MATEMATIKA, 2013, Volume 29, Number 1b, 67-75

Streamflow Forecasting at Ungauged Sites using Multiple Linear Regression

[1]**Basri Badyalina and** [2]**Ani Shabri**
[1,2]Department of Mathematical Sciences, Faculty of Science, Universiti Teknologi Malaysia,
81310 UTM Johor Bahru, Johor, Malaysia
e-mail: [1]basribadyalina@gmail.com, [2]ani@utm.my

Abstract Developing reliable estimates of streamflow prediction are crucial for water resources management and flood forecasting purposes. The objectives of this study are to identifying which the physiographical and hydrological characteristics affected in multiple linear regressions (MLR) model to estimated flood quantile at ungauged site. MLR model is applied to 70 catchments located in the province of Peninsular Malaysia. Three quantitative standard statistical indices such as mean absolute error (MAE), root mean square error (RMSE) and Nash-Sutcliffe coefficient of efficiency (CE) are employed to validate models. MLR model are built separately to estimate flood quantile for T=10 years and T=100 years. The results indicate that elevation, longest drainage path and slope were the best input for MLR model.

Keywords Multiple Linear Regression; Streamflow Forecasting; Ungauged Site.

2010 Mathematics Subject Classification 62P12, 62H10, 62H12

1 Introduction

Accurate estimate of streamflow is important for many engineering project such as flood risk assessment projects, watershed planning and management of hydraulic structures projects such as; dams, roads and design urban drainage system [1, 2]. In order provide reliable estimate of streamflow, historical data at-site of interest is needed for estimate. However, it often happen the historical data at-site of interest not always available. Although at-site of interest may have some available data but the data is not enough to describe the catchment flow because of the changes in watershed characteristics such as urbanization [3]. The UK Flood Estimation Handbook (FEH) notes that "many flood estimation problems arise at ungauged sites which there are no flood peak data" [4]. Typically some site characteristics for the ungauged sites are known. Thus, regionalization is carried out to make estimates of flow statistics at ungauged sites using physiographic characteristics. In streamflow modeling and forecasting, it is hypothesized that incorporating the catchment characteristics variables would improve prediction accuracy and model reliability. The variables affecting the streamflow prediction include catchment characteristics (size, slope, shape and storage characteristics of the catchment), storm characteristics (intensity and duration of rainfall events), geomorphologic characteristics (topology, land use patterns, vegetation and soil types that affect the infiltration) and climatic characteristics (temperature, humidity and wind characteristics) [5,6].

The objective of research paper is to identify which characteristics or input for MLR model that the most effective in estimating flood quantile at ungauged site at Peninsular Malaysia. The five characteristics are rearranged to build 31 combination types of inputs for MLR model. MLR for modeling catchment characteristics against observed (flow) is the most commonly approached used in rainfall runoff modeling [7]. There are some previous researches used multiple linear regression and flood frequency analysis in forecasting flow when historical data not available. MLR is the most consistent method for estimating flood quantiles for unguauged sites [3, 8, 9]. The linear regression based methods of flood regionalization used to make estimates of flow for ungauged sites discussed by Vogel and Kroll [8], Tasker et al. [10] and Pandey and Nguyen [3]. One of the most widely used in regionalization technique is fitting a probability distribution to a flow series, or parameters to a flow duration curve, and then relating the model parameters to physical catchment characteristics [11]. The performances of regression models in estimating the flood quantiles for ungaged sites have been assessed in Pandey and Nguyen [3] by applying jackknife procedure in simulating the ungauged sites. The jackknife procedures are required to simulate gauged station to represent ungauged site.

