Jurnal
Teknologi

Full Paper

NEURO-FUZZY SYSTEMS APPROACH TO INFILL MISSING RAINFALL DATA FOR KLANG RIVER CATCHMENT, MALAYSIA

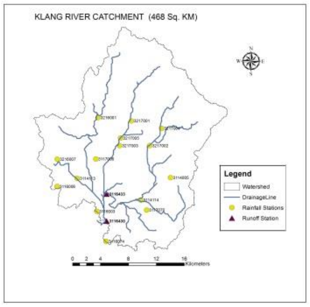
Article history
Received
24 May 2015
Received in revised form
20 November 2015
Accepted
22 December 2015

Nadeem Nawaz[a,c*], Sobri Harun[a], Rawshan Othman[b], Arien Heryansyah[a]

[a]Department of Hydraulics and Hydrology, Faculty of Civil Engineering, Universiti Teknologi Malaysia, 81310 UTM Johor Bahru, Johor, Malaysia
[b]Petroleum Department, Koya Technical Institute, Erbil Polytechnic University, 44001 Erbil, Kurdistan Regional Government, Iraq
[c]Faculty of Water Resources Management, Lasbela University of Agriculture, Water and Marine Sciences, 90150 Uthal, Balochistan, Pakistan

*Corresponding author
nn.engr96@yahoo.com

Graphical abstract



Abstract

Rainfall data can be regarded as the most essential input for various applications in hydrological sciences. Continuous rainfall data with adequate length is the main requirement to solve complex hydrological problems. Mostly in developing countries hydrologists are still facing problems of missing rainfall data with inadequate length. Researchers have been applying a number of statistical and data driven approaches to overcome this insufficiency. This study is an application of neuro-fuzzy system to infill the missing rainfall data for Klang River catchment. Pettitt test, standard normal homogeneity test (SNHT) and Von Neumann Ratio (VNR) tests were performed to check the homogeneity of rainfall data. The neuro-fuzzy model performances were assessed both in calibration and validation stages based on statistical measures such as coefficient of determination (R^2), Root Mean Square Error (RMSE) and Mean Absolute Error (MAE). To evaluate the performance of the neuro-fuzzy system model, it was compared with a traditional modeling technique known as autoregressive model with exogenous inputs (ARX). The neuro-fuzzy system model gave better performances in both stages for the best input combinations. The missing rainfall data was predicted using the input combination with best performances. The results of this study showed the effectiveness of the neuro-fuzzy systems and it is recommended as a prominent tool for filling the missing data.

Keywords: ANFIS, rainfall, missing data, neuro-fuzzy systems

1.0 INTRODUCTION

Hydrologists are always dealing with the problem of insufficient or missing hydrological time series data. The availability of continuous historical data is important to plan future use of available water resources and improving the calibration and validation of the hydrological models. Inadequate length and presence of gaps are the common deficiencies in hydrological data and usually observed in most of developing countries [1]. There may be a number of reasons for this deficit like faulty equipment, electric shortfall, human ignorance, shortage of finances and so on. Rainfall data has been used in a number of studies as it provides useful information for the hydrological modeling [2]. The presence of gaps in rainfall data is a common issue and hurdle in performing many hydrological studies. To infill the missing rainfall data a number of approaches can be found in literature such as artificial neural network [3, 4], inverse distance weighting method [5], regression method [6], simple arithmetic averages [7] and so on. In contrast with previously used techniques, this study is an application of neuro-fuzzy systems for filling missing rainfall data.

According to the Nauck [8] definition: "A hybrid neuro-fuzzy system is a fuzzy system that uses a learning algorithm based on gradients or inspired by the neural network theory (heuristical learning strategies) to determine its parameters (fuzzy sets and fuzzy rules) through the patterns processing (input and output)". In the parallel architecture of neuro-fuzzy systems, a neural network and a fuzzy logic-based system are integrated appropriately. In this architecture, a layer of hidden neurons correspond to each of the task of a fuzzy inference system (FIS). This allows visualizing the flow of data and error signals through the system. Several architectures have been addressed in literature including the adapted fuzzy inference system, method of implementation, and learning algorithm [9].

