Unmeasured Hydrologic Data Estimation by Artificial Intelligence Methods

M. Erkan Turan and M. Ali Yurdusev

Celal Bayar University Civil Engineering Department Muradiye, Manisa, TURKEY
e-mail: erkanturan1982@yahoo.com and yurdusev@bayar.edu.tr

Abstract

Water recourses management has become more crucial by the increase of the water consumption. An effective management relies on accurate and complete information about the river on which a project will be constructed. Artificial intelligence techniques are often and successfully used to complete the unmeasured data. In this study, feed forward back propagation neural networks, generalized regression neural network, fuzzy logic and regression analysis are used or unmeasured data estimation using the data of the four runoff gauge station on the Birs River in Switzerland. The performance of these models are measured by the mean square error, determination coefficients and efficiency coefficients to chose the best fit model.

Key words: *Artificial Neural Networks, Fuzzy Logic, River flow forecasting*

Introduction

Analyzing streamflow records can give significant ideas for both past and future characteristics of streamflows. Therefore, recording and analyzing streamflow measurements have highly important roles in planning, designing and management of water resources. For this purpose, engineers and scientists are interested in reconstructing unmeasured streamflow data as well as future data of streamflow series. For this purposes, in addition to the conventional time series approaches, Artificial Intelligence (AI) techniques such as Artificial Neural Networks (ANN) and Fuzzy Logic (FL) have been preferred for some time for a variety of water resources engineering applications [Tayfur et al. (2003), Abrahart et al. (2001), Imrie et al. (2000), Cığızoğlu (2005), Keskin et al. (2005), Sen et al. (2006), Kişi et al. (2003)]

This paper describes a case study on the use of various AI techniques for the prediction of the river flows at one location from the upstream flow records as such an application is rare in literature. For this purpose, the flows of Birs river in Switzerland recorded at Soyhiéres flow gauge station (2478) are predicted from the values of the three upstream stations; Moutier (2122), Delémont(2479) and Vicques(2610). The methods used in this study include two types of ANN techniques, Feed Forward Back Propagation Algorithm (FFBP) and Generalized Regression Neural Networks (GRNN), and Fuzy Logic. The flow values measured between 1995 and 2002 are used to train and test the models. The best model is sought using the performance measures including mean square error, determination coefficient and efficiency coefficient.

The methods used

Feed Forward Back Propagation Algorithm (FFBP)

Given a training set of input-output data, the most common learning rule for multi-layer perceptrons is the back-propagation algorithm. Back propagation involves two phases: a feed-forward phase in which the external input information at the input nodes is propagated forward to compute the output information signal at the output unit, and a backward phase in which modifications to the connection strengths are made based on the differences between the computed and observed information signals at the output units (Eberhart and Dobbins, 1990). A schematic diagram of a FFBP architecture is presented in Fig.1.

Generalized Regression Neural Networks (GRNN)
The generalized regression neural network proposed by Specht (1991) does not require an iterative training procedure as in the back-propagation method, but approximates any arbitrary function between the input and output vectors, drawing the function estimate directly from the training data. Furthermore, it is consistent; that is, as the training set size becomes large, the estimation error approaches zero, with only mild restrictions on the function. A schematic diagram of GRNN architecture is presented in Fig.2.

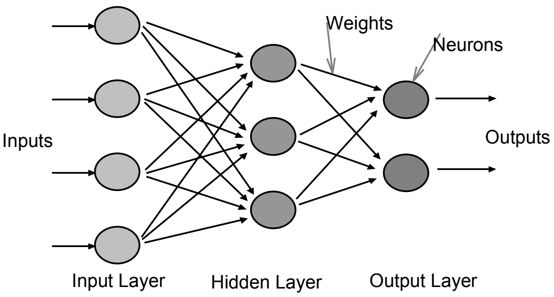

Figure 1. Schematic diagram of feed forward back propagation network architecture

