Australian Journal of Basic and Applied Sciences, 4(11): 5668-5674, 2010
ISSN 1991-8178

Deciding the Embedding Nonlinear Model Dimensions and Data Size prior to Daily Reference Evapotranspiration Modeling

[1]Fatemeh Karimaldini, [2]Lee Teang Shui, [3]Mohammadreza Abdollahi, [4]Najmeh Khalili

[1,2]Department of Agriculture and Biological, Faculty engineering, University Putra Malaysia, Malaysia
[3]Department of Electrical and Electronic Engineering, Faculty of Engineering, University Putra Malaysia, Malaysia
[4]Department of Water Engineering, Faculty of Agriculture, Ferdowsi University of Mashhad, Iran

Abstract: Evapotranspiration is an integral part of the hydrologic cycle and an important component in water resource development and management. It is difficult to obtain an accurate formula for ET_O estimation that is suitable to encompass all environments, because evapotranspiration is an incidental, nonlinear, complex and unsteady process. Soft computing models are able to handle noisy data from a dynamic and nonlinear system such as the evapotranspiration process. But, they do not have the ability of pre-processing before model development. In this study, the Gamma Test (GT) technique is applied to find the best input combination and number of sufficient data points for evapotranspiration modeling under humid and arid conditions. It was found that the minimum required variables to construct a good nonlinear model under arid conditions are the minimum and maximum air temperature and wind speed data. For humid conditions the minimum and maximum air temperature, solar radiation and mean relative humidity are the most effective variables.

Key words: ANFIS model; Empirical formula; Evapotranspiration

INTRODUCTION

The term evapotranspiration (ET) is defined as the "water loss or consumptive use" to the atmosphere by the combined process of evaporation from the soil and transpiration from the plants. In many areas around the world, the accomplishment of agriculture is highly dependent on the determination of the required irrigation water, which mainly relay on estimation of the evapotranspiration, along with knowledge of precipitation and soil moisture storage capacity. In the areas with little rainfall, water loss through evapotranspiration affects water budget and hence the design of water resource and irrigation projects. The important factors affecting evapotranspiration are temperature, solar radiation, relative humidity, and the wind speed. Evapotranspiration estimation is important for agricultural, hydrological, and climatic studies, because it is a major part of the hydrological cycle (Sobrino, 2005). The Food and Agricultural Organization of the United Nations (FAO) assumed the combination equation of Penman–Monteith modified by Allen in 1998 (FAO-56 PM equation) as a standard equation for reference evapotranspiration (ET_O) estimation, and for the calibration of other equations. The major limitation of this method is that it requires many weather variables, including: air temperatures (T), relative humidity (RH), solar radiation (R_s) and wind speed (U_2). Unfortunately, some of these variables are often not available. Most weather stations are equipped with just sensors for air temperature and setting up a station that records all required data for Penman-Monteith equation is expensive. A few empirical or semi-empirical methods have been developed to overcome the problem of absence of meteorological data such as the air temperature-based estimation of radiation and relative humidity (Allen, *et al*., 1998). But, their performances are not satisfactory to all climatic conditions due to the complicated nature of evapotranspiration. The Committee on Irrigation Water Requirements of the American Society of Civil Engineers (ASCE) examined the performance of 20 ET_O equations under different climatic conditions and compared the results with lysimeter data. The results implied the widely varying performance of the methods under different climatic conditions. Among all ET_O equations the Penman Monteith equation showed the best performance for estimating daily and monthly ET_O in all the climates (Jensen, *et al*., 1990).

Corresponding Author: Fatemeh Karimaldinia, Department of Agriculture and Biological, Faculty engineering, University Putra Malaysia, Malaysia
Tel.: +60 147 367830; E-mail: Karimadini.f@ gmail.com.

