NON-LINEAR STATISTICAL MODEL FOR THE DAILY STREAM FLOW PREDICTION IN THE KALU RIVER CATCHMENT IN SRI LANKA

H.A.N.S. Hewavitharana[1], K.Perera[2] and K.D.W. Nandalal[3]

[1]Postgraduate Institute of Science, University of Peradeniya, Sri Lanka.
(hewavitharana.09@gamil.com)

[2] Department of Engineering Mathematics, University of Peradeniya, Sri Lanka.
(kanthip@pdn.ac.lk)

[3]Department of Civil Engineering, University of Peradeniya, Sri Lanka.
(kdwn@pdn.ac.lk)

Abstract
Having a long record of stream flow is very valuable in planning water resources development projects. However, in many occasions, stream flow records are available for very short periods though very long rainfall records are available. Therefore, possibility to relate rainfall over a catchment to the stream flow at its outlet will enable having a long record of stream flow. Besides prediction of stream flow using already available predicted rainfall will permit taking precautionary measures in water related disaster situations such as floods and droughts. This paper presents a research carried out to find a model to predict daily stream flow of Kalu River at Ratnapura. The model, a non-linear regression model based on Marquardt's procedure, was developed using measured daily stream flow at Kalu River at Ratnapura and daily rainfall at eight rainfall gauging stations within the catchment above Ratnapura. Data for the period 1987-1994 were used for the calibration of the model while data for the period 1995-2000 were used for verifying it. The model was validated using Nush-Sutcliffe efficiency and pseudo R^2. Nush-Sutcliffe efficiency (78%) and pseudo R^2 (85%) show the possibility of the fitted model in predicting daily stream flow of Kalu River at Ratnapura.

Keywords: Marquardt's procedure, Non-linear regression, Nash-Sutcliffe efficiency, pseudo R^2

1. Introduction

The determination of the amount of water that would be carried by stream in the future is of crucial importance since it directly affects the design and operation of many water resource structures. Learning the hydrological behavior of a water structure is the first step of design. Because of this reason stream flow data are very important for many areas of water engineering such as dam planning, flood mitigation, operation of water reservoir, distribution of drinking water and drainage water, hydropower generation in dry periods and planning of river transport.

Kalu River is the second largest river in Sri Lanka. It drains an area about 2690 km^2. Magnitude of the annual flow is over 7300 million m^3. Rainfall occurs in Sri Lanka during the North-east, South-west, First inter-monsoon and Second inter-monsoons. During the Southwest monsoon rainfall is mainly confined to the Southwest of the Island, whereas during the Northeast monsoon rainfall occurs in the North and East of the Island. Kalu River catchment receives rain during both of these monsoons. The average annual rainfall of the overall catchment is around 4000 mm and it ranges from 2750 mm in coastal areas to 5000 mm in mountainous areas. Since the catchment is entirely situated in wet zone, it has a high rainfall to runoff response.

The objective of this research is to fit a nonlinear statistical model for average daily streamflow prediction in Kalu River upper cachtment using average daily rainfall. The Theory of regression and correlation is a classical frame work to describe relation between two or more variables. Rainfall-stream flow relationship is a very typical example of the application area of the regression analysis in the science of hydrology in the generation of stream flows using rainfall. This study uses a totally different method, the non-linear regression analysis using marquardt's procedure, to develop a model to compute stream flow. The model is estimated directly using input output data. The resulting model, which was developed with the application of marquardt's procedure in non linear regression is an algorithm for least square estimation of nonlinear parameters.

2. Methodology

Daily stream flow prediction has been done by the several scholars in different locations all over the world. Tabrizi et al (2005) found modeling of Non-linear stochastic systems using effective rainfall by considering non-linear autoregressive moving average with exogenous input. Kluppelberg et al (2008) used electricity spot price modeling with a view towards extreme spike risk by non-linear regression analysis. Farahmand et al (2006) studied daily stream flow prediction using impulse response approach. Amisigo et al (2006) predicted monthly stream flow by considering Single-Input-Single-Output (SISO) non-linear system identification techniques. They employed it to model monthly catchment runoff at selected gauging sites in the Volta Basin of West Africa.