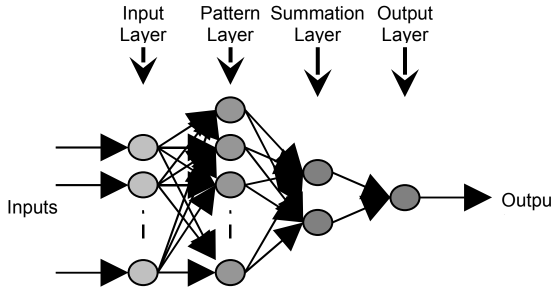

Figure 2. Schematic diagram of generalized regression neural network architecture

Fuzzy Logic [FL]
Fuzzy systems fake the linguistic and verbal information as data and provide the answer accordingly in a vague manner, which includes the crisp desired solutions. This brings to the mind that although the basis of the methodology is fuzzy, the results it yields are precise (Sen, 2004). A general fuzzy system, as shown in Fig. 3, has the components of fuzzification, fuzzy rule base, fuzzy inference engine, and defuzzification. *Fuzzification* converts each piece of input data to degrees of membership by a look-up in one or more several membership functions. In this study, the rule base is constructed by the table-lookup method.

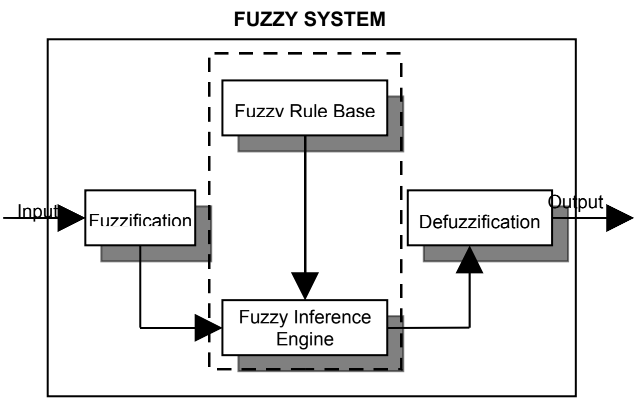

Figure 3. Schematic diagram of fuzzy system

Study area and the data used

In this study, the daily flows of Birs river in Switzerland recorded at Soyhiéres flow gauge station (2478) are predicted from the values of the three upstream stations; Moutier (2122), Delémont(2479) and Vicques(2610). The 2922 flow values measured between 1995 and 2002 are used in building the models. The river Birs is selected because there has been no river regulation work in place yet. The river is a tributary of the River Ren with 73 km length and 924 km^2 drainage area. The statistical characteristics of the measurements used are given in Table 1.

Prediction of river flows from upstream flow records

The daily flow values of the River Birs, Switzerland, recorded at Soyhiéres flow gauge station (2478) are predicted from the values of the three upstream stations; Moutier (2122), Delémont(2479) and Vicques(2610). Seven model structures are experimented to find the best model form to predict the river flows from upstream flows. The models used are described in Table 2. From the observation interval considered, 2192 values between 01.01.1995 and 31.12.2000 are used to train the models whereas 730 values between 01.01.2001 and 31.12.2002 to test the models.

Table 1. The statistical characteristics of the measurements used

Station	Observation interval	Minimum value $x_{\min}$ (m^3/s)	Maximum value $x_{\max}$ (m^3/s)	Mean $\bar{x}$ (m^3/s)	Standard deviation S_x (m^3/s)	Skewness coefficient C_{sx}
2122	01.01.1995–31.12.2002	0.73	29.9	3.40	3.07	3.41
2479	01.01.1995–31.12.2002	0.58	46.4	4.58	4.28	2.63
2610	01.01.1995–31.12.2002	0.20	40.2	1.69	2.32	6.75
2478	01.01.1995–31.12.2002	2.29	128	11.73	10.94	3.09

