MODELING OF MONTHLY PAN EVAPORATION USING M5P MACHINE LEARNING TECHNIQUE

Mohamod Abdul Hassan Joehl AL – Janabi
University of Kufa/ Faculty of Science/ Geology Department
mahmood.jwaihil@uokufa.edu.iq

Abstract:

The purpose of this study is to investigate the ability of M5P model trees machine learning technique for estimating monthly pan evaporation from meteorological data. The M5 method as it is implemented in the WEKA system is used to generate trees models. Three different M5P models comprising various combinations of monthly climatic variables (temperature, wind speed, and relative humidity) are developed to evaluate effect of each of these variables on evaporation estimations. Two error statistics namely root mean squared error and coefficient of determination are used to measure the performance of the developed models. Monthly meteorological data of Emara station in Missan, south of Iraq is used in this study as a case study. The results demonstrated that the M5P models whose inputs are wind speed, relative humidity and temperature performed the best among the input combination tried in the study. It was found that M5P could be employed successfully in modeling evaporation process from the available climatic data.

Keywords: Pan evaporation, fuzzy logic, Iraq, M5P machine learning technique.

نمذجة قيم التبخر الشهري باستخدام تقنية تعليم الآلة M5P

محمود عبد الحسن جويهل الجنابي
جامعه الكوفة /كليه العلوم/قسم علوم الارض

الخلاصة:

ان الغرض من هذه الدراسة هو التحري عن امكانية استخدام تقنية النماذج الشجرية M5P للتنبؤ بقيم التبخر الشهرية من المعلومات المناخية. تم تطوير ثلاث نماذج شجرية اعتماداً على المتغيرات المناخية (درجة الحرارة والرطوبة النسبية وسرعة الرياح) . تم استخدم معياران احصائيان هما معامل التحديد وجذر الخطأ التربيعي لتحديد ادائية النماذج المطورة. استخدمت المعلومات المناخية المتوفرة في محطة العمارة المناخية/ جنوب العراق كدراسة حالة. اظهرت النتائج بان النموذج الشجري الذي تكون متغيرات ادخاله سرعة الرياح والرطوبة النسبية ودرجة الحرارة هو الاحسن من بين النماذج الاخرى. وجد ايضاً بان تقنية M5P يمكن تطبيقها بنجاح في نمذجة عملية التبخر من المعلومات المناخية.

الكلمات المفتاحية:. حوض التبخر ، المنطق الضبابي، العراق، تقنية تعليم الألة M5P .

1. Introduction

The evaporation process involves the transfer of water from a liquid state into a gaseous form in the atmosphere. Evaporation is a major component of the hydrological cycle. It is difficult to measure directly and there are various estimation techniques. These range from water budget techniques, such as evaporation pans and lysimeters, to modeling techniques, such as the Penman-Monteith equation. As a process, evaporation is extremely variable in space and time. This variability leads to difficult in moving measurements to area estimation such as required for a catchment study. Three primary common means to estimate evaporation has been used during the past century. These are (1) pan evaporation measurements (2) an estimate of potential evaporation based on weather data, and (3) a reference evapotranspiration.

Pan evaporation has been widely used as an index of evaporation and for estimating lake and reservoir evaporation [1]. An evaporation pan is used to hold water during observations for the determination of the quantity of evaporation at a given location. Such pans are of varying sizes and shapes, the most commonly used is the USA class A pan, which is 1.21 m in diameter and 254 mm deep, constructed of stain less steel, and placed above a 0.15 m tall open timber framework such that the top the pan is about 0.4 m above the surrounding ground level [2]. The other two commonly used pans are the Russian (Soviet) GGI-3000 pan and the GCI tank, both placed in the soil with only 0.075 to 0.1 m of rim above the soil surface [3].

It is impractical to place evaporation pans at every point where there is a planned or existing reservoir and irrigation project [4], and even where there is a pan, the measurements may be vitiated by poor maintenance, leading to errors due to many reasons including growth of algae in the water, weed- growth nearby and an incorrect water level measurement. In view of these difficult it would be useful and cheaper to have some means of estimating pan evaporation with reasonable accuracy from reliable climate measurements such as temperature [5].