Soft computing models have the capability of modeling nonlinear system such as evapotranspiration, even in the case of unknown underlying physical relationships. This advantage makes them more suitable over conventional models. Soft computing is an approach that "parallels the remarkable ability of the human mind to reason and learn in an environment of uncertainty and imprecision" (Zadeh, 1994). Some researcher, used some paradigms of soft computing (i.e., Artificial Neural Networks (ANNs), and adaptive neuro-fuzzy based inference system (ANFIS)) for evapotranspiration modeling (Aytek, 2009; Kisi, 2006; Kisi, *et al*., 2007; Landeras, *et al*., 2008; Rahimi Khoob, 2008; Zanetti, *et al*., 2007).

A major drawback of soft computing techniques is that they do not have the ability of pre-processing before model development. Applying some inputs that are noisy, insignificant, irrelevant or dependent on other inputs will increase the complexity of the model and does not affect the precision of learning (Jang(1996); (Pomares, *et al*., 2002)). Using Gamma test would help to find the best input combination and the required size of data points before constructing a model. This can prevent many trial and error procedures of the modeling process and improve the performance of the model.

Moghaddamnia *et al*. 2009 used the Gamma test technique before modeling evaporation under arid areas of Iran. Remesan, *et al*., 2008 used Gamma test before solar radiation modeling. They found the capability of Gamma test in determining the importance rule of inputs in modeling, the require data length and preventing the over fitting problem during the trained period. From the literature no published studies are available to utilize pre-processing techniques, for ET_O modeling and thus the motivation for the present study.

The objective of this study is to find the proper input data combination and number of data points that are enough for model construction using Gamma technique.

Methodology:

2.1 Geographical Situation and Meteorological Data Acquisition:

Iran is a country with four-season weather and an area of 1648000 and most of the areas are subject to hot, dry climate. Based on the Demartone's climatic classification, two arid conditions (Kerman and Isfahan stations) and two humid conditions (Rasht and Ramsar stations) are considered in this study. The description of the stations that are used in this research and the aridity coefficients are shown in Table 1.

Table 1: Summary of the Weather Station Sites

Climate	Station	*P (mm)*	*T*(°C)	*I(mm/°c)*	Elevation(mm)	Latitude	Longitude
Arid	Kerman	152.9	15.8	5.92	1753.8	30°15 N	56°58 E
	Isfahan	122.8	16.2	4.68	1550.4	32°37 N	51°40 E
Humid	Rasht	1359	15.9	52.47	-6.9	37°15 N	49°36 E
	Ramsar	1217.8	16	46.83	-20	36°54 N	50°40 E

In this table, P, T and I are the average precipitation, temperature and aridity coefficient respectively. The climatic data of 5 years covering the period of (21March 2000-20 March 2005), and sourced from the Central Meteorological Organization of Iran are used in this study. The correlation of meteorological parameter with ET_o for four stations is provided in Table 2.

Table 2: Correlation of Meteorological Variables with Evapotranspiration for the four stations

Variables	Kerman	Isfahan	Rasht	Ramsar
(T)	0.87	0.80	0.63	0.71
(R_s)	0.68	0.71	0.64	0.66
(RH)	93.00	0.64	0.53	0.39
(U_2)	0.19	0.23	0.14	0.05

2.2 Determining the Best Input Combination with Gamma Test Technique:

Having prior knowledge of the model can help to increase the performance of the model. The Gamma test is a nonlinear continuous modeling and analysis tool, which estimates the minimum mean square error (MSE) during modeling the unseen data and allows for examining the input/output relationship in a numerical data set. It can help to find the required size of data and best input combination to achieve a particular target output. The Gamma test was first introduced by Koncar and Agalbjrn (Agalbj.rn, *et al*., 1997) and after that many researchers discussed this issue in detail (Chuzhanova, *et al*., 1998; de Oliveira, 1999; Durrant, 2001; Jones, *et al*., 2002; Tsui, 1999; Tsui, *et al*., 2002).