Table 2. The model structures experimented

Model	Inputs	Output
M1	$Q_{2122}(t)$	$Q_{2478}(t)$
M2	$Q_{2479}(t)$	$Q_{2478}(t)$
M3	$Q_{2610}(t)$	$Q_{2478}(t)$
M4	$Q_{2122}(t)+Q_{2479}(t)$	$Q_{2478}(t)$
M5	$Q_{2122}(t)+Q_{2610}(t)$	$Q_{2478}(t)$
M6	$Q_{2479}(t)+Q_{2610}(t)$	$Q_{2478}(t)$
M7	$Q_{2122}(t)+Q_{2479}(t)+Q_{2610}(t)$	$Q_{2478}(t)$

The flow values used in ANN modellings were normalized between -0.9 and 0.9 based on the network structure used in Matlab using the Eqn.1 whereas the normalization between 1 and 3 in FL models was undertaken by the Eqn.2.The aim of such transformation is to be able to compare the different quantities.

$$z_i = 1.8\left[\frac{x_i - x_{\min}}{x_{\max} - x_{\min}}\right] - 0.9 \tag{1}$$

$$x_{nor} = \left[\left(\frac{2\left(x_i - x_{\min}\right)}{x_{\max} - x_{\min}}\right) - 1\right] + 2 \tag{2}$$

The performance measures; determination coefficient8 R^2), mean square error (MSE) and efficiency coefficients (E, E_1 and E_2) are defined in Eqns. 3, 4 and 5.

$$R^2 = \left[\frac{\sum_{t=1}^{N} \left(Q_g(t) - \overline{Q_g} \right) \left(Q_h(t) - \overline{Q_h} \right)}{\sqrt{\sum_{t=1}^{N} \left(Q_g(t) - \overline{Q_g} \right)^2 \left(Q_h(t) - \overline{Q_h} \right)^2}} \right]^2 \tag{3}$$

$$OKH = \frac{\sum_{t=1}^{N} \left(Q_g(t) - Q_h(t) \right)^2}{N} \tag{4}$$

$$E = \frac{E_1 - E_2}{E_1} \qquad E_1 = \sum_{t=1}^{N} \left(Q_g(t) - \overline{Q_g} \right)^2, \qquad E_2 = \sum_{t=1}^{N} \left(Q_g(t) - Q_h(t) \right)^2 \tag{5}$$

where Q_g are the recorded flow values, Q_h are the predicted flow values, $\overline{Q_g}$ and $\overline{Q_h}$ are the mean values.

Feed Forward Back Propagation (FFBP) Models

All 7 models have been designed to have one input layer, one output layer and one hidden layer. As can be seen from Table 1, the output layer in each model contains one node, which is the flow value at the station #2478. The input layers contain varying number of nodes depending on the number of upstream stations considered. The determination of number of nodes in hidden layers is the issue of trial, which has been verified by altering the iteration number to achieve the best performance values (MSE, R^2 and E) for both training and testing phases. The activation function used is sigmoid function. The outcome of trials for each model, the best network structure, is given in Table 3. Table 3 also includes the values of performance measures for training and testing for each model. As seen from Table 3, the best fit model for predicting the flow values of the station #2478 is M7 model with the lowest MSE value and the highest R^2 and E values for both training and testing. The model consists of 3 input nodes and 6 nodes in the hidden layer. The graphical representations of M7 model are given in Figure 4 for training and in Figure 5 for testing.

Table 3. The performance values of FFBP models

Model	Number of nodes in hidden layer	MSE for training	MSE for testing	R^2 for training	R^2 for testing	E for training	E for testing
M1	2	9.564	16.844	0.909	0.923	0.909	0.893
M2	3	2.966	17.060	0.972	0.960	0.972	0.892
M3	3	10.110	31.753	0.903	0.874	0.903	0.798
M4	5	1.973	11.861	0.981	0.971	0.981	0.925
M5	4	4.581	15.286	0.956	0.945	0.956	0.903
M6	5	1.491	2.821	0.986	0.984	0.986	0.982
M7	**6**	**1.111**	**1.858**	**0.989**	**0.989**	**0.989**	**0.988**