2 Methodology

2.1 MLR Based Method of Regionalization

The performances of regression models in estimating the flood quantiles for ungaged sites have been assessed in Pandey and Nguyen [3] by applying jackknife procedure in simulating the ungauged sites. In order to estimates streamflows at ungauged sites, power form function such as:

$$Q_T = \alpha_0 A_1^{\alpha_1} A_2^{\alpha_2} \ldots A_m^{\alpha_m} \varepsilon_0 \tag{1}$$

is commonly used to build relation between streamflow and the catchment characteristics [12, 13, 14]. Here, $\alpha_0, \alpha_1, \ldots, \alpha_m$ are the model parameters, $A_1, A_2, \ldots A_m$ are the catchment characteristics, ε_0 is the multiplicative error term, m is the number of catchment characteristics and Q_T represents flood quantile of T-year return period. Eq. (1) can be solved using linear regressions by linearizing the power form model using a logarithmic transformation to the form. The linearized power form model becomes as follow:

$$\ln(Q_T) = \ln(\alpha_0) + \alpha_1 \ln(A_1) + \alpha_2 \ln(A_2) + \ldots + \alpha_n (A_n) + \ln(\varepsilon_0). \tag{2}$$

Eq. 2 can be solved using MLR. MLR attempts to model the relationship between two or more explanatory variables and a response variable, by fitting a linear equation to the observed data [15]. The dependent variable y is given by

$$y = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \varepsilon \tag{3}$$

where are the explanatory variables, β_i are regression coefficient, and ε is the error that associated with the regression and assumed to be normally distributed with expectation value zero and constant variance. The sample estimate of the parameter vector β, is given by

$$\beta = (X^T X)^{-1} X^T Y \tag{4}$$

where X is the design matrix that contains the levels of explanatory variables:

$$X = \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{k1} \\ 1 & x_{12} & x_{22} & \cdots & x_{k2} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{kn} \end{bmatrix} \tag{5}$$

n i the sample size and

$$Y = \begin{pmatrix} y_1 & y_2 & \dots & y_n \end{pmatrix}^T \tag{6}$$

is the vector of observations of the response variable.

3 Experimental Design

3.1 Data

The annual maximum flow series from 88 stations obtained were used in this study. The data obtained from Department of Irrigation and Drainage, Ministry of Natural Resources and Environment, Malaysia. Fig. 1 shows the location of the study region. The stations include wide variety of basins region ranging from 16.3 km^2 to 19,000 km^2. The period of the flow series for different sites vary from 11 -50 years starting from 1959 – 2009. Two types of data, physiographical and hydrological data are used in this study. Five variables including four physiographical variables and one hydrological variable were implemented in this work. The four physiographical variables are catchment area (AREA), mean catchment slope (MCS), elevation (ELV) and longest drainage path (LDP). The hydrological variable is annual mean total rainfall (AMR). Probability distributions such as generalized extreme value (GEV), generalized pareto (GPA) and generalized logistic (GLO) distributions were fitted to the flow series using L-moments estimator (Hosking, 1990). The generalized extreme value (GEV) statistical model used in this study to estimate flood quantile for 10- and 100- years return period. This model was found suitable for flood patterns in Malaysia [16].

Figure 1 Map showing location of stream flow stations used in the study

3.2 Evaluation Criteria

3.2.1 Evaluation Criteria

To assess the performance of each regional flood frequency analysis model, the following numerical indices are used: mean absolute error (MAE), root mean square error (RMSE), Nash-Sutcliffe coefficient of efficiency (CE) and coefficient of determination (r^2). The definitions of MAE, RMSE, CE and R are provided in Eq. (7) - (10).

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|Q_{T,i} - \hat{Q}_{T,i}\right| \tag{7}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(Q_{T,i} - \hat{Q}_{T,i}\right)^2} \tag{8}$$

$$CE = 1 - \frac{\sum_{i=1}^{n}\left(Q_{T,i} - \hat{Q}_{T,i}\right)^2}{\sum_{i=1}^{n}\left(Q_{T,i} - \overline{Q}_{T,i}\right)^2} \tag{9}$$

$$r^2 = \left(\frac{\sum_{i=1}^{n}(Q_{T,i} - \overline{Q}_{T,i})(\hat{Q}_{T,i} - \overline{\hat{Q}}_{T,i})}{\sqrt{\sum_{i=1}^{n}(Q_{T,i} - \overline{Q}_{T,i})^2}\sqrt{\sum_{i-1}^{n}(\hat{Q}_{T,i} - \overline{\hat{Q}}_{T,i})^2}}\right)^2 \tag{10}$$

