Marsh Bulletin 3(2)(2008)125-135

MARSH BULLETIN

Amaricf_Basra office@yahoo.com
abdulalwan@yahoo.com
.marshbulletin@yahoo.com

Rainfall-Runoff modeling by using M5 model trees technique: an example of Tigris catchment area in Baghdad, Middle of Iraq

A.M. Atiaa[a] and H.B. Ghalib[b]

Dept. of Geology, Coll. of Sci., Univ. of Basra, Basra, Iraq

[a]e-mail: alaaatiaa@gmail.com

[b]e-mail: hbggeo@hotmail.com

Abstract

This paper investigates the applicability of the M5 model trees technique to emulate rainfall-runoff transformation of Tigris catchment area in Baghdad city, Middle of Iraq. For building M5 model, a free of charge open –leading machine learning and data mining weka software is used. Four models are build firstly to study the interdependency among the input variables and to select the effective variables. The applicability of this technique is studied by predicting runoff (discharge) of Tigris River one and two months ahead. The results show the high accuracy of the M5 technique to identify low values and some of high values of flow with very high accuracy, but most of the high flows were underestimated. M5 model tree and other data-driven models could be used alone or corporation with physically-based models such as HEC-HMS to manage water resources of Iraq after a detailed monitor hydrological programming surveys are employed.

Key words Model tree, Rainfall-runoff transformation, Tigris River, Iraq

1-Introduction

The rainfall-runoff transformation process is one of the most complex and non-linear relationship in hydrology and water resources. The transformation of rainfall to watershed runoff involves many highly complex processes such as interception, depression storage, evaporation, base flow, and other many hydrological components. For many years, water resources practitioners and

engineers have attempted to understand and investigate the control processes of transformation of a given rainfall to runoff in order to predict stream flow for hydrological studies. Many models were developed during the last decades to capture and simulate this highly non-linear relationship. These models can be basically divided into three groups: (1) the deterministic models seek to simulate the physical process in the catchment involved in the transformation of rainfall to runoff. (2) the stochastic model describe the hydrological time series of the several measured variable such as rainfall, evaporation and stream flow involving distribution in probability (Shaw, 1999). (3) data-driven models which attempt to handle hydrological system as black box and try to find a relationship between causal variables such as rainfall and consequent output such as stream flow. These models do not require an explicit well-defined representation of the physical processes and govern equations of the process. Several data driven models have been applied successfully to simulate the rainfall-runoff dynamics of the catchment for example by using artificial neural network (Solomatine and Dulal, 2003), adaptive neuro-fuzzy system (Vernieuwe et al., 2005), evolutionary neural networks (Nazemi et al., 2003), genetic programming (Whigham and Crapper, 2001) and M5 model trees (Solomatine and Xue, 2004).

Application of M5 model trees in hydrology is limited; only some of the published papers are founded, for example to model stage-discharge relationship (Bathettchery and Solomatine, 2005) and for simulating rainfall-runoff process (Solomatine and Xue, 2004). Works of Solomatine and others proved that M5 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 M5 model tree for modeling rainfall-runoff relationship of Tigris catchment area in Baghdad city, capital of Iraq. Also, this paper tries to introduce these techniques to field of water resources of Iraq as alternative tool to manage water resources of country with inexpensive techniques.

M5 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 be used to predict the value of a numeric attribute (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 M5 model tree algorithm was originally developed by Quinlan (1992). Detail

description of this technique is beyond of this paper and can be found in Witten and Frank (1999). 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: (Quinlan, 1994)

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

where T represents a set of example that reaches the node; T_i represents the subset of examples that have the i^{th} 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.

Study area and available data set

The Tigris is one of the largest rivers of the Middle East stretching for over 1900 km, of which, 1415 km are within Iraq with a catchment area of 235000 km^2 Fig.2. The River emerges from the south-west part of Turkey (form lake Hazar) which fed by a series of small watercourses originating from Geldjuk lake at an altitude of about 1200 m above sea level (Iraqi Ministries of Environment, water resources, and Municipalities and public works, 2006). Three large tributaries join the river within the area of Turkey; these are Batman, Garzan, and Bokhtanch. These tributaries provide most of the water in the upper reach of the Tigris. Five large tributaries join the Tigris on its left bank within the area of Iraq: the Greater Zab River, the Rawanduz River, the Lesser Zab River, the Adhaim River, and the Diyala River. In its lower reaches, the Tigris traverses the Mesopotamian plain. The adjoining area is composed of alluvial deposits. From Kut, the river runs through the marshland area. From the Hawizhe Marshes, water return to the Tigris River through the Kassarah and Swaib canals, which join the Tigris respectively 59 km and 23 km upstream of the city of Al-Qurna.

Baghdad is situated in the middle of Iraq; this city lies on the Tigris River at its closest point to the Euphrates, 40 km to the west. The terrain surrounding Baghdad is a flat alluvial plain 34 m above sea level. Historically, the city has been inundated by periodic floods. The city is located on a vast plain bisected by the Tigris River. The Tigris splits Baghdad in half,

with the eastern half being called 'Al-Risafa' and the western half known as 'Al-Karkh'. The land on which the city is built is almost entirely flat and low-lying, being of alluvial origin due to the periodic large floods which have occurred on the river.