Recently, the outstanding results using artificial intelligent techniques such as artificial neural networks, fuzzy inference system, adaptive neuro-fuzzy inference system, and genetic programming in the field of evaporation and evapotranspiration have been obtained [6] [7] [8] [4] [9] [10] and [11]. Application of M5P model trees in hydrology is limited; just few published papers are found, for example to model stage-discharge relationship [12] and to simulate rainfall-runoff process [13]. Works of Solomatine and others proved that M5P is a very effective technique and more understandable and allows one to build a family of models of varying complexity and accuracy.

The aim of this paper is to investigate the applicability of the M5P model tree to estimate monthly pan evaporation from climatic variables easy to measure, for Emara meteorological station, south of Iraq.

2. M5P model trees

A decision tree is a logical model represented as a binary (two-way split) tree that shows how the values of a target (dependent) variable can be predicted by using the values of a set of predictor (independent) variables. These are basically two types of decision trees: (1) classification trees are the most common

and are used to predict a symbolic attribute (class).
(2) regression trees which are used to predict the value of a numeric attribute [14] (Witten and Frank, 1999). If each leaf in the tree contains a linear regression model, that is used to predict the target variable at that leaf, is called a model tree.

The M5P model tree algorithm was originally developed by Quinlan [15]. Detail description of this technique is beyond of this paper and can be found in Witten and Frank [14]. A bit description of this technique follows. The M5 algorithm constructs a regression tress by recursively splitting the instance space using tests on a single attributes that maximally reduce variance in the target variable. Figure 1 illustrates this concept. The formula to compute the standard deviation reduction (SDR) is: [15]

$$SDR = sd(T) - \sum \frac{|T_i|}{|T|} sd(T_i)$$

where T represents a set of example that reaches the node; T_i represents the subset of examples that have the ith outcome of the potential set; and sd represents the standard deviation.

After the tree has been grown, a linear multiple regression is built for every inner node using the data associated with that node and all the attributes that participate for tests in the subtree to that node. After that, every subtree is considered for pruning process to overcome the overfitting problem. Pruning occurs if the estimated error for the linear model at the root of a subtree is smaller or equal to the expected error for the subtree. Finally, the smoothing process is employed to compensate for the sharp discontinuities between adjacent linear models at the leaves of the pruned tree.

3. Data Description

The monthly climatic data of Emara weather station (Latitude 31°78′44.47″ N, Longitude 47°08′27.72″ E) are used in this study. The location of this station is shown in Fig.2. The data consisted of 26 years (1980-2006) of monthly average records of air temperature (T), wind speed (W), relative humidity (RH), and pan evaporation (E). The monthly statistical parameters of the climatic variables are given in Table 1. The $\bar{x}$, x_{min} , x_{max} , S_x , C_k , and R denote the mean, minimum, maximum, standard deviation, coefficient of skewness, and correlation coefficient, respectively. It can be seen from the correlation coefficient between climatic variables and evaporation, Table 1, that temperature has a significant effect on evaporation.

Table 1: The monthly statistical parameters of data set.

Data set	*Unit*	$\bar{x}$	x_{min}	x_{max}	S_x	C_k	R^2
W	*m/s*	*4.03*	*1.60*	*9.8*	*1.47*	*1.00*	*0.506*
RH	*%*	*46.66*	*15.00*	*83.0*	*18.24*	*0.22*	*0.746*
T	*℃*	*24.44*	*8.40*	*39.6*	*9.49*	*-0.05*	*0.768*
E	*Mm*	*280.52*	*24.00*	*959.6*	*195.2*	*0.69*	*1.000*