The basic idea of the Gamma test is presented as follows: Assume a set of *M* input output observations of the form:

$$\{(x_i, y_i) : 1 \le i \le M\} \tag{1}$$

where x R and y R denote the input and output vectors respectively. The relationship between an input x and corresponding output y can be presented as:

$$y=f(x)+r \tag{2}$$

where f is a smooth unknown function and r is a random variable that represents the noise and has an expectation of Zero. The Gamma test estimates what proportion of the variance of y is related to the function f and what proportion is related to the random variable r. Gamma test can estimate $Var(r)$ directly from the data even when there is no information about underlying smooth function. Having this estimate is much easier than discover the model. In fact, Gamma test does not measure the fitness of data to a line, so it is not related to the shape of function as an alternative it distinguishes noise and smooth relationships. The Gamma test is based on the fact that if two points x and x' are close together in the input space, then their corresponding outputs y and y' should also be close together in the output space. If the outputs are not close together, then this difference is considered as a noise. This estimate can be calculated by these two following functions: The Delta function of the input vectors,

$$\delta_M(k)=1/M\sum \left| x_{N[I,k]}-x_i \right|^2 \ (1 \le k \le p) \tag{3}$$

where $|.|$ denotes the Euclidean distance, $x_{N[I,k]}$ represents the kth ($1 \le k \le p$) nearest neighbours, for each vector x_i ($1 \le i \le M$), P is the number of near neighbours.
The corresponding Gamma function of the output values,

$$\gamma_M(k)=1/2M\sum \left| y_{N[I,k]}-y_i \right|^2 \ (1 \le k \le p) \tag{4}$$

where, $y_{N[I,k]}$ is the corresponding y-value for the kth nearest neighbour of x_i in equation (3). By plotting the linear regression line of $_M(k)$ versus $\delta_M(k)$ for ($1 \le k \le p$), very useful information can be explored:

$$\gamma = A\delta + \Gamma \tag{5}$$

First the intercept of this line on the vertical axis (δ=0) reveals the Gamma statistic value denoted by Γ The small value of Γ, indicates a strong predictive relation between the input variables and the output. Large values of Γ shows that the output is the result of random variation, and the inputs are irrelative to the output. This is because of neglecting some important input variables or using insufficient data availability. Second, the gradient (A) proposes an indication of the model's complexity (the steeper gradient the greater model complexity). Besides gradient, the results can be standardize by considering some other indicators such as v_{ratio} and standard error of Γ. The v_{ratio} is defined as:

$$v_{ratio}= \Gamma/(\sigma^2(y)) \tag{6}$$

Where $\sigma^2(y)$ is the variance of output y. A v_{ratio} close to zero shows that there is a high degree of predictability of the given output y. The standard error (SE) close to zero, the reliability value of the Gamma statistic as an estimator of noise in the given output will increase.

In this study, different combination of inputs are examined with the aid of Gamma test to find their influence on the evapotranspiration in the arid and humid climates at four stations of Iran.This is done by using Win Gamma software. By running a series of Gamma test, this software gives the values of gamma, gradient, standard deviation and v-ratio in each experiment.

2.3 Determing Size of Data Points Using the M-test Technique:

The Win Gamma software also can help to have confidence about sufficient data being present for training and calibration a model by running a series of the M-test. M-test line can be performed by measuring the amount of change in the Gamma values for increasing small step in the number of data points (M). By plotting M-test line and presenting all input variables, the underlying smooth function can be achieved after a point. When the M-test line becomes flat, adapting a model to fit the data should be stopped. This can prevent the over fitting problem. Thus, non stabilized the M-test line indicates that more data is required to construct a smooth model. Finding sufficient size of data points for training requires a lot of trial and error. Furthermore, using a part of data for validation may result in lake of data for the following training or testing.

RESULTS AND DISCUSSIONS

The best combination of inputs can be determined among 2^n-1 meaningful combination of inputs by observing the Gamma value. In this study, the nature of importance of eight input variables (T_{min} , T_{max} , T_{mean}, R_s , RH_{min} , RH_{max} , RH_{mean} , U_2) for ET_O modeling were analyzed under humid and arid conditions in Iran by assessing the effect of different input combinations on error variance. For selecting the best input combination, only the Gamma value is considered. The other three factors (gradient, v-ration and standard error of Γ) represent how good the Gamma value is and how complex the model should be developed. In Tables 3 and 4 some of the examined combinations for arid and humid climatic are presented.