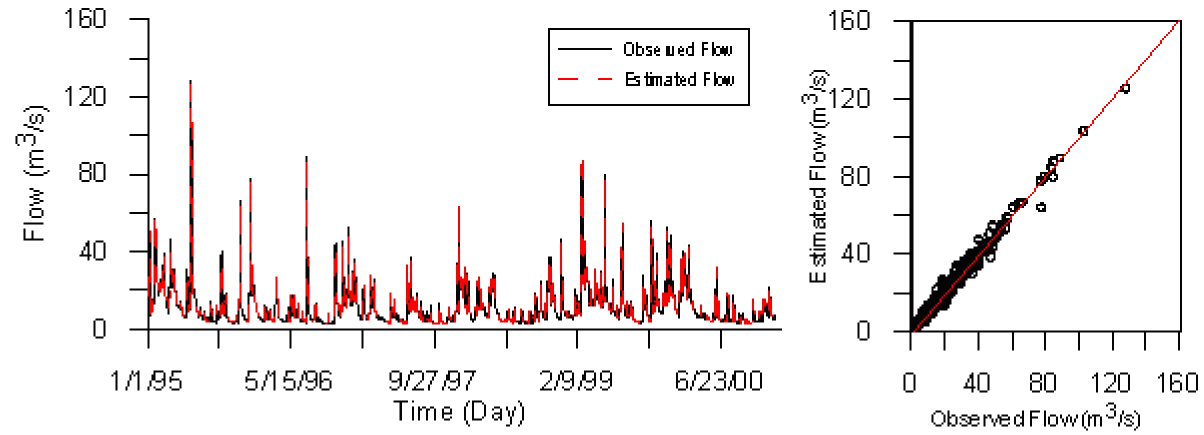

Figure 4. Training Set Performance of FFBP for M7 Model

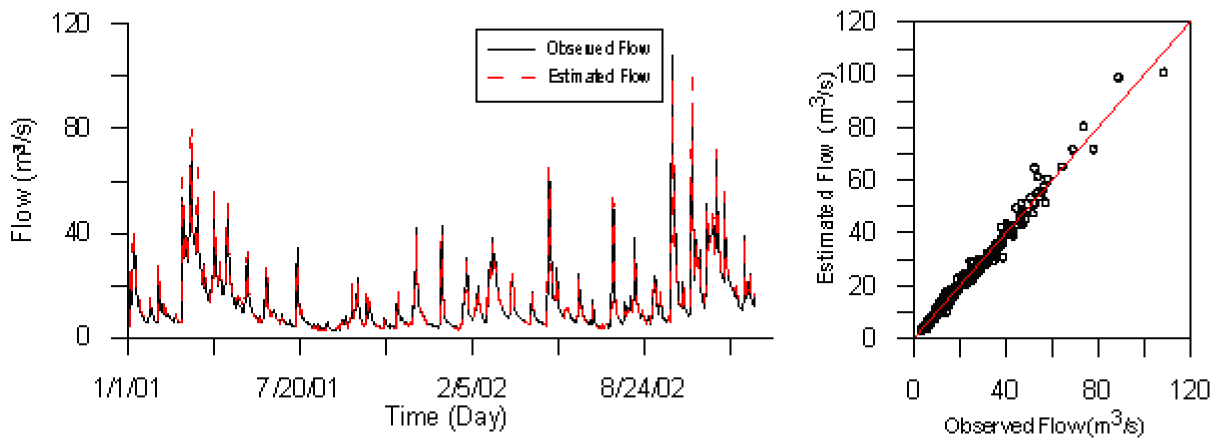

Figure 5. Test Set Performance of FFBP for M7 Model

Generalized Regression Neural Networks (GRNN) Models

The best fit model in this type of ANN is sought to assess the correction parameter which will yield the most appropriate performance values. The results of Matlab trials for each of 7 models previously mentioned are given in Table 4 again for both training and testing. Comparing the performance values, the best fit model for this case is found to be M6. The graphical representations of M6 model are given in Figure 6 for training and in Figure 7 for testing.

