A Monthly Rainfall–Runoff Modelling using Multiple Linear Regression and Artificial Neural Network Techniques

Bachir Sakaa [1,2], Nabil Brahima[3], Hicham Chaffai[2], and Azzeddine Hani [2]

[1] Scientific Research Center on Arid Regions CRSTRA. BP 1682 RP Biskra. 07000, Algeria.
[2] Laboratory of Water Resource and Sustainable Development (REED), Annaba University, Algeria.
[3] Laboratoire des réservoirs souterrains pétroliers, gaziers et aquifères
sakaabachir@yahoo.fr hichamchaffai@yahoo.fr haniazzedine@yahoo.fr nabilbra@yahoo.fr

Abstract—The study presents multiple linear regression MLR coupled with multilayer perceptron MLP for predicting monthly runoff (Q_t). The data set including total 348 data records is divided into two subsets, training and testing. Various models depending on the combination of antecedent values of monthly runoff and rainfall are constructed and the best fit input structure is examined. The performance of models in training and testing phases are compared with the observed monthly runoff values to select the best fit forecasting model. For this purpose, some performance criteria such as Root Mean Square Error (RMSE) and coefficient of determination (R^2) are evaluated for different models (MLR, MLP_BFGS and MLP_SCG). The results indicated that MLP_BFGS outperforms all other models (MLP_SCG and MLR) in the forecasting of monthly runoff.

Key-Words— MLR, BFGS algorithm, SCG algorithm, monthly runoff, modeling

I. INTRODUCTION

The modeling of the hydrological behavior of watersheds is essential therefore that we are interested in the issues concerning the management of water resources, land, or one of the various aspects of the hydrological risk. Most hydrological processes are marked by questions of physical nonlinearity and uncertainties in parameter estimates [1]. This non-linearity makes forecasting floods an exercise far from obvious.

Many models have been created to simulate the nonlinear rainfall-rainoff relationship as well as to predict flows. These models are classified as deterministic or stochastic [2]. The deterministic approach is based on the calculation of the physical processes of the hydrological system. This calculation is often limited by the lack of data related to the physical processes of the hydrological system on the first hand and by the limited scientific knowledge of the natural systems on the other hand. These limitations negatively influence hydrological modeling and flood forecasting, degrading the reliability of the models carried out, especially as the specialists in the field demand precise results associated with a minimal calculation time. These points are surpassed by the stochastic approach that functions as a black box regardless of the internal structure of the system. This approach characterizes rainfall-rainoff relationships by analyzing time series without the use of physical data in the watershed. The use of the artificial neural network in hydrological modeling falls into this category of black box models [3], it produces accurate calculations, but without any understanding of the internal structure of the basin. This technique positively influences the prediction of flows using a minimum of parameters, requiring a short calculation time and producing more accurate results.

The application of ANN in hydrological modeling has been discussed by the American Society of Civil Engineers (ASCE) in a working committee on the application of ANN in Hydrology [4]. The

main advantages of ANN modeling are its nonparametric nature and its simple adaptation to data of different types.The study described in this paper is to apply an artificial neural network ANN to monthly rainfal-runoff modeling in Wadi Khmakham basin.

II. Multiple Linear Regression Model (MLR)

The objective of MLR analysis is to study the relationship between several independent or predictor variables and a dependent or criterion variable. The assumption of the model is that the relationship between the dependent variable Y_i and the p vector of regressors X_i is linear. The following represents a MLR equation [5]:

$$Y = a + \beta_1 X_1 + \dots\dots\dots + \beta_k X_k \quad (1)$$

where a is the intercept, β is the slope or coefficient, and k is the number of observations. For forecasting purposes, the linear regression equation will fit a forecasting model to an observed data set of Y and X values. The fitted model can be used to make a forecast of the value of Y with new additional observed values of X.