2.1 Study area and data used

Ratnapura is located in the south-western part of Sri Lanka within the wet zone. Ratnapura is the main city of Sabaragamuwa Province and it is located in a valley, which is approximately 21 m above mean sea level and it is surrounded by mountain ranges.

Daily average rainfall and daily average stream flow are required as input to the model. Data for the period 1987-2000 at the stations given in Table 1 were used in this study.

Table 1. Rainfall and stream flow gauging stations

	Gauging Station	Data Period	Latitude (N)	Longitude(E)
Rainfall (mm)	Ratnapura	1987-2000	6.68	80.40
	Pinnawala	1987-2000	6.73	80.86
	Kuruwita	1985-2000	6.80	80.38
	Lellopitiya Estate	1987-2000	6.68	80.50
	Alupola Group	1987-2000	6.72	80.58
	Balangoda (Post office)	1987-2000	6.65	80.70
	Wellandura Estate	1987-2000	6.53	80.57
	Hapugastanna	1987-2000	6.72	80.52
Steam flow (m^3/s)	Ratnapura	1987-2000	6.68	80.40

2.2 Thiessen polygon method

Thiessen polygon network was drawn by connecting eight rainfall stations and drawing perpendicular lines through mid points of each line segment. The method assumes that the area covered by a polygon gets an average rainfall similar to the rainfall received at the station inside the polygon. Average rainfall for the catchment was calculated as weighted average rainfall using following equation.

$$\overline{R} = \frac{(\sum_{i=1}^{n} R_i A_i)}{A}$$

Where, Ri is the daily average rainfall of the i^{th} station in mm

$\overline{R}$ is the average daily rainfall over the catchment in mm

Ai is the area covering i^{th} station in km^2

A is the total area of the catchment in km^2

2.3 Daily average rainfall and daily average stream flow modeling

Many daily average rainfall and daily average stream flow were studied using various approaches available for modeling.

Non-Linear regression model

General form of the non-linear regression model is Farahmand et al (2006), Neter,4th editon:

$$y_i = f(x_i, \gamma) + \epsilon_i$$

Where;

$$x_i = \begin{pmatrix} x_{i1} \\ .. \\ x_{iq} \end{pmatrix} \qquad \gamma = \begin{pmatrix} \gamma_0 \\ .. \\ \gamma_p \end{pmatrix}$$

ε_i =Random error $\qquad$ Y_i = Response variable

2.4 Estimation of regression parameters

Data from 1987 to 1995 have used to estimate the model parameters of the average daily stream flow model and 1996 to 2000 data have used to check the validation of the predicted model.

In many nonlinear regression problems, it is more practical to find the least estimates by direct numerical search procedures. Direct numerical search procedures are Gauss-Newton, steepest descent and Marquardt's procedures.

Marquardt's procedure

The Marquardt procedure is an iterative procedure. To start a minimization, the user has to provide an initial guess for the parameter vector. This procedure seeks to utilize the best features of the Gauss-Newton and steepest descent procedures and occupies a middle ground between these two procedures.

In one hand uninformed initial values will work on the other hand the procedure converge only initial guess close to the final solution. In each step the parameter vector is replaced by a new estimator. If reduction of sum of square is rapid, a smaller value can be used for the iterative procedure. The reduction of the sum of squares from the least parameter vector, then iteration stops and the last parameter vector is to be the solution Neter et al 4th edition.

2.5 Model validation

Nash-Sutcliffe efficiency

The efficiency provide by the Nash and Sutcliffe et al (2005) is defined the one minus sum of the absolute squared difference between the predicted and observed values normalized by the variance.

$$E = 1 - \frac{\sum_{i=1}^{n}(Oi - Pi)^2}{\sum_{i=1}^{n}(Oi - \widehat{O})^2}$$

Where,
E = the Nash-Sutcliffe efficiency.
Oi =Observed value
Pi =Predicted value
$\overline{O}$ =Average value of the observed value

The normalization of the variance of the observation series results in relatively higher value of E in catchment with high dynamics. The range of Nash-Sutcliffe efficiency lies between 1 and –infinity. If the value is closer to 1 model, it would make perfect fit model.