where $Q_{T,i}$ is the observed flows, $\hat{Q}_{T,i}$ is the predicted flows, $\overline{Q}_{T,i}$ is the mean of the observed flows, $\overline{\hat{Q}}_{T,i}$ is the mean of the predicted flows and n is the number of flow series that have been modeled. The MAE is related with the prediction bias whereas the RMSE is associated with the model error variance. Both of MAE and RMSE evaluate how closely the predictions match the observations by judging the best model based on the relatively small MAE and RMSE values. The coefficient of efficiency (CE) provides an indication of how good a model is at predicting values away from the mean. CE ranges from $-\infty$ in the worst case to 1 (perfect fit). An efficiency of lower than zero indicates that the mean value of the observed flow would have been a better predictor than the model. Coefficient of determination can also be expressed as the squared ratio between the covariance and the multiplied standard deviations of observed and predicted values. The range of r^2 lies between 0 and 1, and it describes how much of the observed dispersion is explained by the prediction. A value zero means no correlation at all whereas a value of 1 means that dispersion of the prediction is equal to that observation.

3.2.2 MLR Implementation

A jackknife multiple linear regression is simulated using MATLAB software. The observed flow data are expressed as a function of catchment area (km^2), annual mean rainfall (mm), elevation (m), longest drainage path (m) and mean catchment slope (%). From Eq. (2) the observed flow and five explanatory variables are converted into the natural logarithm form. The model then is fitted by regular least squares procedures.

4 Results and Discussion

The objective of this paper is to investigate the effects of variables towards the performance of multiple linear regressions in estimating the flood quantiles for ungauged sites. To this end, effort is focused on selecting best input variables for MLR in order MLR to perform a good estimation. As stated earlier, there are five variables using in this study. The five variables are area (x_1), elevation (x_2), longest drainage path (x_3), mean catchment slope (x_4) and annual mean total rainfall (x_5). The performance of each model depend on it prediction quantiles. The prediction quantiles compared in the real domain and not the logarithm transformation [3]. The data set split into two sets of data which are training and testing data sets. The training data set is used to fit the model and obtain the model parameters while the testing data set used to evaluate the performance of the model. In this study jackknife procedure was implemented for simulating the ungauged sites. Jacknife procedure required to move one site form the data set and the parameters models are estimated using the remaining site in data set. This process is repeated until all sites are removed at least once [3].