Table 3: The Gamma Test Results on the Evapotranspiration Estimation Data Sets (Arid Climatic)

	Location	Kerman				Isfahan			
Exp	Input	Gamma (г)(mm^2)	Gradient (A)	Standard Error of г(mm^2)	V-Ratio	Gamma (г)(mm^2)	Gradient (A)	Standard Error of г(mm^2)	V-Ratio
1	T_{mean}, R_s, RH_{mean}, U_2	0.061	0.024	0.010	0.004	0.157	0.032	0.016	0.014
2	T_{min}, T_{max}, R_s, RH_{min}, RH_{max}, U_2	0.135	0.016	0.011	0.009	0.251	0.021	0.023	0.022
3	T_{min}, T_{max}, R_s, RH_{mean}, U_2	0.054	0.026	0.019	0.004	0.117	0.046	0.011	0.010
4	T_{mean}, R_s, RH_{min}, RH_{max}, U_2	0.125	0.022	0.024	0.008	0.194	0.028	0.031	0.017
5	T_{min}, T_{max}, RH_{mean}, U_2	0.157	0.033	0.016	0.011	0.114	0.067	0.013	0.010
6	T_{min}, T_{max}, R_s, U_2	0.177	0.044	0.012	0.012	0.268	0.049	0.019	0.023
7	T_{min}, T_{max}, R_s,RH_{mean},	1.154	0.008	0.029	0.078	1.437	0.013	0.050	0.125
8	R_S, RH_{mean}, U_2	1.128	0.058	0.026	0.076	0.783	0.069	0.026	0.068
9	T_{min}, T_{max}, RH_{mean}	1.289	0.013	0.030	0.087	1.905	-0.003	0.037	0.166
10	T_{min}, T_{max}, R_s	1.299	0.028	0.033	0.088	1.658	0.010	0.056	0.145
11	T_{min}, T_{max}, U_2	0.366	0.068	0.017	0.025	0.326	0.136	0.025	0.028
12	T_{min}, T_{max}	1.587	0.013	0.032	0.106	2.164	0.030	0.036	0.189

Table 4: The Gamma Test Results on the Evapotranspiratin Estimation Data Sets (Humid Climatic)

	Location	Rasht				Ramsar			
Exp	Input	Gamma (г)(mm^2)	Gradient (A)	Standard Error of г(mm^2)	V-Ratio	Gamma (г)(mm^2)	Gradient (A)	Standard Error of г(mm^2)	V-Ratio
1	T_{mean}, R_s, RH_{mean}, U_2	0.136	0.013	0.019	0.034	0.023	0.009	0.007	0.008
2	T_{min}, T_{max}, R_s, RH_{min}, RH_{max}, U_2	0.175	0.005	0.008	0.044	0.013	0.005	0.008	0.005
3	T_{min}, T_{max}, R_s, RH_{mean}, U_2	0.131	0.009	0.016	0.033	0.010	0.007	0.007	0.004
4	T_{mean}, RH_{min}, RH_{max}, U_2	0.395	0.008	0.014	0.098	0.021	0.006	0.009	0.008
5	T_{min}, T_{max}, RH_{mean}, U_2	0.524	0.008	0.021	0.131	0.230	0.017	0.013	0.083
6	T_{min},T_{max}, R_s, U_2	0.376	0.002	0.031	0.094	0.154	0.015	0.009	0.055
7	T_{min}, T_{max}, RH_{mean}, R_s	0.200	0.023	0.020	0.050	0.109	0.007	0.008	0.039
8	R_S, RH_{mean}, U_2	0.384	0.018	0.014	0.096	0.376	0.016	0.019	0.135
9	T_{min}, T_{max}, RH_{mean}	0.461	0.012	0.029	0.165	0.323	0.016	0.010	0.116
10	T_{min}, T_{max}, R_s	0.442	0.021	0.048	0.155	0.278	0.014	0.017	0.100
11	T_{min}, T_{max}, U_2	0.702	0.040	0.025	0.175	0.392	0.045	0.016	0.141
12	T_{min}, T_{max}	1.031	0.138	0.027	0.257	0.508	0.027	0.014	0.182

From Tables 3 and 4 the minimum Gamma value was observed when all meteorological inputs were observed under arid and humid conditions, but it is important to identify the best embedded dimensions for ET_O modeling in the absence of missing weather data.