III. Artificial Neural Network (ANNs)

Use The most popular neural network model is the Multilayer Perceptron (MLP). The MLP is a layered feed forward network, which is typically trained with BFGS back propagation (Broyden Fletcher Goldfarb Shanno Quasi-Newton) and SCG back propagation (Scaled Conjugate Gradient). The number of neurons in a hidden layer is decided after training and testing. Multi layered network, trained by back propagation [6] are currently the most popular and proven [7] and has been used in this study. In the MLP, the neurons are organized in layers, and each neuron is connected only with neurons in contiguous layers. The input signal propagates through the network in a forward direction, layer by layer. The mathematical form of a three-layer feedforward ANN is given as [8]

$$O_k = g_2\left[\sum_j V_i w_{ji} g_1\left(\sum_i w_{ji} I_i + w_{jo}\right) + w_{ko}\right] \quad (2)$$

Where I_i is the input value to node i of the input layer, V_j is the hidden value to node j of the hidden layer, and O_k is the output at node k of the output layer. An input layer bias term I_0 =1 with bias weights w_{j0} and an output layer bias term V_0 = 1 with bias weights w_{k0} are included for the adjustments of the mean value at each layer.

The performance of each of the selected models (MLR, SCG & BFGS) was determined using the criteria, such as the Root Mean Square Error (RMSE), the coefficient of determination (R^2).

IV. THE STUDY AREA AND DATABASE

In this study, the monthly flow data of Khmakham Station on Saf-Saf river basin in the Eastern region of Algeria were used. The location of Wadi Khmakham is shown on Figure1.

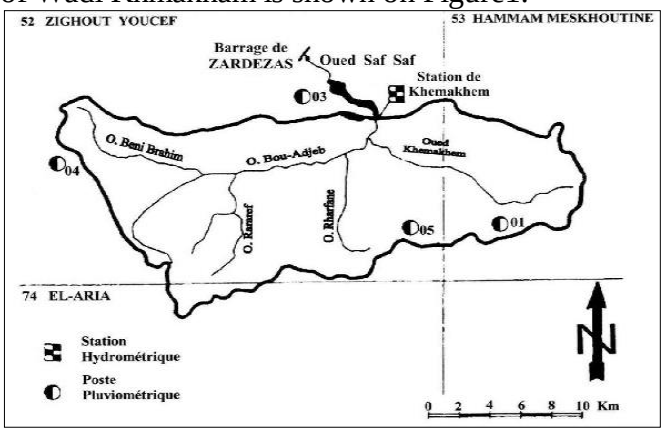

Figure1. Geographical location of study area

The monthly statistical parameters of the rainfall and runoff data for the Wadi of Khémakham are given in table 1. In the table, the X_{mean}, *SD*, *Skewness*, *Cv*, X_{min}, X_{max} denote the mean, standard deviation, skewness, coefficient of variation, minimum and maximum, respectively. In the calibration flow data, X_{min} and X_{max} values for runoff fall in the ranges 0 –19.18 mm/month for the Khémakham station. However, the testing flow dataset extremes are X_{min}=0, X_{max}=12.40 mm/month.

Table1. Statistical parameters of the rainfall and runoff data

	Rainfall (mm)			Runoff (m^3s^{-1})		
	Training	Testing	All data	Training	Testing	All data
X_{mean}	53.768	42.764	50.479	0.974	0.908	0.95448
X_{min}	0.000	0.00	0.00	0.00	0.00	0.000
X_{max}	452.50	225.300	452.50	19.180	12.40	19.1800
SD	55.832	42.676	52.431	2.734	2.061	2.282
Skew	2.546	1.728	2.483	4.526	3.543	4.328
Cv	103.838	99.794	103.865	243.632	226.973	239.080

V. RESULTS AND DISCUSSION

The inputs and outputs of the data sets were normalized for performance improvement of the model.

In this study, we built various models based on different combination of input variables and compared their RMSE and R^2 values so as to estimate the degree of effect of each input variable on the Qt. Five models were created and compared. The five models are the two-variable input vector model (P_t, Q_{t-1}); the three-variable input vector model (P_t, P_{t-1}, Q_{t-1}); the four-variable input vector model (P_t, P_{t-1}, Q_{t-1}, Q_{t-2}); and the five-variable input vector model (P_t, P_{t-1}, P_{t-2}, Q_{t-1}, Q_{t-2}, respectively (Table 2).

