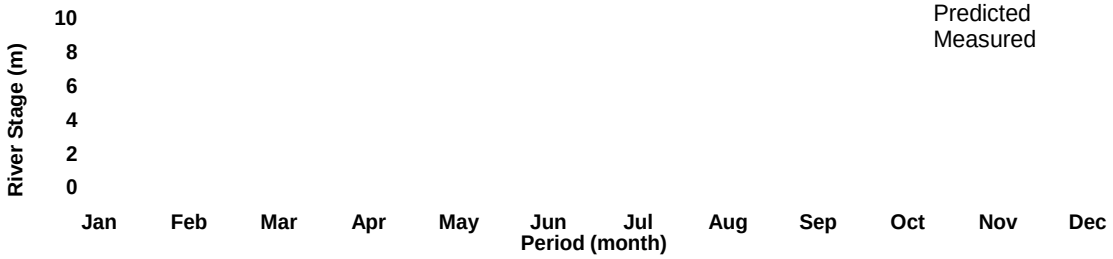

Fig 1.5: Comparison between Measured and Predicted River Stage in Obigbo, 2009

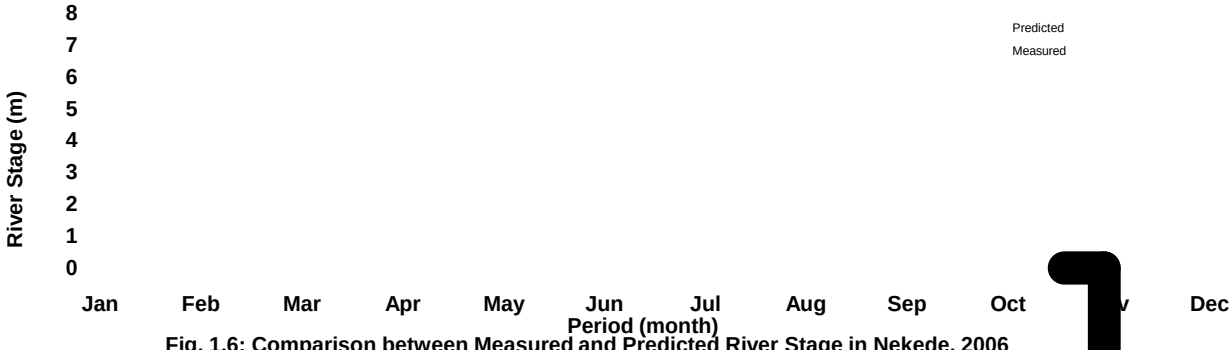

Fig. 1.6: Comparison between Measured and Predicted River Stage in Nekede, 2006

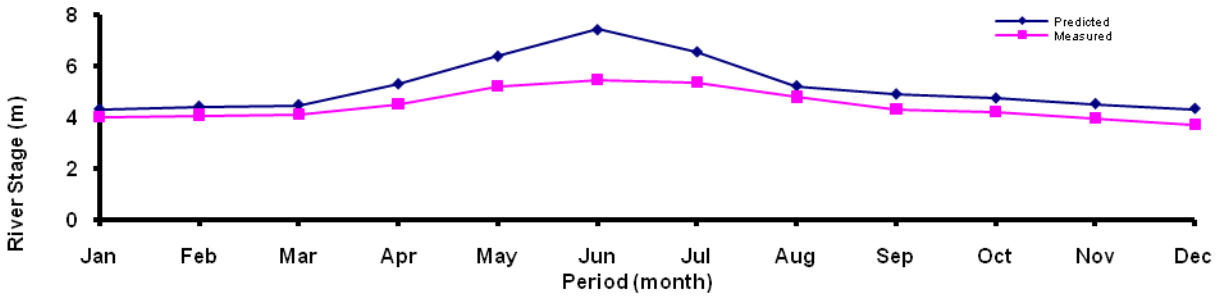

Fig. 1.7: Comparison between Measured and Predicted River Stage in Nekede, 2007

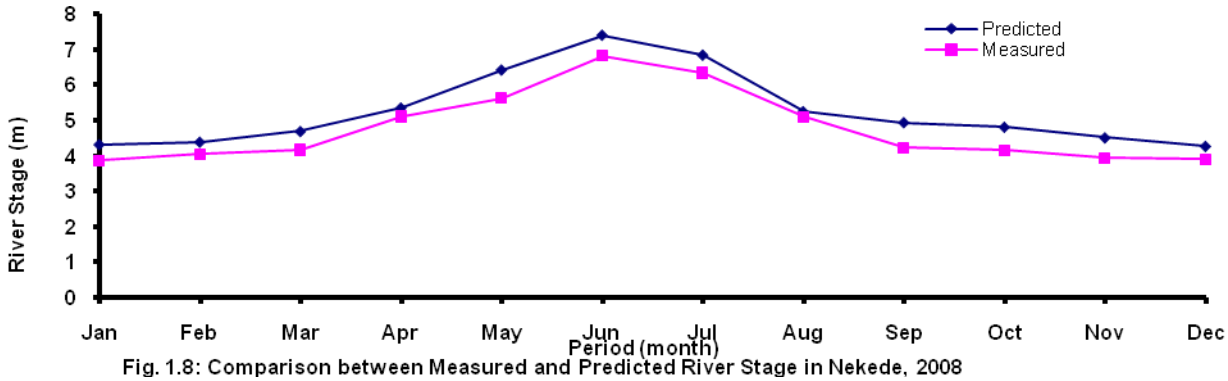

Fig. 1.8: Comparison between Measured and Predicted River Stage in Nekede, 2008

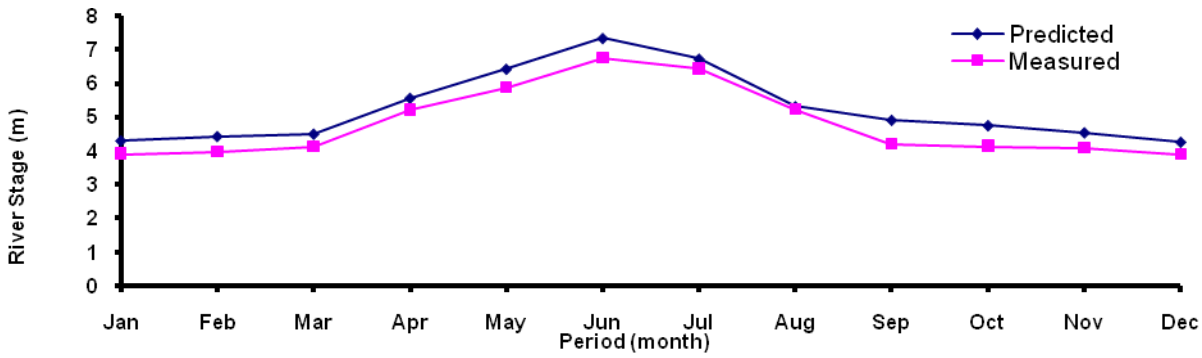

Fig. 1.9: Comparison between Measured and Predicted River Stage in Nekede, 2009

V. CONCLUSION

This study presents a solution for prediction of river stage in ungauged stream using the Principal Component Analysis (PCA). The research was illustrated by using the Imo River with a station at OBIGBO as a case study, upon which data was collected for analysis and possible development of a model. Twelve input variables were considered in the analysis; the most important contribution of PCA in this study was the identification of the key factors responsible for the changes in river stage. The amount of precipitation and the run-off discharge into the stream were the factors identified by the PCA, which can reasonably reflect the status of river stage in many streams. The developed model did not only predict the river stage of OBIGBO but also show great level of accuracy in predicting that of NEKEDE with an average correlation coefficient of 0.95. It can be concluded that the model has a great ability to predict river stage in a homogenous catchments and the predicted results provide a useful guidance or reference for flood control operations. Yet, more substantial improvement certainly should be pursued through further research to improve the forecast results.

REFERENCES

[1] Ahsan, M. & K.M. O'Connor (1994). A simple non-linear rainfall-runoff model with a variable gain factor. J. Hydrol., 155. pp. 151 – 183.

[2] Ayub, S., et al (2005). "Flood and drought forecasting and early warning program (for the Nile Basin)". Nile Basin Capacity