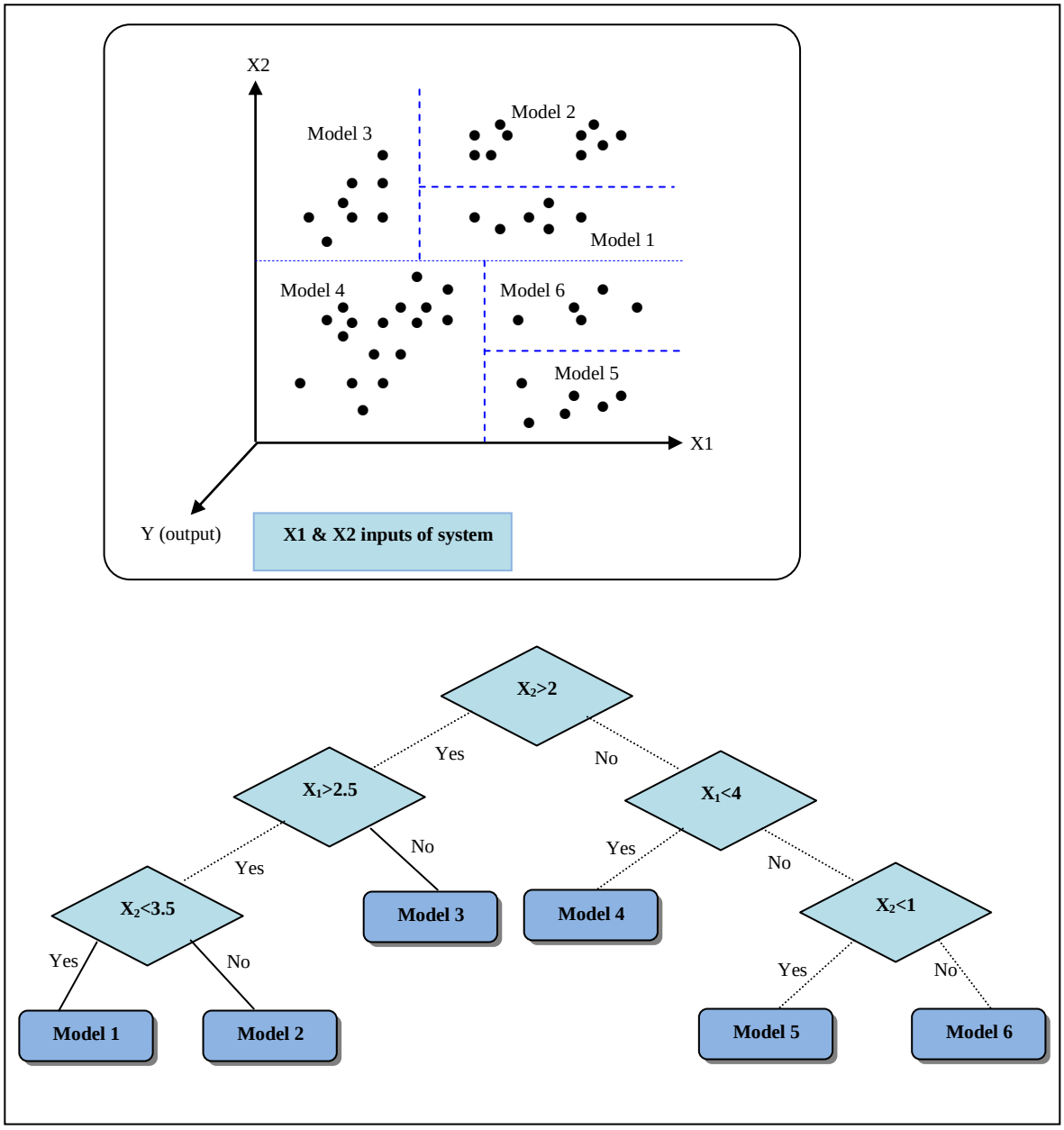

Fig.1: Examples of M5 model 1-6 are linear regression models (modified after Solomatine and Xue (2003)).