The Gamma results of experiment 1 to 4 under arid and humid climatic indicated that the effect of using T_{max} and T_{min} in the combination of inputs was more than T_{mean}. For the parameter *RH,* on the other hand, RH_{mean} was more effective than RH_{min} and RH_{max} for estimating ET_O value.

From Table3, under arid conditions, the results of the Gamma test were not in line with our expectations. Table 2 suggests that after temperature, two parameters *RH* and R_s have the highest correlation with ET_o. The lowest correlation was observed for the wind speed. However, from Table 3, the Gamma test indicated that it does not matter to eliminate any one of *RH* and R_S but after temperature, the wind is the most effective variable that should be included in the model input. Therefore, according to the Gamma values, the best result could be achieved when *T* and U_2 are considered in the combination of inputs under arid conditions (see the Gamma value of experiment 5, 6 and 11) which is comparable to the combination of all inputs. This difference between Tables 2 and 3 reveals that, the combination of insignificant variable (U_2) with other variables may provide a very significant new variable. This fact is rooted from the highly nonlinear nature of relations between variables. In addition, because *RH* and R_S can be derived from the information of the temperature, they are not as informative as wind speed.

From Table 4, under humid conditions, it has been found that the relative importance of the inputs is *T*, R_s, *RH*, U_2 in descending order. The significance of the wind speed data was relatively small when compared with other meteorological variables since by excluding this input from the combination of inputs, small variation in the Gamma values is observed.

The number of data points to construct a reliable model for daily evapotranspiration estimation was determined using the M-test. If the *M*-test line becomes stable and asymptote to the estimated Gamma value, then we have confidence that sufficient data has been presented to construct a smooth model. The results of *M*-test analysis and the response of the models to the removing of one input, and consequently, the significant role of each variable in the combination of inputs are illustrated in Figures 1 and 2, for arid conditions(Kerman and Isfahan) and Figures 3 and 4, for humid conditions (Rasht and Ramsar) respectively.

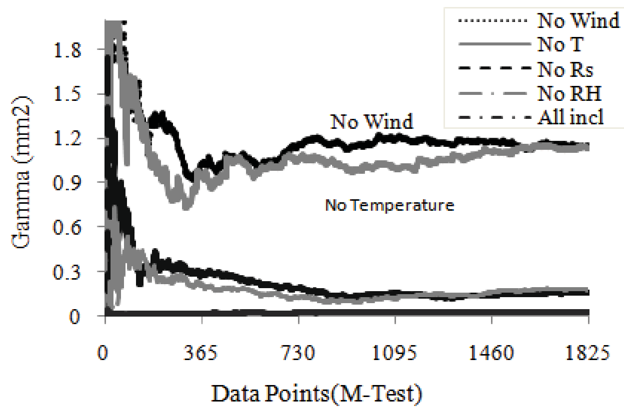

Fig. 1: Variation of the Gamma statistic related to removing one variable from combination of inputs for different training data length (Kerman station)

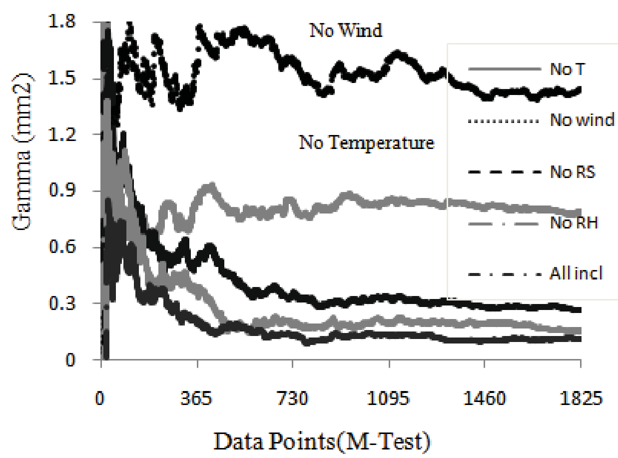

Fig. 2: Variation of the Gamma statistic related to removing one variable from combination of inputs for different training data length (Isfahan station)