Neuro-fuzzy systems are generally classified into two main groups. The first group is linguistic neuro-fuzzy systems which employ Mamdani-type inference system [10] in their structures. In this group of neuro-fuzzy systems, linguistic output data is produced from linguistic input data. The second group is precise neuro-fuzzy systems, which employs Takagi-Sugeno-type inference system [11] and is able to produce numerical (non-linguistic) output from input data. Neuro-fuzzy systems have been widely used in time series modeling in hydrology such as rainfall-runoff simulation [12, 13], streamflow forecasting [14-16], water quality [17, 18], rainfall forecasting [19] and so on. The results of these studies have proven the promising potential of the neuro-fuzzy systems in prediction and simulation of hydrological time series. Adaptive network-based neuro fuzzy system (ANFIS) is one of the mostly used neuro-fuzzy modeling techniques and has its successful applications in diverse fields. ANFIS can also be used with an option for model validation as a check for over fitting. The objectives of this study were: (a) to infill the missing rainfall data for Klang River catchment in Malaysia; and (b) to check the capabilities of ANFIS for infilling rainfall data in a tropical catchment.

2.0 METHODOLOGY

2.1 ANFIS

Jang [20] implemented Takagi-Sugeno fuzzy rules by ANFIS and its architecture consists of five layers as illustrated in Figure 1. The actions of these layers are: Layer 1 generates fuzzy membership values for input variable; Layer 2 multiplies the incoming signals from the previous layer and calculate the firing strength of the rule (T-norm operation); Layer 3 computes the normalized firing strength; Node *k* in this layer 4 calculates the contribution of the *k*th rule in the model output based on first-order Takagi-Sugeno rules; and Layer 5 calculates the weighted global output of the system.

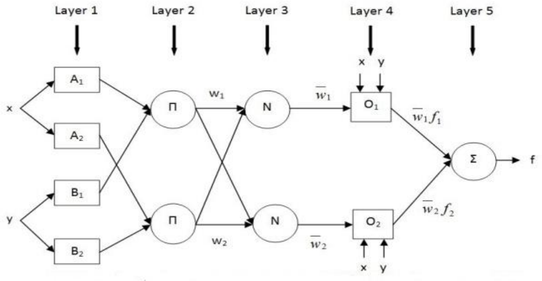


Figure 1 Architecture of ANFIS

The reason for selecting ANFIS to infill missing rainfall data was due to its capability of simulating complex input and output relationship. It uses a combination of the least-squares method and the back propagation gradient descent method for training FIS membership function parameters for a given training data set.

2.2 Study Area and Data Used

Klang river catchment is located in the central part of Peninsular Malaysia as can be seen in Figure 2. The catchment size is 468 km^2 and fully urbanized and densely populated as it surrounds the capital city of Malaysia. Heavy rainfall events are recorded in Malaysia because of its presence in tropical zone. Peninsular Malaysia receives approximately 2400mm rainfall per annum [21].

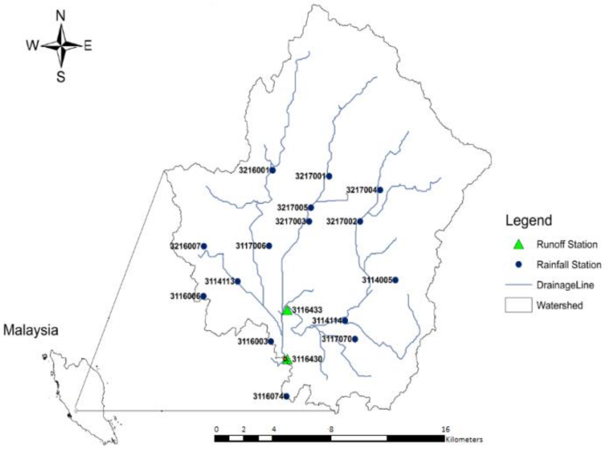


Figure 2 Map of Klang River catchment

The northeast monsoon contributes heavy rainfall events in the eastern part of Peninsular Malaysia that occur during November-February and the western part of Peninsular Malaysia receives southwest monsoon during May-August. Peninsular Malaysia receives the most rainy days in both monsoons. Klang River catchment also receives the inter-monsoon period during the months of March-April and September-October [22]. Flood events have also been recorded and proper planning of drainage remains the key concern during the heavy storm events. The rainfall data of the presently active rainfall stations was arranged from department of irrigation and drainage (DID), Malaysia.