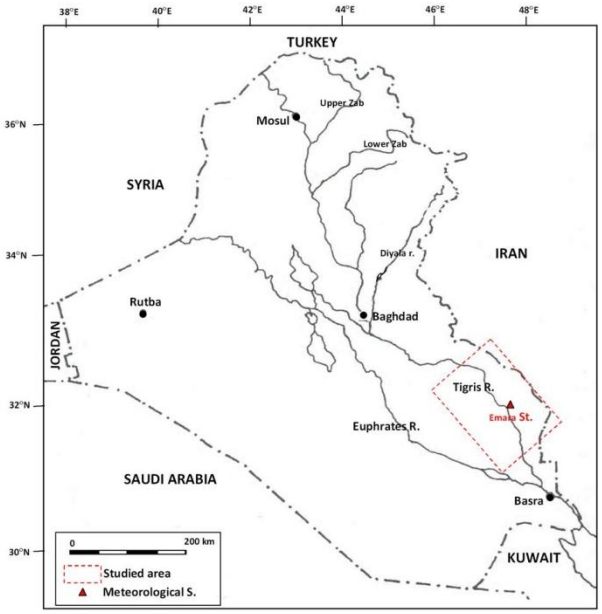

Fig.2: Location of Emara meteorological station.

4. Application and Results

A total of 262 monthly average observations for each climatic variable are used to build decision trees models. Input data is divided into two groups including training set (172 observation points) and testing set (90 observation points). to building M5 model, Weka software was used. Weka is open-source machine learning/data mining software written in Java [14]. The software contains a comprehensive set of pre-processing tools, learning algorithm and evaluation methods.in this study, the parameters of M5 algorithm were set to their default values; pruning factor 2.0 and smoothing option. The data set splits into two groups: 70% for training and reaming for testing. Several input combinations, Table 2, were tried to estimate monthly pan evaporation. The performance of the various models was evaluated by using statistical parameters namely root mean squared error (RMSE) and coefficient of determination (R^2). The RMSE statistic indicates the model ability to predict away from the mean. The optimal value is 0. It is defined as:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{x})^2} \qquad (3)$$

where

x = the measured values

$\hat{x}$ = the predicted values

n = number of observations

The coefficient of determination R^2 measures the linear correlation between the measured and predicted values; the optimal value is 1. It is defined as:

$$R^2 = 1 - \frac{SSE}{SSy} \qquad (4)$$

Where

$$SSE = \sum_{i=1}^{n}(x_i - \hat{x})^2 \qquad (5)$$

$$SSy = \sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad (6)$$

$\bar{x}$ = Mean of the measured values.

5. Results and discussion

Table 2 summarized error statistics of the experiments and Fig.3 showes a comparison between the measured and simulated monthly averaged pan evaporation. The final three model trees were build in this study for evaporation prediction are follows:

Table 2: The RMSE and R^2 of M5P models in training and testing periods.

Model No.	Variables combination	Training		Testing		Time taken to build model (sec)
		RMSE	R^2	RMSE	R^2	
Model 1	W, RH, T	66.30	0.88	61.60	0.89	0.56
Model 2	RH, T	61.60	0.89	71.97	0.84	0.47
Model 3	T	67.70	0.88	67.37	0.88	0.53

For model 1

T <= 21.9 : LM1 (114/10.931%)

T > 21.9 : LM2 (147/43.858%)

LM1 (linear model number 1)

E = 10.8879 * W- 2.2732 * RH + 5.4823 * T+ 131.85

LM2 (linear model number 2)

E = 32.0648 * W- 5.0622 * RH + 9.955 * T+ 112.8265

For model 2

T <= 21.9 : LM1 (114/11.397%)

T > 21.9 :

| RH <= 32.5 :

| | T <= 34.55 :

| | | RH <= 25.5 : LM2 (5/37.48%)

| | | RH > 25.5 : LM3 (19/34.429%)

| | T > 34.55 : LM4 (59/46.943%)

| RH > 32.5 :

| | T <= 29.4 : LM5 (49/27.134%)

| | T > 29.4 : LM6 (15/16.982%)

LM num: 1

E = -2.5387 * RH+ 5.864 * T+ 177.3874

LM num: 2

E = -8.2284 * RH+ 2.896 * T+ 605.5619

LM num: 3

E = -6.7146 * RH+ 2.896 * T+ 544.6922

LM num: 4

E = -8.7175 * RH- 26.1389 * T + 1725.3666

LM num: 5

E = -2.5589 * RH+ 5.9565 * T + 216.8014

LM num: 6

E = -1.5255 * RH + 8.934 * T + 140.0229

For model 3

T <= 21.9 : LM1 (114/14.197%)