Table 4. The performance values of GRNN models

Model	S Correction parameter	MSE for training	MSE for testing	R^2 for training	R^2 for testing	E for training	E for testing
M1	0.02	8.319	17.470	0.921	0.913	0.921	0.889
M2	0.02	2.774	15.201	0.974	0.963	0.974	0.904
M3	0.02	9.390	36.139	0.911	0.855	0.910	0.771
M4	0.03	1.410	9.225	0.987	0.959	0.987	0.941
M5	0.03	3.436	19.024	0.967	0.928	0.967	0.879
M6	**0.05**	**2.030**	**3.971**	**0.982**	**0.975**	**0.981**	**0.975**
M7	0.06	1.495	4.455	0.987	0.972	0.986	0.972

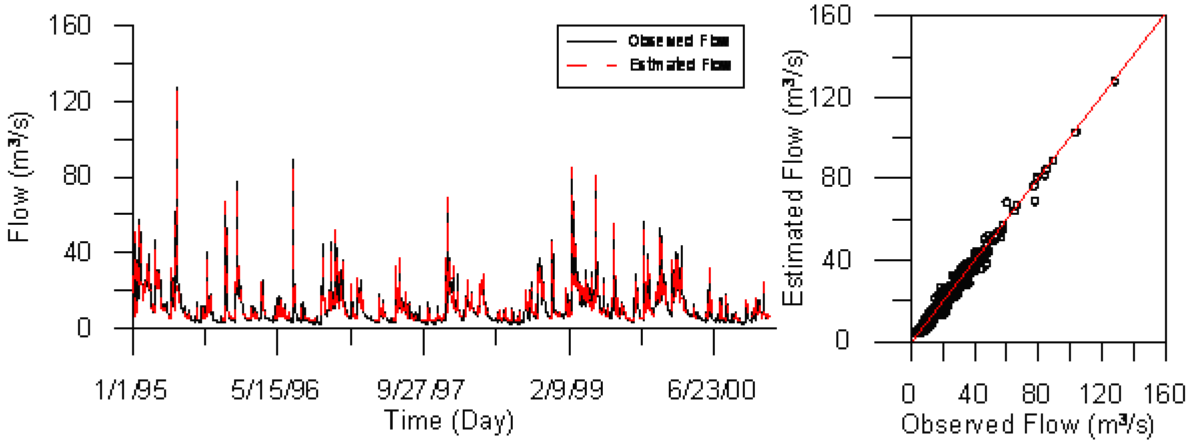

Figure 6. Training Set Performance of GRNN for M6 Model

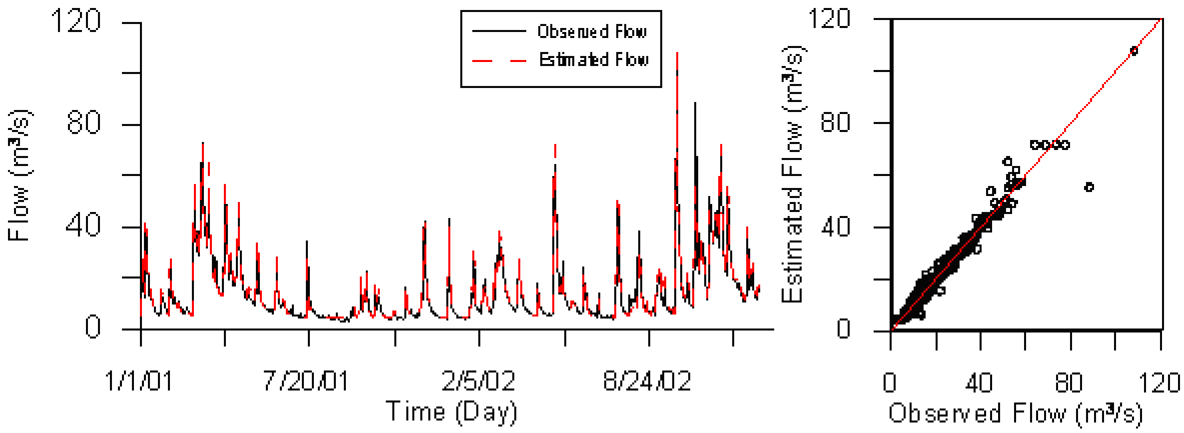

Figure 7. Test Set Performance of GRNN for M6 Model

Fuzzy Logic (FL) Models

The table-look up memory algorithm is used to construct the fuzzy system for the 7 models mentioned previously. The number of data division, N, was altered to obtain the most appropriate fuzzy system for each of 7 model. The Mamdani type inference model, triangle membership function, AND type fuzzy operator and the centroid defuzzification method were used for all fuzzy systems. The number of membership function and the performance values for training and testing obtained from a series of Matlab exercises for all models are given in Table 5. As for GRNN exercises, the model M6 is found to be the best fit model exhibiting the most satisfying performance values. The input-output data of fuzzy system constructed for M6 model includes 19 fuzzy sub-sets. The rule base of the system comprises 53 rules. The graphical representations of M6 model are given in Figure 8 for training and in Figure 9 for testing.