The daily rainfall data of the fifteen stations were used in this study. The data provided by the DID was varied in terms of record length as some satiations were installed lately. Most of the stations had rainfall data from 2007 to 2013, so it was decided to perform this study to infill the missing rainfall record of the fifteen stations for the same period. Table 1 shows the detail of the stations with their geographical coordinates, record length and percentage missing data used for this study.

Homogeneity is important to check the variability in rainfall data as it is always affected by the changes made in measurement techniques and environmental characteristics. Homogeneity tests were performed on all daily rainfall data from the fifteen stations. Three commonly used approaches were adopted to perform homogeneity tests that included: (a) Pettitt test developed by [23]; (b) standard normal homogeneity test (SNHT) developed by [24]; and (c) Von Neumann Ratio (VNR) test developed by [25]. The performance of the tests was evaluated on annual mean and annual median.

After performing the homogeneity tests the correlation analysis were performed for each station having missing data with all other stations. It was found that there is good correlation between the neighboring rainfall stations for 3216007 and 3217005. On the other hand the correlation for station 3116074 was not strong compared with neighboring stations. This could be because the neighboring stations were at longer distance in relation to the others. Based on the correlation analyses results, the stations were selected to be used as input for developing the ANFIS model. Excluding the missing values the rainfall data was distributed for calibration and validation datasets.

Table 1 Locations, data record length and missing data periods of the rainfall stations in Klang River catchment

Station ID	Coordinates		Record length		Missing data		Missing data
	Latitude	Longitude	From	To	From	To	(%)
3114005	3.1947	101.7797	Jan, 2007	Dec, 2013	_	_	_
3114113	3.1938	101.6594	Jan, 2007	Dec, 2013	_	_	_
3114114	3.1660	101.7413	Jan, 2007	Dec, 2013	_	_	_
3116003	3.1514	101.6847	Jan, 2007	Dec, 2013	_	_	_
3116006	3.1833	101.6333	Jan, 2007	Dec, 2013	_	_	_
3116074	3.1126	101.6966	Jan, 2007	Dec, 2013	1/1/2011	8/2/2011	5.60
					11/3/2013	16/6/2013	
3117006	3.2189	101.6833	Jan, 2007	Dec, 2013	_	_	_
3117070	3.1531	101.7489	Jan, 2007	Dec, 2013	_	_	_
3216001	3.2722	101.6861	Jan, 2007	Dec, 2013	_	_	_
3216007	3.2186	101.6336	Jan, 2008	Dec, 2013	1/1/2007	31/12/2007	14.28
3217001	3.2681	101.7292	Jan, 2007	Dec, 2013	_	_	_
3217002	3.2361	101.7528	Jan, 2007	Dec, 2013	_	_	_
3217003	3.2361	101.7139	Jan, 2007	Dec, 2013	_	_	_
3217004	3.2583	101.7681	Jan, 2007	Dec, 2013	_	_	_
3217005	3.2458	101.7153	Jan, 2007	Dec, 2013	23/3/2008	7/6/2008	6.06
					25/5/2009	15/7/2009	
					1/10/2009	29/10/2009	

All data to be used in developing ANFIS model were normalized. The normalization procedure adopted in this study followed [26] which can be given by:

$$x_n = F_{min} + \left[\frac{x_i - x_{min}}{x_{max} - x_{min}}\right] \times (F_{max} - F_{min}) \quad (1)$$

where F_{min} and F_{max} are the required minimum and maximum of the new domain (e.g. 0.1-0.9); x_n is the normalized data; x_{min} and x_{max} are the minimum and maximum of the observed data respectively; and x_i is the observed data.

As different input combination of the neighbouring stations were selected based on the correlation analyses to find out appropriate antecedents for model development. Table 2 shows the detail of selected rainfall stations for input and datasets for calibration and validation. Table 3 shows the different input combinations used for developing ANFIS model with their performances in calibration and validation stages against them.