Pseudo-R^2

In nonlinear regression, such a measure is not readily defined. One of the problems with the R^2 definition is that it requires the presence of an intercept, which most nonlinear models do not have. A measure relatively closely corresponding to R^2 in the nonlinear case is:

$$Pseudo-R^2 = 1 - \left(SSE / SST_{Corrected} \right)$$

where;

SSE= Sum of square error

SST corrected= Total sum of square corrected

3. Analysis of the system

3.1 System description

The system under investigation consists of eight rainfall stations and one stream flow station. Eight rainfall stations are Allupola group, Balangoda, Kuruwita, Pinnawala, Rathnapura, Lellopitiya Estate, Wellandura Estate, Hpugastanna and steam flow station is Ratnapura.

3.2 Data analysis

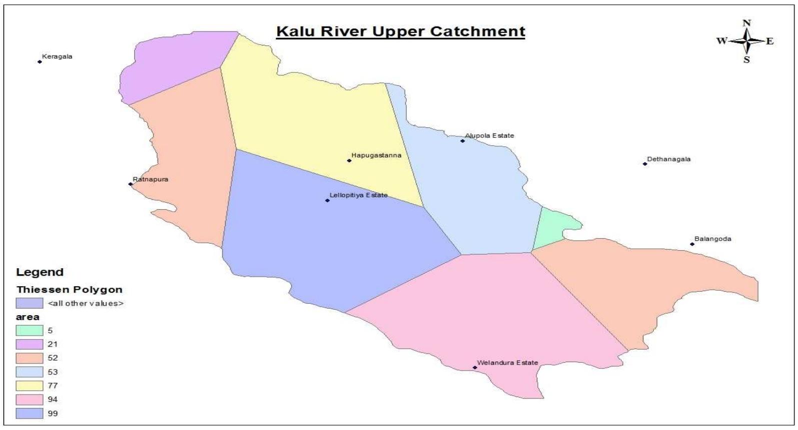
The data analysis began by investigating the characteristics of the data. The procedure included analyzing the statistical properties of data, normality of data, mean analysis, non stationary analysis and Thiessen polygon method to calculate average rainfall over the catchment.

The above statistical analysis suggested that the rainfall data set is not well correlated and that there may be significant non stationarity in the data. Therefore, it is essential to pre-process by shifting the data, removing the mean, transformation or use of proper rainfall filter. Logarithmic transformation is used to minimize the variation of the average daily stream flow.

The data during the period from 1987 to 1994 were used in the calibration of the model while data from 1995 to 2000 were used to validate the model.

4. Results and discussion

Daily rainfall of eight rainfall stations were used to calculate average daily rainfall over the catchment based on the Thiessen polygon. The Thiessen polygon network and the polygon areas are given in Figure 1.



Rainfall Station	Area(km^2)
Ratnapura	51.55
Pinnawala	4.63
Kuruwita	21.21
Hapugastanna	76.76
Lellopitiya	99.45
Alupola	53.48
Balangoda	51.77
Wellandura	94.00

Figure 2. Thiessen polygon Network

The modeling of average daily rainfall and average daily stream flow processes is a major task for hydrologist who requires such model applications such as flood forecasting, waste water flow rate forecasting and water table evaluation and such models received considerable attention in the hydrological literature. However, the average daily rainfall and average stream flow are clearly dynamic with non-linear characteristics.

Since the antecedent average daily rainfall condition clearly affects the subsequent flow behavior, the physical process should be non-linear. If the prior average daily rainfall has been sufficient to thoroughly wet soil in the catchment area then the average daily stream flow in the river will be significantly higher than the condition when the soil had been dried out due to the lack of rainfall.

Since there is seasonal variation, two cosine terms with time trend were used to catch the seasonality of the data set. Also lag 1, lag 2, lag3 of average daily rainfall (r_{t-1}, r_{t-2} and r_{t-3}), and lag 1 of logarithmic transformation of average daily stream flow (z_{t-1}) were used.

$$Z(t) = 0.0239 * \cos(3*(22/7)*t/950) + 0.0239 * \cos(2*(22/7)*t/3) + 0.9316 * Z(t-1)$$
$$+ 0.00704 * r(t-1) - 0.00093 * r(t-2) - 0.00279 * r(t-3)$$

The number of previous days average daily rainfall and logarithmic transformation of average daily stream flow were decided based on the improvement on the model statistics. The predicted and the observed average daily stream flow of the Kalu River at Ratnapura for the calibration period is given in Figure 2.