Building Network for River Engineering (NBCBN - RE).

[3] Bengraine, K and Marhaba, T.F. 2003. Using principal component analysis to monitor spatial and temporal changes in water quality. Journal of Hazardous Material, 3: 179 – 195

[4] Betson, R.P., (1964) "What is watershed runoff?" JGR, 69: 1541 – 1551.

[5] Chang, F.J. & Chen, Y.C. (2001) A counterpropagation fuzzy-neural network modeling approach to real time streamflow prediction. J.Hydrol., 245, 153 – 164.

[6] CHOW, V.T. (1964) "Runoff" section 14 in Chow, V.T. (ed.). Handbook of Applied Hydrology, McGraw_Hill, New York.

[7] Clarke, R.T. (1994). Statistical Modeling in Hydrology. John Wiley and Sons.

[8] Cunge, J.A. (1969). "On the subject of a flood propagation computation method (Muskingum method), Journal of Hydraulic Research, 7 (2).

[9] Dalrymple, T. (1960) "Flood frequency analysis": Manual of Hydrology, part 3, flood flow techniques, USGS Wat. Sup. Pap.,1543-A 80 pp.

[10] Dawdy, D.R. & O'Donnell, T., (1965). Mathematical Models of Catchment Behaviour. J. of the Hydraulics Div., A.S.C.E., Vol. 91, Number HY4 Proc. Paper 4410. pp. 123 – 137.

[11] De Roo, A & G. Schmuck (2003). European Flood Alert System. Joint Research Center, Institute for Environment and Sustainability. European Commissions.

[12] De Zhou, Rougquu, Z., Liming, L., Linghing, G. and Simin, C. (2009). "A study on water resources consumption by principal component analysis in Qingtongxia irrigation areas of Yinchuau plain, China". Journal of food, Agriculture and environmental Vol. 7. (3 & 4): 734 – 738.

[13] Elzein, A.S. & I.S. Adam (1996). Sudan Flood Early Warning System (FEWS). ElMoutandis, Vol. 2, No. 5, pp. 23 – 24.

[14] Hu, T.S., Lam, K.C. & Ng. S.T. (2001) River flow time series prediction with a range-dependent neural network. Hydrol. Sci. J., 46, 729 – 745.

[15] Mesfin, H. T (2008). "Flood routing in ungauged catchments using Muskingum Methods". Dissertation.com, Boca Raton, Florida, USA.