Table 2 Input stations for developing ANFIS model

Station ID	Training (days)	Validation (days)	Stations used as input
3116074	1690	724	3116003(4.511km), 3117070(7.35km), 3114114(7.74km)
3216007	2192	658	3217003(3.707km), 3217001(2.913km), 3217002(4.301km), 3117006(4.643km), 3216001(4.370km)
3217005	2402	721	3114113 (3.972km), 3116006(3.920km), 3117006(5.520km)

Table 3 ANFIS performances in calibration and validation stages

Station ID	Input Combinations	Calibration			Validation		
		R^2	MAE	RMSE	R^2	MAE	RMSE
3116074	3116003	0.89	2.49	4.97	0.91	2.67	5.23
	3117070	0.76	3.45	6.45	0.82	3.11	5.97
	3114114	0.81	3.17	6.34	0.84	3.02	5.92
	3116003, 3117070	0.91	2.42	4.83	0.93	2.43	5.02
	3116003, 3114114	0.87	2.6	5.12	0.84	2.98	5.72
	3114114, 3117070	0.83	3.03	6.02	0.85	2.79	5.67
	*3116003, 3114114, 3117070	0.94	2.27	5.12	0.95	2.21	4.93
3216007	3116006	0.93	2.27	4.91	0.94	2.12	4.53
	3117006	0.87	2.41	5.23	0.89	2.49	5.03
	3114113	0.92	2.23	4.63	0.91	2.35	4.89
	3116006, 3117006	0.89	2.37	5.11	0.91	2.32	4.67
	*3116006, 3114113	0.96	1.92	4.21	0.97	1.78	4.16
	3117006, 3114113	0.87	2.43	5.29	0.85	2.83	5.12
	3116006, 3117006, 3114113	0.95	2.18	4.42	0.95	1.96	4.32
3217005	3217001	0.89	2.82	4.39	0.91	2.23	4.76
	3217003	0.88	2.89	5.78	0.89	2.45	5.03
	3217001, 3217002	0.87	3.01	6.02	0.86	2.82	5.64
	3217001, 3217003	0.89	2.79	5.43	0.92	2.11	4.52
	3217001, 3217002, 3217003	0.91	2.44	5.02	0.92	2.13	4.59
	*3217001, 3217002, 3217003,3216001	0.95	2.02	4.34	0.95	1.98	4.29
	3217001, 3217002, 3217003, 3216006	0.92	2.33	4.79	0.91	2.22	4.63
	3217001, 3217002, 3216001, 31176006	0.93	2.23	4.65	0.94	2.04	4.36
	3217001, 3217003, 3216001, 31176006	0.87	2.92	5.81	0.89	2.49	5.12
	3217001, 3217002, 3217003, 3216001, 31176006	0.79	3.42	6.34	0.83	3.12	6.04

* Input combination with best performance

2.3 Model performances

The performances of DENFIS model in this study were evaluated based on several statistical measures such as coefficient of determination (R^2), Root Mean Square Error (RMSE) and Mean Absolute Error (MAE)

$$R^2 = \left[\frac{\sum_{i=1}^{n}(P_i - \overline{P})(\widehat{P}_i - \widetilde{P})}{\sqrt{\sum_{i=1}^{n}(P_i - \overline{P})^2} \times \sqrt{\sum_{i=1}^{n}(\widehat{P}_i - \widetilde{P})^2}}\right]^2 \qquad (2)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(P_i - \widehat{P}_i)^2}{n}} \qquad (3)$$

$$MAE = \frac{\sum_{i=1}^{n}|P_i - \widehat{P}_i|}{n} \qquad (4)$$

where n is the total number of the observations, P_i is the observed precipitation, $\overline{P}$ is average observed precipitation, $\widehat{P}_i$ is the simulated Precipitation rate and $\widetilde{P}$ is average simulated precipitation.

3.0 RESULTS AND DISCUSSIONS

To assess the homogeneity of rainfall data obtained from DID for Klang River catchment, the critical values were adopted from [27], that are 57, 6.95 and 1.30 for Pettitt test, SNHT and VNR respectively. The results of homogeneity tests showed that the rainfall data from all stations were homogeneous and found suitable for further analyses. The initial analysis was done with the one neighboring station for each station with missing data. The results revealed that two triangular functions were appropriate for the development of ANFIS model. It was also found that model performance is suitable with 40 number of epoch. The model performances were first checked with the each selected neighboring station as input and later with input combination of different neighboring stations to select the appropriate combination for prediction of missing data. Out of different input combinations used for developing ANFIS model the best input combination was selected based on their performances in calibration and validation stages against them. It can be seen from the Table 3 that coefficient of determination values were found above 0.75 for the all stations during model calibration using several input combinations. It can also be seen that the MAE and RMSE values are less in validation stage comparing with calibration stage. This shows that model performances were even better in validation stage. The missing values were predicted with the trained ANFIS model using the input selections with best performances in calibration and validation stages. The predicted missing rainfall data and comparison between observed and simulated values can be seen in Figure 3 for the stations 3116074, 3216007 and 3217005.