Conclusions

Various AI techniques are used for the prediction of the river flows at one location from the upstream flow records in the case of Birs River in Switzerland. The measured flow values at Soyhiéres flow gauge station (2478) are predicted from the values of the three upstream stations; Moutier (2122), Delémont(2479) and Vicques(2610). Three AI methods, Feed Forward Back Propagation Algorithm, Generalized Regression Neural Networks and Fuzy Logic, are used to achieve this end. The best fit models are selected based on the values of a series of performance measures; mean square error, determination and efficiency coefficient. Comparing the performance values of the 7 models based on the three methods mentioned, given in Tables 3, 4 and 5, the M7 model of FFBP algorithm can/should be selected over the other models if the flows of Soyhiéres station would be predicted. Based on the findings of this study, it can be concluded that the best method should be sought to model river flows rather than picking up a single method even if it is freshly developed.

Table 5. The performance values of FL models

Model	Number of membership function	MSE for training	MSE for testing	R^2 for training	R^2 for testing	E for training	E for testing
M1	17	18.176	23.967	0.885	0.898	0.826	0.848
M2	19	9.150	22.232	0.960	0.928	0.913	0.859
M3	15	30.814	31.661	0.866	0.832	0.706	0.799
M4	17	9.296	17.434	0.961	0.908	0.911	0.889
M5	19	11.002	20.441	0.922	0.916	0.895	0.870
M6	**19**	**7.100**	**15.034**	**0.970**	**0.917**	**0.932**	**0.905**
M7	17	8.872	16.234	0.965	0.910	0.915	0.897