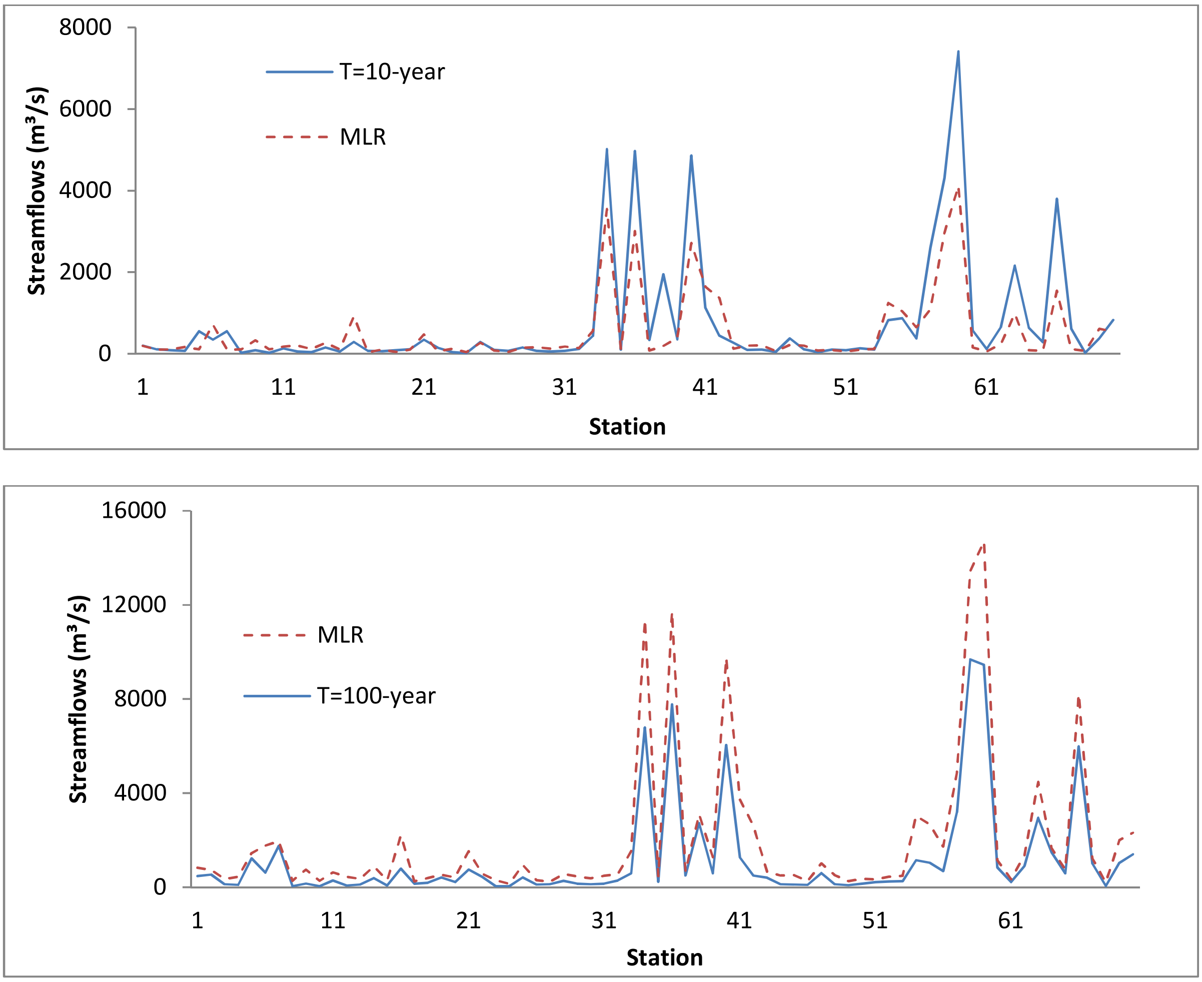

Figure 2 Observed and best predicted streamflow by MLR models of stations in Peninsular Malaysia for 10 year and 100 year return periods

Table 1 Performance MLR using different variables obtained from the jackknife procedure for T=10-year.

Variables implement in model	T = 10-year			
	RMSE	MAE	CE	R
x_1	1002.2000	457.0000	0.5168	0.7225
x_2	1536.2000	678.2000	-0.1354	0.0358
x_3	898.2000	438.9000	0.6118	0.7684
x_4	1426.8000	626.2000	0.0206	0.1795
x_5	1539.9000	669.6000	-0.1408	0.0391
x_1, x_2	944.9000	442.3000	0.5704	0.6567
x_1, x_3	797.9000	388.4000	0.6937	0.7672
x_1, x_4	1005.4000	460.6000	0.5137	0.7171
x_1, x_5	1055.7000	469.4000	0.4638	0.6564
x_2, x_3	836.5000	407.9000	0.6633	0.8086
x_2, x_4	1425.9000	629.8000	0.0218	0.1660
x_2, x_5	1539.0000	671.8000	-0.1396	0.0245
x_3, x_4	883.3000	440.2000	0.6246	0.7832
x_3, x_5	955.6000	453.2000	0.5606	0.7225
x_4, x_5	1441.5000	628.7000	0.0003	0.1215
x_1, x_2, x_3	818.6000	395.6000	0.6776	0.7401
x_1, x_2, x_4	948.8000	446.4000	0.5669	0.6496
x_1, x_2, x_5	995.3000	456.3000	0.5233	0.6131
x_2, x_3, x_4	746.2000	383.1000	0.7321	0.8697
x_2, x_3, x_5	905.7000	422.9000	0.6054	0.7604
x_3, x_4, x_5	940.2000	454.6000	0.5747	0.7408
x_1, x_3, x_4	1872.9000	1028.7000	-0.6877	0.0283
x_1, x_3, x_5	866.1000	405.8000	0.6391	0.7189
x_2, x_4, x_5	1441.9000	633.0000	-0.0003	0.7408
x_3, x_4, x_5	866.1000	405.8000	0.6391	0.1133
x_1, x_2, x_3, x_4	758.5000	386.5000	0.7232	0.7889
x_1, x_2, x_3, x_5	879.5000	411.9000	0.6278	0.6972
x_2, x_3, x_4, x_5	819.1000	399.2000	0.6772	0.8350
x_1, x_3, x_4, x_5	816.2000	398.2000	0.6795	0.7560
x_1, x_2, x_4, x_5	999.3000	460.5000	0.5196	0.6050
x_1, x_2, x_3, x_4, x_5	820.3000	401.7000	0.6763	0.7538

Table 2 Performance MLR using different variables obtained from the jackknife procedure for T=100-year.