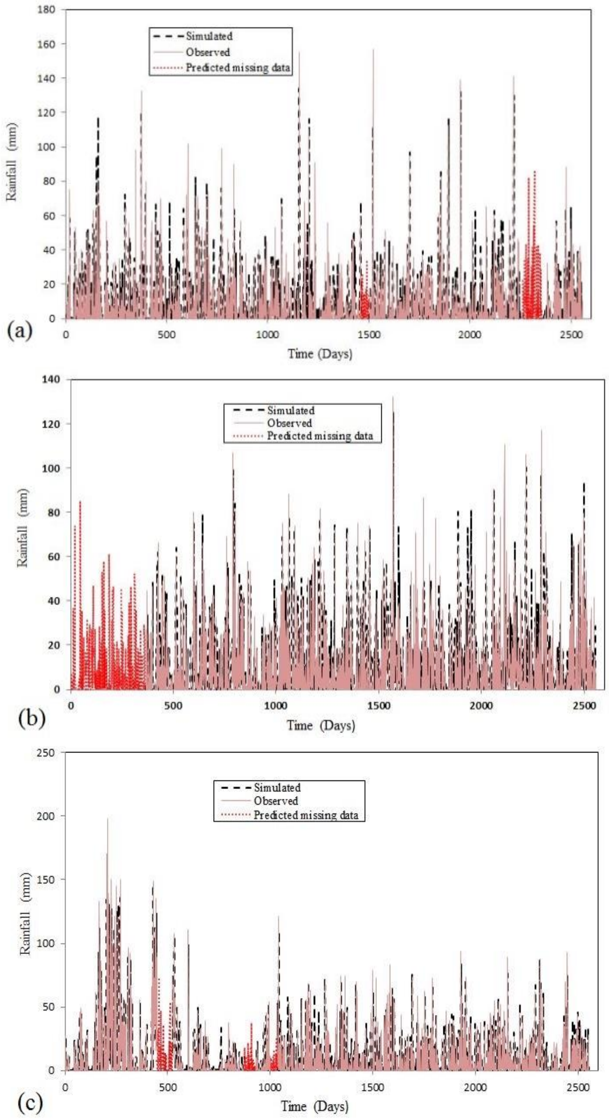


Figure 3 Observed, simulated and predicted missing rainfall data from ANFIS model for the station: (a) 3116074; (b) 3216007; (c) 3217005

To validate the ANFIS model performances in filling missing rainfall data, it was also compared with autoregressive model with exogenous inputs (ARX). The ARX model was developed with the same input combinations which gave better performances in calibration and validation phases for ANFIS model as the initial selection criteria is same for both models. Figure 4 shows comparisons of the ANFIS model and ARX model based on model performances for the three stations. As can be seen R^2 values obtained by ANFIS varies widely with those obtained from ARX model for all stations. ANFIS model gave much higher R^2 values comparing with ARX model. Similarly for RMSE and MAE the values obtained from ANFIS model are much lower than that ARX model. On basis of all statistics it can be said that ANFIS model completely out performed ARX model.