According to these figures, the number of data points was found at around 1460 data points (3 years) where, the stable part of the *M*-test line converged to the Gamma values. From Figures 1 and 2, the largest modeling error was observed when the wind and after that, the temperature were excluded from the model inputs. When these two variables were both included, the influence of *RH* and R_s were not significant and the three lines for '*No RH*, *No* R_S and *All incl*' were not distinguished. Here, "*All incl* "denotes all weather inputs (T, R_S, RH, U_2) include in the model. However, from Figures 3 and 4, under humid climatic the difference between *All incl* line and *No wind* line is not as much as other lines. This low effect of wind speed on evapotranspiration under humid conditions may be due to the fact that under humid weather conditions, the

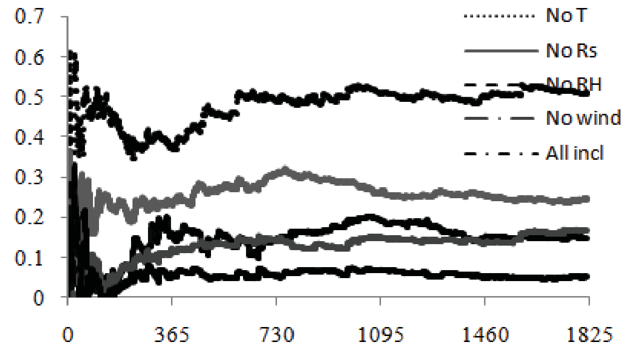

Fig. 3: Variation of the Gamma statistic related to removing one variable from combination of inputs for different training data length (Rasht station)

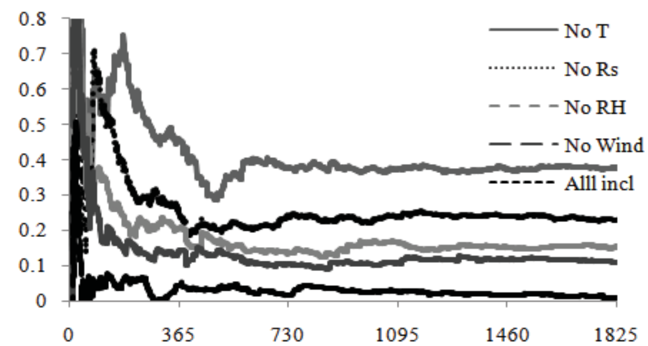

Fig. 4: Variation of the Gamma statistic related to removin one variable from combination of inputs for different training data length (Ramsar station)

wind can only replace saturated air with slightly less saturated air. Consequently, the effect of wind speed on evapotranspiration under humid conditions is lesser than under arid conditions where variation in the wind may lead to larger variation in the evapotranspiration.

Conclusion:

The major challenges associated with the prediction and modeling of nonlinear variables such as evapotranspiration and many other hydrological parameters, using non linear techniques like ANNs and ANFIS are the gaps in finding sufficient data length for a meaningful training and choosing most relevant inputs to make a reasonable model. This problem is solved using the Gamma test technique to better embedding nonlinear model dimensions and meaning full data length for evapotranspiration modeling, before model construction and evaluation under arid and humid conditions.

For arid conditions, the low value of Γ, v-ratio, A and SE was observed when temperature and wind speed were considered as inputs. These values all together can give a clear indication of a reasonably accurate smooth model with the minimum complexity and good fit to construct a nonlinear predictive model. Under humid conditions, the low effect of wind speed on the evapotranspiration is observed. In both climatic conditions, the number of data points to construct a reliable model for daily evapotranspiration model was determined using the M-test, which has indicated that around three years data points is sufficient.