Variables implement in model	T = 100-year			
	RMSE	MAE	CE	R
x_1	1549.7000	735.8000	0.4665	0.6588
x_2	2257.1000	1022.5000	-0.1317	0.0416
x_3	1466.7000	726.2000	0.5221	0.6914
x_4	2094.8000	944.4000	0.0251	0.1851
x_5	2260.8000	1008.1000	-0.1355	0.0236
x_1, x_2	1509.1000	723.1000	0.4941	0.5765
x_1, x_3	1339.9000	663.2000	0.6012	0.6974
x_1, x_4	1555.6000	741.8000	0.4624	0.6553
x_1, x_5	1599.0000	754.0000	0.4320	0.6234
x_2, x_3	1384.3000	689.9000	0.5743	0.7485
x_2, x_4	2095.9000	948.5000	0.0241	0.1667
x_2, x_5	2259.3000	1012.5000	-0.1339	0.0140
x_3, x_4	1455.3000	729.9000	0.5295	0.6960
x_3, x_5	1521.2000	746.4000	0.4860	0.6655
x_4, x_5	2110.3000	943.4000	0.0107	0.1383
x_1, x_2, x_3	1388.4000	677.9000	0.5718	0.6462
x_1, x_2, x_4	1517.2000	730.6000	0.4886	0.5687
x_1, x_2, x_5	1554.0000	743.5000	0.4635	0.5537
x_2, x_3, x_4	1287.3000	662.7000	0.6318	0.8012
x_2, x_3, x_5	1443.7000	710.3000	0.5370	0.7292
x_3, x_4, x_5	1509.4000	749.1000	0.4939	0.6724
x_1, x_3, x_4	1302.5000	656.6000	0.6231	0.7056
x_1, x_3, x_5	1399.6000	686.8000	0.5649	0.6724
x_2, x_4, x_5	2113.0000	948.9000	0.0081	0.1255
x_3, x_4, x_5	1509.4000	749.1000	0.4939	0.6724
x_1, x_2, x_3, x_4	1338.0000	667.9000	0.6023	0.6770
x_1, x_2, x_3, x_5	1438.0000	699.2000	0.5406	0.6256
x_2, x_3, x_4, x_5	1347.8000	681.7000	0.5965	0.7928
x_1, x_3, x_4, x_5	1357.1000	678.9000	0.5909	0.6880
x_1, x_2, x_4, x_5	1561.1000	751.1000	0.4586	0.5461
x_1, x_2, x_3, x_4, x_5	1396.0000	705.6000	0.5671	0.6561

Table 1 and Table 2 showed the performance of MLR using different combination variables as input for MLR. The assessment of the performance MLR are based on RMSE, MAE, CE and r^2. From Table 1 for T=10-year, the best MLR performance is when using elevation, longest drainage path and mean catchment as input for MLR. The RMSE, MAE, CE and r^2 obtained are 746.2000, 383.1000, 0.7321 and 0.8697. The RMSE and MAE are the smallest compare to others and the CE and r^2 close to 1. From Table 1 also, there are several variables when implement in MLR the prediction produce a negative value of CE. The negative value of CE indicated the mean of the observed are better than prediction. From Table 2 for T=100-year, the best MLR performance is also the same with T=10 year, and that is elevation, longest drainage path and mean catchment slope as input for MLR. The RMSE, MAE, CE and r^2 are 1443.7000, 710.3000, 0.5370 and 0.7292. The RMSE is the smallest but for MAE it was the second smallest. For CE and r^2 it was the most closed to 1 compare to others. In overall, a conclusion can be reached such that the variables affect the performance of MLR in estimating flood qunatiles at ungauged sites. The best input for MLR through this study is combination of three variables that are elevation, longest drainage path and mean catchment slope.