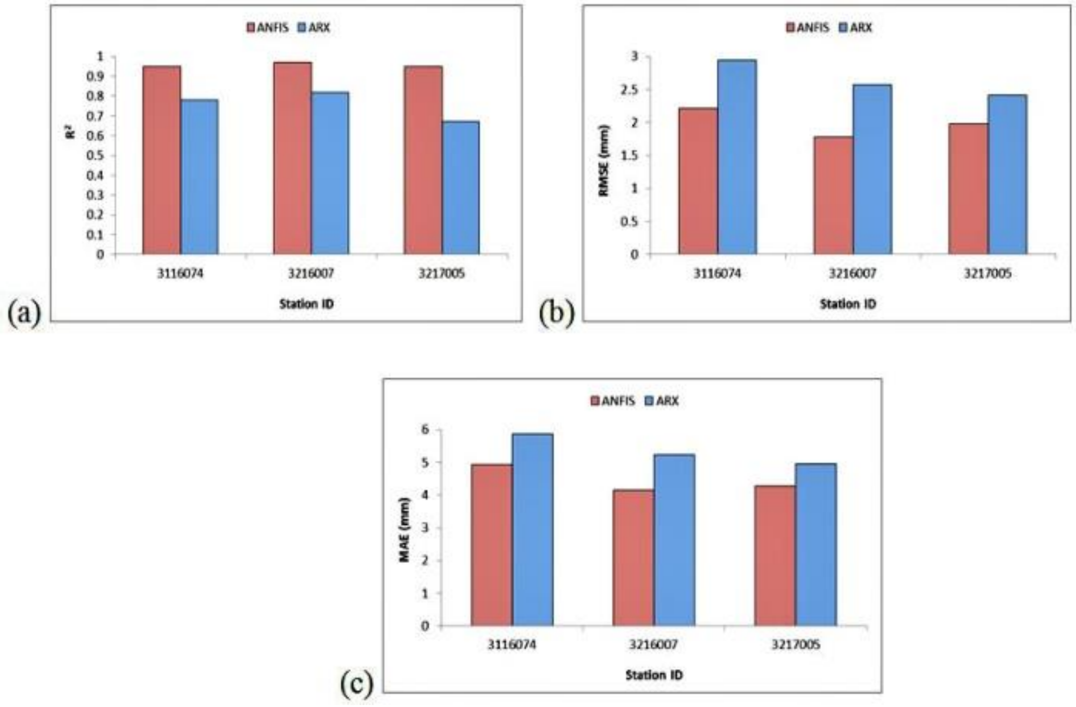


Figure 4 Comparison of ANFIS model and ARX model in validation stage: (a) R^2; (b) RMSE; and (c) MAE

4.0 CONCLUSION

The performance of the ANFIS model for filling rainfall data for Klang River catchment has been evaluated in present study. The ANFIS model was calibrated with the neighboring rainfall stations located in the catchment. The model performance was evaluated by different statistical parameters namely R^2, MAE and RMSE. The study found that ANFIS model is proficient for the prediction of missing rainfall data. More investigations on this approach will require the confidence of the hydrologists in dealing with the problems of filling missing data. The study explored the importance of the availability of more numbers of rainfall stations to achieve required outcome for different hydrological purposes. The availability of long period data can make the performance of the ANFIS modeling practices more precise. The study highlighted the importance of homogeneity tests to check the variability of the available rainfall data as the doubted stations can effect on the performances.

Acknowledgement

The authors would like to express their appreciation to the Department of Irrigation and Drainage for provision of rainfall data for Klang River catchment and profusely thank to Lasbela University of Agriculture, Water and Marine Sciences for financial support to the scholar (PD-MS/LUAWMS/315).

References

[1] Ilunga, M. and Stephenson, D. 2007. Infilling Streamflow Data Using Feed-Forward Back-Propagation (BP) Artificial Neural Networks: Application Of Standard BP And Pseudo Mac Laurin Power Series BP Techniques. *Water SA*. 31(2): 171-176.

[2] Brath, A. Montanari, A. and Toth, E. 2004. Analysis Of The Effects Of Different Scenarios Of Historical Data Availability On The Calibration Of A Spatially-Distributed Hydrological Model. *Journal of Hydrology*. 291(3): 232-253.

[3] Nkuna, T. and Odiyo, J. 2011. Filling Of Missing Rainfall Data In Luvuvhu River Catchment Using Artificial Neural Networks. *Physics and Chemistry of the Earth*. Parts A/B/C. 36(14): 830-835.

[4] Mwale, F. Adeloye, A. and Rustum, R. 2012. Infilling Of Missing Rainfall And Streamflow Data In The Shire River Basin, Malawi–A Self Organizing Map Approach. *Physics and Chemistry of the Earth, Parts A/B/C*. 50: 34-43.

[5] Simanton, J. R. and Osborn, H. B. 1980. Reciprocal-Distance Estimate Of Point Rainfall. *Journal of the Hydraulics Division*. 106(7): 1242-1246.

[6] Lynch, S. 2004. Development of a Raster Database of Annual, Monthly and Daily Rainfall for Southern Africa: Report to the Water Research Commission. Water Research Commission.

[7] Dinpashoh, Y. Jhajharia, D. Fakheri-Fard, A. Singh, V. P. and Kahya, E. 2011. Trends In Reference Crop Evapotranspiration Over Iran. *Journal of Hydrology*. 399(3): 422-433.