Table 2 R^2 and RMSE Statistics of Each Model in Training period

	RMSE			R^2		
Model imputs	MLR	SCG	BFGS	MLR	SCG	BFGS
P_t, Q_{t-1}	0.801	0.867	0.721	0.794	0.741	0.812
P_t, P_{t-1}, Q_{t-1}	0.706	0.711	0.650	0.812	0.749	0.836
P_t, P_{t-1}, Q_{t-1}, Q_{t-2}	0.658	0.650	0.575	0.831	0.827	0.891
P_t, P_{t-1}, P_{t-2}, Q_{t-1}, Q_{t-2}	0.606	0.514	0.438	0.904	0.942	0.964

The main data set is divided into two sub-data sets: (i) a training set and (ii) a testing set. Among the 348 data, 278 input-output pairs (80 %), randomly chosen from the data sequence, were used in the training set and the remaining 70 data (20 %) of the available data set were reserved for testing the developed models.

The results obtained from the best model for each type of forecasting method are presented in Table 2, and the variables for the best model of each forecasting method are shown in Table 3. All the models were developed in the same way via an iterative procedure involving successively adding variables and keeping them if they improved the forecasting performance.

Table 3 R^2 and RMSE Statistics of Each Model in Testing period

	RMSE			R^2		
Model imputs	MLR	SCG	BFGS	MLR	SCG	BFGS
P_t, Q_{t-1}	0.861	0.871	0.811	0.776	0.721	0.847
P_t, P_{t-1}, Q_{t-1}	0.746	0.728	0.755	0.824	0.732	0.912
P_t, P_{t-1}, Q_{t-1}, Q_{t-2}	0.688	0.692	0.588	0.850	0.801	0.954
P_t, P_{t-1}, P_{t-2}, Q_{t-1}, Q_{t-2}	0.606	0.544	0.468	0.937	0.951	0.970

All the models have the minimal root mean square error RMSE for the six-variable input vector model M (iv).in addition, the maximum coefficient of determination R^2 in the testing phase of M (iv) was obtained in BFGS medel (5-9-1), while other best models were obtained in SCG model (5-10-1) and MLR, respectively.

The best performances between the three models (BFGS, SCG and MLR) were compared; the optimal results of RMSE, and R^2 were obtained in BFGS model (5-9-1) not only in the training phase but also in the testing phase (Table 2). Table 2 shows that the minimum RMSE (0.438 mm/month) and the maximum R^2 (0.964) in the training phase were observed from BFGS model (5-9-1). The best performance in the testing phase (RMSE = 0.468 mm/month and R^2 = 0.970) was recorded from BFGS model (5-9-1). As the model was purposed for rainfall-runoff modelling, the performance in testing phase is the crucial index for selecting models. Therefore, BFGS model (5-9-1) with architecture (5-9-1) was selected as the best fit rainfall-runoff prediction model for Saf-Saf river basin.

Figure 2 shows the time-series graphical plots of both the observed Q_t values and the best-fit Q_t values by BFGS, SCG, and MLR models of the training and testing phases. It is clearly seen from the graphs that the predicted Q_t values of BFGS model is closer to the corresponding observed Q_t value than the value of the SCG and MLR models.