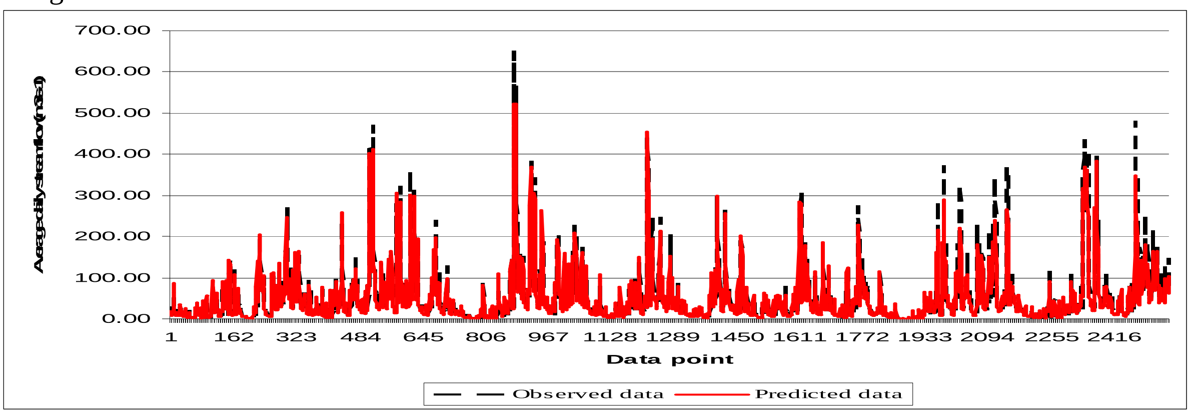


Figure 2. Plot of observed and predicted stream flow for the calibration period

Testing of Hypothesis

According to the analysis of variance (Table 2), p-value less than 5% significant level it can be conclude that the model is significant. Table 4 presents the significance of parameters.

Table 2. Analysis of variance

Source of Variation	DF	Sum of Square	Mean Square	F-value	P-value
Regression	6	7567.7	1261.3	4770.05	<.0001
Error	2547	155.5	0.0454		
Uncorrected total	2553	7683.2			
Corrected total	2552	1197.4			

Table 3. Significance of the parameters

Parameters	Estimates	Approximate Standard error	Approximate 95% confidence interval	
b0	0.02390	0.00534	0.01340	0.0344
b1	0.9316	0.00516	0.9215	0.9417
b2	-0.00093	0.00036	-0.00164	-0.00023
b3	0.00704	0.00035	0.00635	0.00773
b4	0.00407	0.00032	0.00344	0.00407
b5	-0.00279	0.00034	-0.00346	-0.00213

According to the Table 3, seven parameters do not have confidence interval including zero. Therefore, all parameters are significant.

Model validation

Table 4 shows test statistics for the calibration period and validation period of the model.

Table 4. Measure of predictability of the model

	Calibration	Validation
R^2-Pseudo	0.90	0.90
Nash-Sutcliffe efficiency	0.95	0.84
Press statistic	40.68	49.89
Root MSE	0.21	0.35

Figure 3 is the plot of daily average stream flow for the validation period of 1996-2000. It shows that the deviations of the two curves are not significant and the distribution patterns of the two data sets are almost similar. Figure 4 shows that the plot of monthly total of the stream flows for the validation period. Both observed data set and predicted data set follow same distribution.