5 Conclusions

There are many physiographical and hydrological characteristics exist at ungauged site that used in estimating flood quantile. Although there are a lot of characteristics can be used, not all of the characteristics are useful for estimating the flood quantiles at ungauged sites. In this study, five physiographical and hydrological characteristics were implemented. From five variables, total 31 combinations of variables used as input for MLR model. From the result obtained the suitable characteristics used as input for MLR model are elevation, longest drainage path and mean catchment. This characteristic is suitable only estimating flood quantile for ungauged site located at Peninsular Malaysia only. Although there are exist other catchments characteristics but from this study there are only some catchment characteristics is suitable use as input to estimate flood quantile.

Acknowledgment

The authors are thankful to Ministry of Higher Education and Ministry of Science, Technology and Innovation (MOSTI) for the financial support for this research under vot 4F088. High appreciation is also given to Department of Irrigation and Drainage, Ministry of Natural Resources and Environment, Malaysia for providing the stream flow data to complete the research.

References

[1] Besaw, L. E., Rizzo, D.M., Bierman P.R. and Hackett W.R. Advances in ungauged streamflow prediction using artificial neural network. *Journal of Hydrology*. 2010. 386: 27-37.
[2] Seckin, N. Modelling flood discharge at ungauged sites across Turkey using neuro fuzzy and neural network. *Journal of Hydroinformatics*, 2011. 13(4): 842 - 849.
[3] Pandey, G. R. and Nguyen, V.-T.-V. A comparative study of regression based methods in regional flood frequency analysis. *Journal of Hydrology*. 1999. 225: 92-101.
[4] Reed, D. W. and Robson, A.J. *Flood Estimations Handbook*. Centre for Ecology and Hydrology, UK. 1999.
[5] Hosking, J. R. M. and Wallis, J.R. *Regional frequency analysis: An approach based on L-Moments*, Cambridge, University Press, UK. 1997.

[6] Jain, A. a. Kumar, A.M. Regional flood frequency analysis at ungauged sites using the adaptive neuro-fuzzy inference system. *Journal of Hydrology*. 2008. 349: 31-43.
[7] McIntyre N, L. H., Wheater H. S., Young A and Wagener T. Ensemble predictions of runoff in ungauged catchments. *Water Resource Research*. 2005. 41.
[8] Shu, C. Ourda, T. b. M. J. Regional flood frequency analysis at ungauged sites using the adaptive nuero-fuzzy inference system. *Journal of Hydrology*. 2008. 349: 31-43.
[9] Vogel, R. M. and Kroll, C. N. Generalized low-flow frequency relationships for ungauged sites in Massachusetts. *American Water Resources Association*. 1990. 26(2): 241-253.
[10] Tasker, G. D., Hodge, S.A. and Barks, C.S. Region of influence regression for estimating the 50-year flood at ungaged sites. *American Water Resources Association*. 1996. 32(1): 163-170.
[11] Dawson, C. W., Abrahart, R. J., Shamseldin, A. Y. and Wilby, R. L. Flood estimation at ungauged sites using artificial neural networks. *Journal of Hydrology*. 2006. 319: 319-409.
[12] Thomas, D. M., Benson, M. A. Generalization of streamflow characteristics from drainage-basin charateristics. *US Geological Survey, Water Supply Paper*. 1970.
[13] Fennessey, N. and Vogel, R. M. Regional flow-duration curves for ungauged sites in Massachusetts. *Journal Water Resource Planning Management*. 1990. 116(4).
[14] Mosley, M. P., and McKerchar, A. "*Streamflow''Handbook of hydrology*, New York, McGraw-Hill. 1993.
[15] Pires J.C.M., M. F. G., Sousa S.I.V., Alvim-Ferraz M.C.M. and Pereira M.C. Selection and validation of parameters in multiple linear and principal component regressions. *Journal Environmental Modeling & Software*. 2007. 23: 50-55.
[16] Department of Irrigation and Drainage (DID). Magnitude and frequency of floods in Peninsular Malaysia. Hydrological Procedure No. 4, Ministry of Agriculture, Kuala Lumpur, Malaysia.