T > 21.9 :

| T <= 28.8 : LM2 (47/27.661%)

| T > 28.8 :

| | T <= 34.35 : LM3 (38/42.806%)

| | T > 34.35 : LM4 (62/51.919%)

LM num: 1

E = 10.1738 * T - 49.8769

LM num: 2

E = 6.9394 * T + 80.9011

LM num: 3

E = 8.8512 * T + 126.9639

LM num: 4

E = -13.4091 * T+ 1033.7665

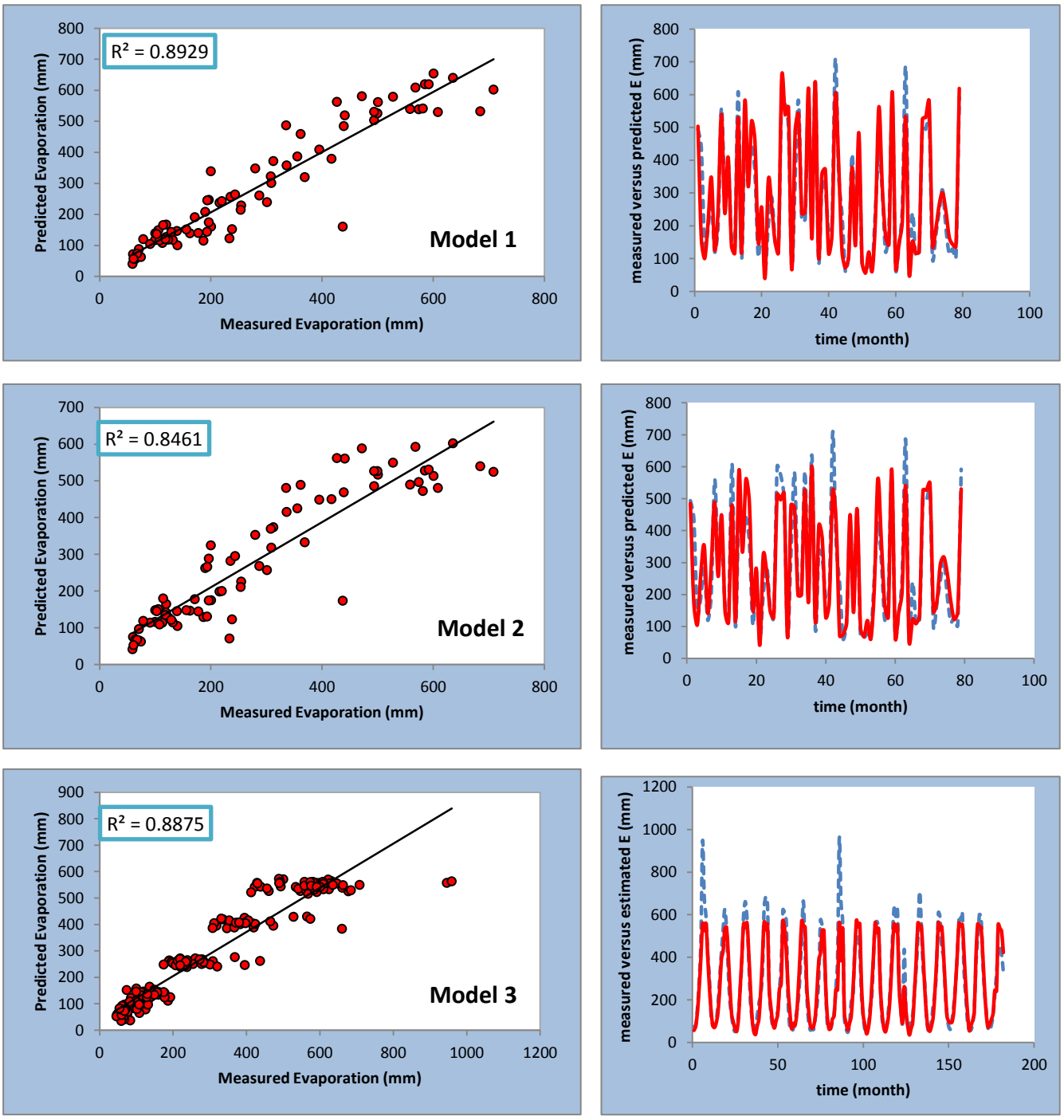

Fig. 3: Measured versus estimated monthly pan evaporation using M5P models.