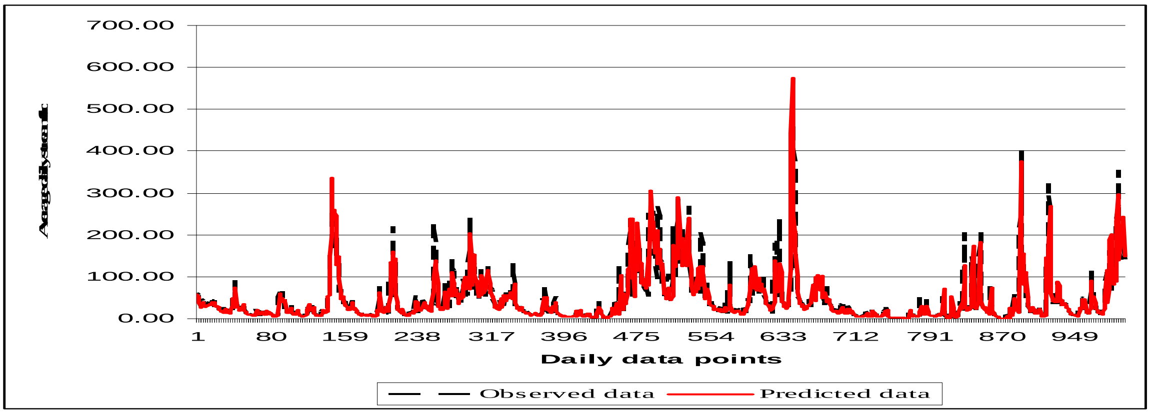


Figure 3. Average daily stream flow for the validation period

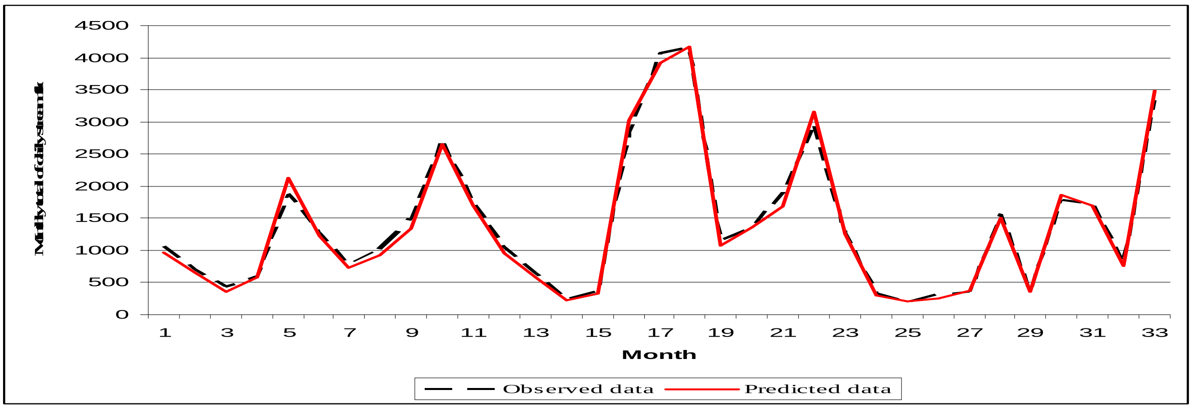


Figure 4. Monthly total of the stream flow for the validation period

5. Conclusion

Model has been estimated using marquardt's procedure. The comparison of the model predicted output with the actual output over both the estimation and test data set were used to measure the model performance.

However, the model validity test suggested that there were missing model terms. This may be due to the difficulty involved in fitting models to rainfall and stream flow data which are affected by many unknown factors, such as soil moisture, temperature, etc.

Finally, non-linear regression model was developed by using marquardt's procedure for the Kalu River upper catchment outlet for the calibration period 1987-1995. Since the model is significant according to the p-value, the fitted model can be suggested as suitable for the generation of daily average stream flow.

6. References

1. Tabrizi, H.N., Mohammad, J. and A.S. Billing (2004) Modeling of Non-linear stochastic systems.
2. Kluppelberg, C., Brandis, T.M. and A. Schmidt (2008) Electricity spot price modeling with a view towards extreme spike risk.
3. Amisigo, B.A., Rogers, C., Andah, W.E.I. and J. Friesen (2006) Monthly Streamflow Prediction in the Volta Basin of West Africa: A SISO NARMAX Polynomial Modelling
4. P.Krause, D.P. Boyle, F. Base (2005), "Comparison of different efficiency criteria for hydrological model assessment".
5. Gordon K. Smyth Edited by Abdel H. El-Shaarawi and Walter W. Piegorsch(2002),"Nonlinear regression", Volume 3.
6. Prajneshu, Indian Agriculture Statistical Research Institute, Library Avenue, New Dilhi, "Non-linear regression models and their application".
7. Neter, Katner, Nachtsheim, Wasserman, "Applied Linear Statistical Model", 4th Edition, 531-559.
8. Farahmand, T., Fleming, S.W., and E.J. Quilty(2006) Detection and visualization of storm hydrograph changes under urbanization: An impulse response approach