[8] Nauck, D. Klawonn, F. and Kruse, R. 1997. *Foundations Of Neuro-Fuzzy Systems*. John Wiley & Sons, Inc. New York, NY, USA.

[9] Karray, F. O. and De Silva, C. W. 2004. *Soft Computing And Intelligent Systems Design: Theory, Tools And Applications*. Addison Wesley Longman.

[10] Mamdani, E. H. 1974. Application Of Fuzzy Algorithms For Control Of Simple Dynamic Plant. *Electrical Engineers, Proceedings of the Institution of*. 121(12): 1585-1588.

[11] Takagi, T. and Sugeno, M. 1985. Fuzzy Identification Of Systems And Its Applications To Modeling And Control. *Systems, Man and Cybernetics, IEEE Transactions on*. SMC-15(1): 116-132.

[12] Ali, T. B. and Dechemi, N. 2004. Daily Rainfall-Runoff Modelling Using Conceptual And Black Box Models; Testing A Neuro-Fuzzy Model. *Hydrological Sciences Journal-Journal Des Sciences Hydrologiques*. 49(5): 919-930.

[13] Nasr, A. and Bruen, M. 2008. Development Of Neuro-Fuzzy Models To Account For Temporal And Spatial Variations In A Lumped Rainfall–Runoff Model. *Journal of Hydrology*. 349(3-4): 277-290.

[14] Chang, F.-J. and Chen, Y.-C. 2001. A Counterpropagation Fuzzy-Neural Network Modeling Approach To Real Time Streamflow Prediction. *Journal of Hydrology*. 245(1-4): 153-164.

[15] Aqil, M. Kita, I. Yano, A. and Nishiyama, S. 2007. Analysis And Prediction Of Flow From Local Source In A River Basin Using A Neuro-Fuzzy Modeling Tool. *Journal of Environmental Management*. 85(1): 215-223.

[16] Firat, M. and Turan, M. E. 2010. Monthly River Flow Forecasting By An Adaptive Neuro-Fuzzy Inference System. *Water and Environment Journal*. 24(2): 116-125.

[17] Yan, H. Yan, H. Zou, Z. and Wang, H. 2010. Adaptive Neuro Fuzzy Inference System For Classification Of Water Quality Status. *Journal of Environmental Sciences*. 22(12): 1891-1896.

[18] Noori, R. Safavi, S. and Nateghi, S.A. 2013. A Reduced-Order Adaptive Neuro-Fuzzy Inference System Model As A Software Sensor For Rapid Estimation Of Five-Day Biochemical Oxygen Demand. *Journal of Hydrology*. 495: 175-185.

[19] Partal, T. and Kişi, Ö. 2007. Wavelet And Neuro-Fuzzy Conjunction Model For Precipitation Forecasting. *Journal of Hydrology*. 342(1-2): 199-212.

[20] Jang, J. S. R. 1993. ANFIS: Adaptive-Network-Based Fuzzy Inference System. *Systems, Man and Cybernetics, IEEE Transactions on*. 23(3): 665-685.

[21] Che-Ani, A.L. Shaari, N. Sairi, A. Zain, M.F.M. and Tahir, M.M. 2009. Rainwater Harvesting As An Alternative Water Supply In The Future. *European Journal of Scientific Research*. 34(1): 132-140.

[22] Yakubu, M. L. Yusop, Z. and Fulazzaky, M.A. 2014. The Influence of Rain Intensity on Raindrop Diameter and the Kinetics of Tropical Rainfall: A Case study of Skudai, Malaysia. *Hydrological Sciences Journal*: 1-8.

[23] Pettitt, A. 1979. A Non-Parametric Approach To The Change-Point Problem. *Applied Statistics*: 126-135.

[24] Alexandersson, H. 1986. A Homogeneity Test Applied To Precipitation Data. *Journal of climatology*. 6(6): 661-675.

[25] Von Neumann, J. 1941. Distribution Of The Ratio Of The Mean Square Successive Difference To The Variance. *The Annals of Mathematical Statistics*. 12(4): 367-395.

[26] Van Ooyen, A. and Nienhuis, B. 1992. Improving The Convergence Of The Back-Propagation Algorithm. *Neural Networks*. 5(3): 465-471.

[27] Wijngaard, J. Klein, T.A. and Können, G. 2003. Homogeneity of 20th century European daily temperature and precipitation series. *International Journal of Climatology*. 23(6): 679-692.