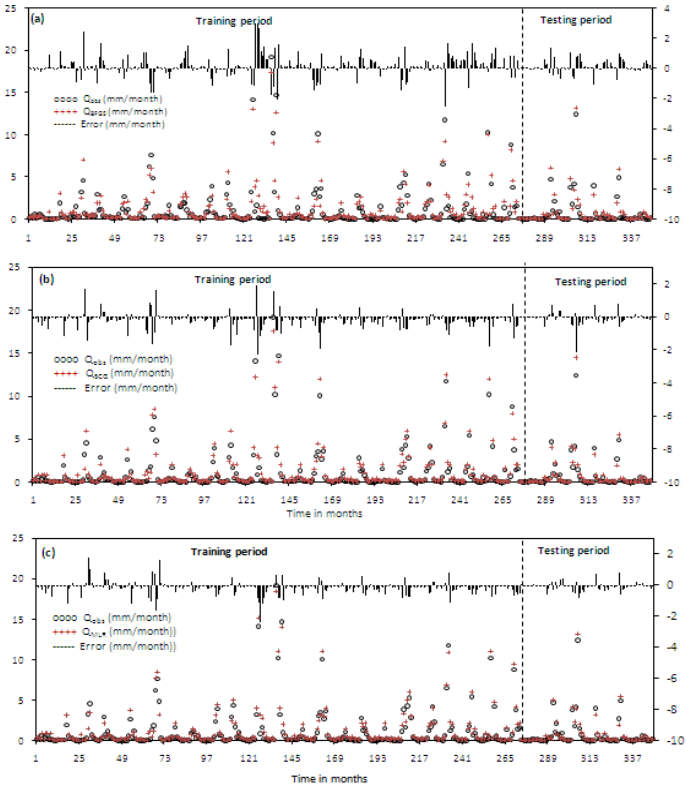

Figure 2 Monthly observed vs simulated discharge over the training and testing periods, (a): BFGS model, (b): SCG model, (c): MLR model.

The scatterplots of the observed versus predicted value of Q_t of the BFGS, SCG and MLR models analyzed herein are shown in Figure 3 for the training and testing phases, respectively. This figure nicely demonstrate that for all phases (training and testing), (i) the models' performances are, in general, accurate; (ii) the BFGS model is closer to the exact fit line than those of the SCG and MLR; (iii) SCG model is consistently superior to MLR model.

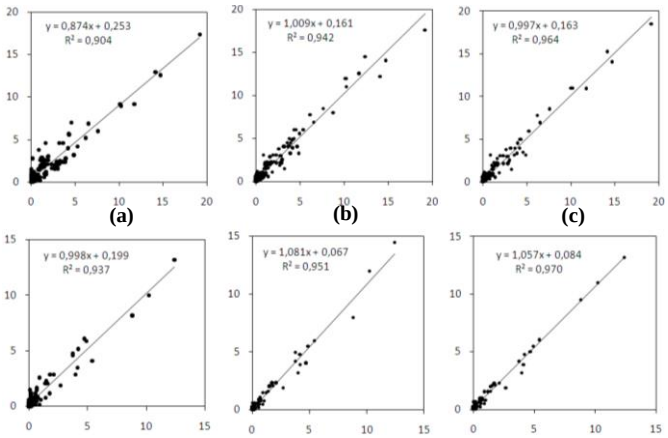

Figure 3 Scatterplots of predicted versus observed Q_t (in mm/month) in training and testing phases for (a): MLR, (b): SCG and (c) BFGS models.

VI. Conclusion

In this research, we developed a methodology based on the combination of ANNs and MLR to simulate Q_t for wadi Khmakham. Three different models were trained and tested. The predictive capability of the three models is determined using two criteria namely RMSE and R^2. From the results comparing the performance of the three different models (BFGS, SCG and MLR), it shows that BFGS model perform well than the other models (SCG and MLR). The results demonstrate that BFGS model have great potential to forecast Q_t when it is difficult to acquire the monitoring data.

References

[1] Birikundavyi S., Labib R., Trung H.T., Rousselle J. (2002). Performance of neural networks in daily stream flow forecasting. Journal of Hydrological Engineering 7 (5): 392 – 398.

[2] Lauzon, N., Rousselle, J., Birikundavyi, S., Trung, H.T. (2000). Real-time daily flow forecasting using black-box models, diffusion processes, and neural networks. Can. J. Civil Engng 27, 671 – 682

[3] Dawson C.W., and Wilby R.L., (2001) Hydrological modeling using artificial neural networks, Progress in Physical Geography, vol. 25 (1), pp. 80–108.

[4] ASCE Task Committee on Application of Artificial neural networks in Hydrology (2000). Artificial neural networks in hydrology (I & II, Journal of Hydrologic Engineering, 5(2), pp. 115-136.
[5] Pedhazur E.J. (1982). Multiple Regressions in Behavioral Research: Explanation and Prediction, Holt, Rinehart and Winston, New York.
[6] Rumelhart D.E., Hmton G.E., & Williams R.J. (1986a). Learning representations by error propagalion. In I). E. Rumelhart, J. L. McClelland and the PDP Research Group (Eds.).Cambridgc, MA: MIT Pres
[7] Ozbek F.S., and Fidan H. (2009). Estimation of pesticides usage in the agricultural sector in Turkey using artificial neural network (ANN), J. Animal Plant Sci., 4(3), 373–378.
[8] Hagan M.T., Demuth H.B., Beale M.H. (1996). Neural Network Design. PWS Publishing Company, Boston, MA.