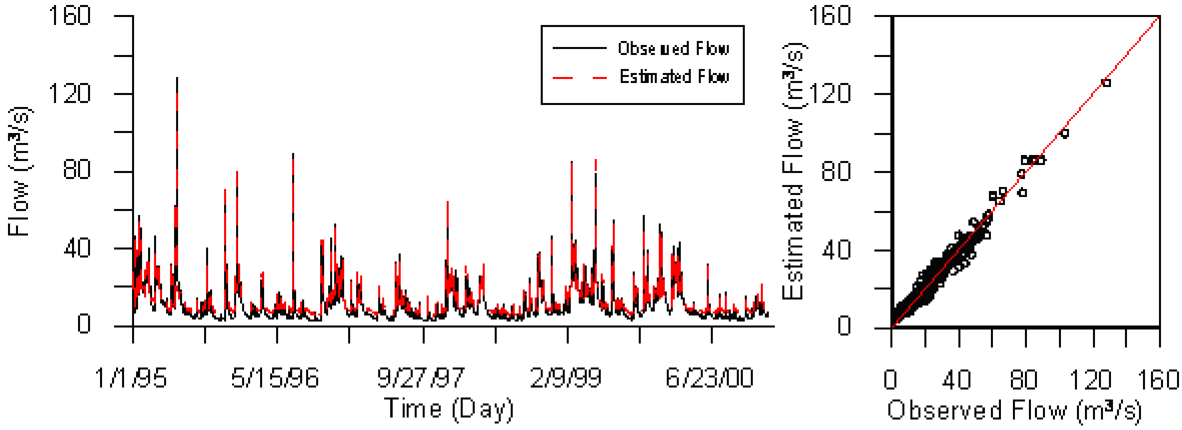

Figure 8. Training Set Performance of FL for M6 Model

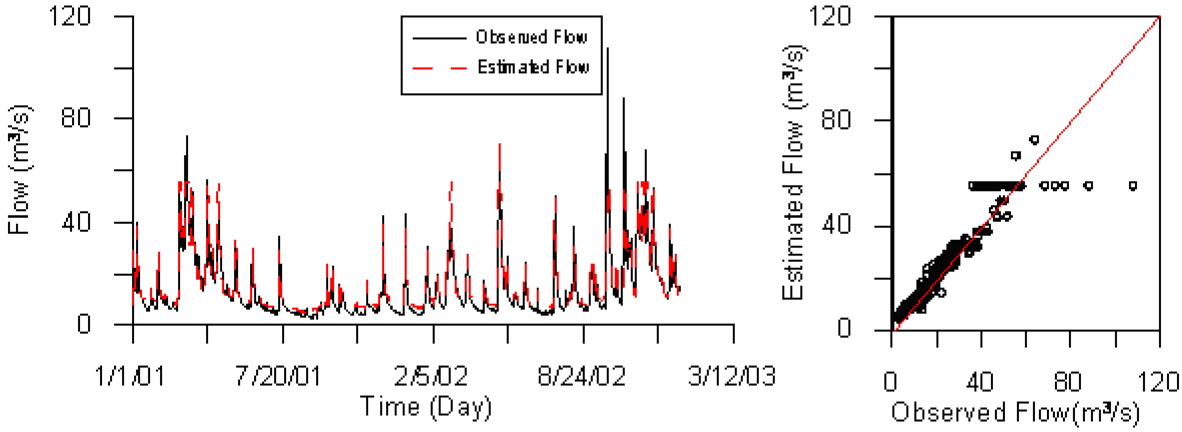

Figure 9. Test Set Performance of FL for M6 Model

REFERENCES

Abrahart, R.J., See, L. ve Kneal, P.E. (2001). Investigating the role of saliency analysis with neural network rainfall-runoff model, *Computers & Geosciences*, vol. 27, pp. 921-928.

Cığızoğlu HK, (2005). Application of generalized regression neural networks to intermittent flow forecasting and estimation. *Journal of Hydrologic Engineering*, vol. 10(4), pp. 336-341.

Eberhart, R.C. and Dobbins, R.W., (1990). *Neural Network PC Tools: A Practical Guide*. Academic Press, San Diego, 414 pp.

Imrie, C. E. Durucan, S. Korre, A. (2000), “River flow prediction using artificial neural networks: generalisation beyond the calibration range” *Journal of Hydrology*, Vol. 233, pp. 138-153, Elsevier Science B.V.

Keskin, M.E., Terzi, Ö., Taylan, D., (2005), Fuzzy Logic approach for daily pan evaporations. 2nd National Water Engineering Symposium, 673-681, IZMİR, Turkey (In Turkish).

Kişi, Ö., Karahan, M.E., Şen, Z., (2003). Suspended sediment modeling using fuzzy logic, *ITU Journal*, Vol: 2(3), p. 43-54 (In Tukish).

Şen, Z. (2004) *Fuzzy logic and system models in water sciences*. Turkish Water Foundation Publication, Istanbul, Turkey.

Specht, D.F.,(1991) A general regression neural network, IEEE Transactions on Neural Networks, 2, 6, 568-576.

Tayfur, G. , Ozdemir, S., Singh, V.P. (2003). Fuzzy Logic Algorithm for Runoff-induced Sediment Transport from Bare Soil Surfaces, Advances In *Water Resources*, vol. 26(12), pp. 1249-1256