From table 2, it is obvious that model 1 has the lowest *RMSE* (61.60) and highest R^2 (0.89) for testing period. The M5 model whose input was temperature only also performed very well. Due to the fact that temperature parameter is a very easy to measure, estimation of monthly evaporation from just this parameter is robust and significant issue.

6. Conclusions

M5 model tree is an efficient tool to estimate monthly pan evaporation. The results of M5 model tree are understandable and could be used as predictive tool. We recommended using M5 model tree and other data driven models to manage water resources of Iraq alone or as a hybrid model with physically models to improve situation of that worse managed resource.

7. Reference

[1] Richard G. Allen (2005) Evaporation modeling: potential. Encyclopedia of hydrological sciences, John Wiley & Sons Ltd., 3174 p, 623-633.

[2] Shaw, M. E. (1999) Hydrology in practice. Stanly Thomes Lts, Glos, UK, 569p.

[3] World Meteorological Organization (1970) Guide to hydrometeorological practices, WMO, No. 188. TP. 82. Geneva.

[4] Kisi, O. (2006) Daily pan evaporation modeling using a neuro-fuzzy computing technique. J. Hydrology, 329, 636-646.

[5] Cahoo, J. E., Castello, T. A. and Ferquson, J. A. (1991) Estimating pan evaporation using limited meteorological observations. Agricultural and Forest Meteorology, 55, 181-190.

[6] Sudheer, K. P., Gosain, A. K., Rangan, D. M., and Saheb, S. M. (2002) Modelling evaporation using an artificial neural network algorithm. Hydrologic Process. 16, 3189-3202.

[7] Kumar, M. Raghuwanshi, N. S, Singh, R., Wallendew, W. W., and Pruitt, W. O. (2002) Estimating evapotranspiration using artificial neural network. J. Irrig. Drain Engr., ASCE 128(4), 224-233.

[8] Keskin, M. E., Terzi, O, and Trylon, D. (2004) Fuzzy logic model approaches to daily pan evaporation estimation in western Turkey. Hydrological Sciences, 49(6), 1001-1010.

[9] Parasuraman, K., Elshorbagy, A. and Corry, S. K. (2007) Modelling the dynamics of the evapotranspiration process using genetic programming. Hydrological Sciences, 53(3), 565-578.

[10] Kim, S., and Kim, H. S. (2007) Neural networks and genetic algorithm approach for nonlinear evapotranspiration and evapotranspiration modelling. J. Hydrology, 351, 299-317.

[11] Aytek, A., Guren, A., Yuce, M. I. and Aksoy, H. (2008) An explicit neural network formulation for evapotranspiration. Hydrological Sciences, 53(4), 893-904.

[12] Bhattachary, B. & Solomatine, D. P. (2005) Neural networks and M5 model trees in modeling water-level-discharge relationship. Neurocomputing 63, 381-396.

[13] Solomatine, D. P. & Xue, Y. (2004) M5 model trees and neural networks: application to flood forecasting in the upper reach of the Huai River in China. J. of Hydrol. Engine., ASCE.

[14] Al-Dahaan, S. A. M. (2015) A Model for the Relashinshipe between Boron, Fluoride and Salinity of Groundwater at Safwan. S. Iraq. Journal of Environmental and Earth Science. Vol. 5, No 5, 2015.

[15] Al-Dahaan, S. A. M. (2016) Influence of Groundwater Hypothetical Salts on Electrical Conductivity Total Dissolved Solids. Scientific Research Publishing, Engineering, 2016, 8,823, 830.