Using Gamma test can help to save a great amount of time and effort for a modeler to choose proper combinations of inputs and appropriate number of data points that are required for calibration of a model. The Gamma test employed in this research would have a great potential to solve other types of hydrological problems.

REFERENCES

Agalbjرrn, S., N. Konar and A. Jones, 1997. A note on the gamma test. Neural Computing and Applications, 5(3): 131-133.

Allen, R., L. Pereira, D. Raes and M. Smith, 1998. Crop evapotranspiration-Guidelines for computing crop water requirements-FAO Irrigation and drainage paper 56. FAO, Rome, 300.

Aytek, A., 2009. Co-active neurofuzzy inference system for evapotranspiration modeling. Soft Computing-A Fusion of Foundations, Methodologies and Applications, 13(7): 691-700.

Chuzhanova, N., A. Jones and S. Margetts, 1998. Feature selection for genetic sequence classification. Bioinformatics, 14(2): 139.

de Oliveira, A., 1999. Synchronization of chaos and applications to secure communications. PhD thesis, Department of Computing, Imperial College of Science, Technology and Medicine, University of London, 1999.

Durrant, P., 2001. winGamma TM: a non-linear data analysis and modelling tool with applications to flood prediction. PhD thesis, Department of Computer Science, Cardiff University, Wales, UK.

Jensen, M., R. Burman and R. Allen, 1990. Evapotranspiration and irrigation water requirements. ASCE Manuals and Reports on Engineering Practice No. 70. The Society, New York, NY, 360.

Jones, A., A. Tsui and A. Oliveira, 2002. Neural models of arbitrary chaotic systems: construction and the role of time delayed feedback in control and synchronization. Complexity International, 9: 1-9.

Kisi, O., 2006. Evapotranspiration estimation using feed-forward neural networks. Nordic Hydrology, 37(3): 247-260.

Kisi, O. and O. Ozturk, 2007. Adaptive neurofuzzy computing technique for evapotranspiration estimation. Journal of Irrigation and Drainage Engineering, 133: 368.

Koncar, N., 1997. Optimisation methodologies for direct inverse neurocontrol: University of London.

Landeras, G., A. Ortiz-Barredo and J. López, 2008. Comparison of artificial neural network models and empirical and semi-empirical equations for daily reference evapotranspiration estimation in the Basque Country (Northern Spain). Agricultural water management, 95(5): 553-565.

Moghaddamnia, A., M. Ghafari Gousheh, J. Piri, S. Amin and D. Han, 2009. Evaporation estimation using artificial neural networks and adaptive neuro-fuzzy inference system techniques. Advances in Water Resources, 32(1): 88-97.

Pomares, H., I. Rojas, J. Gonzalez and A. Prieto, 2002. Structure identification in complete rule-based fuzzy systems. IEEE transactions on fuzzy systems, 10(3): 349-359.

Rahimi Khoob, A., 2008. Comparative study of Hargreaves's and artificial neural network's methodologies in estimating reference evapotranspiration in a semiarid environment. Irrigation Science, 26(3): 253-259.

Remesan, R., M. Shamim and D. Han, 2008. Model data selection using gamma test for daily solar radiation estimation. Hydrological processes, 22(21): 4301-4309.

Sobrino, J., 2005. A simple algorithm to estimate evapotranspiration from DAIS data: Application to the DAISEX campaigns. Journal of hydrology, 315(1-4): 117-125.

Tsui, A., 1999. Smooth data modelling and stimulus-response via stabilisation of neural chaos.

Tsui, A., A. Jones and A. Guedes de Oliveira, 2002. The construction of smooth models using irregular embeddings determined by a Gamma test analysis. Neural Computing & Applications, 10(4): 318-329.

Zadeh, L.A., 1994. Fuzzy logic, neural networks, and soft computing. Commun. ACM, 37(3): 77-84.

Zanetti, S., E. Sousa, V. Oliveira, F. Almeida and S. Bernardo, 2007. Estimating evapotranspiration using artificial neural network and minimum climatological data. Journal of Irrigation and Drainage Engineering, 133: 83.