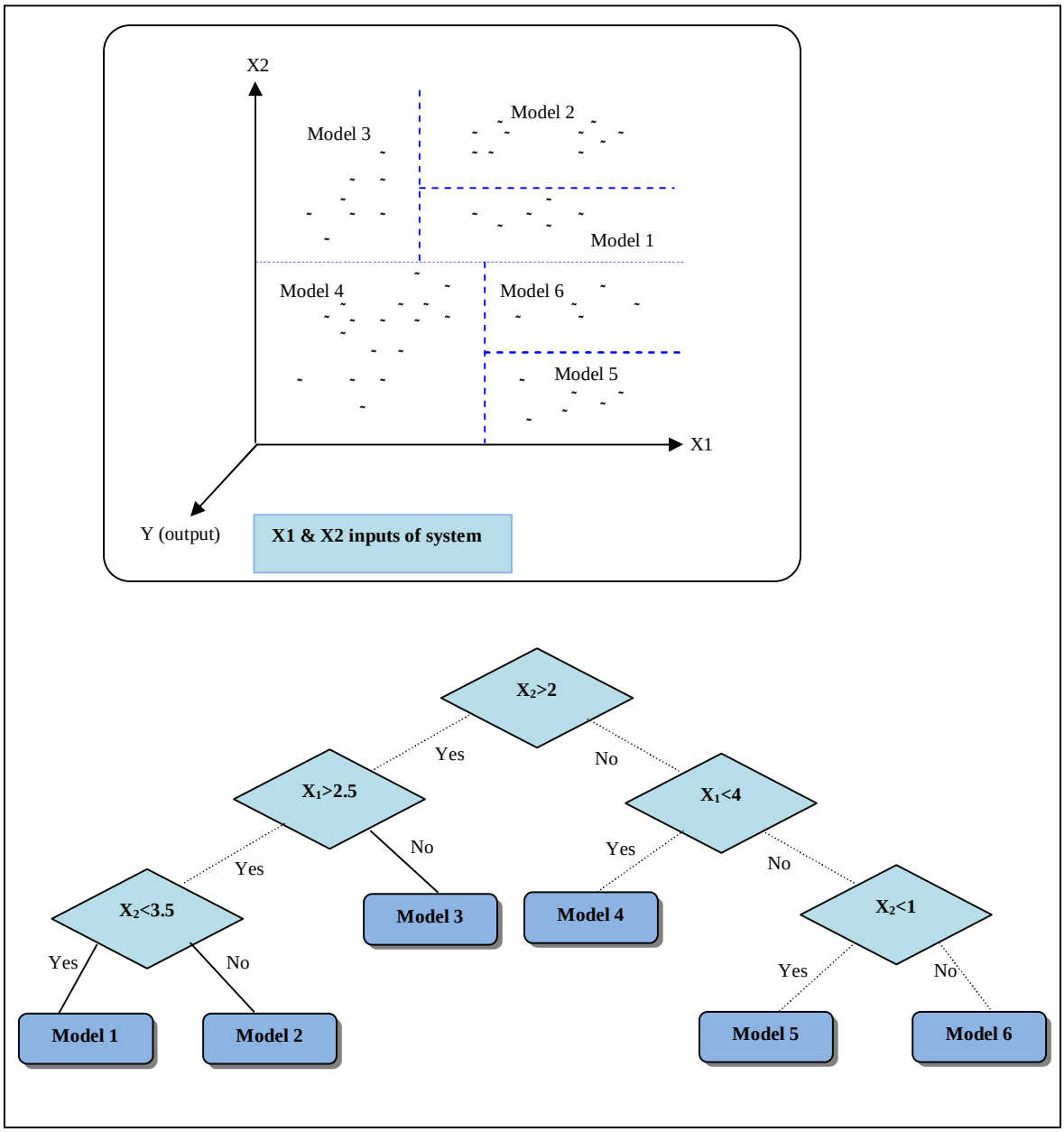

Fig.1: Examples of M5 model 1-6 are linear regression models (modified after Solomatine and Xue (2003))

All data of rainfall was obtained from the report of meteorological survey of Iraq for period 1887-1958 published in 1975, for six stations distributed over the catchment area of river Fig.2. These stations are: Mosul, Khanaqin, Kirkuk, Mandily, Arbil, and Baghdad. Discharge of Tigris River in Baghdad was obtained from previous literature for period 1930-1975 (Al-Ansari, 1975). The data set then tabulated in Excel sheet and for every water year from 1939-1958, the average areal monthly rainfall (calculated by Thessin polygon method) was assigned versus the corresponding monthly discharge. Hence 261 instants are ready to use.

Methodology

For building M5 model, Weka software is used. Weka is open-source machine learning/data mining software written in Java (Witten and Frank, 1999). The software contains a comprehensive set of pre-processing tools, learning algorithm and evaluation methods. It is free via www.weka.com. For 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: 66% for training and reaming for testing.

Data preparation and select the effective input variable are so important to develop a reasonable data-driven model. Therefore, four models are adapted firstly to select the input and output variables and to explore the effect of lag time on the prediction accuracy. This is necessary to analysis the interdependencies between variables and the lag τ. Taking into account the time lags for the input variables is the way to bring the catchment characteristics into data-driven model (Solomatine and Xue, 2004). These models are

1. $Q_{t+1} = f(Q_t, R_t)$
2. $Q_{t+1} = f(Q_t, Q_{t-1}, R_t, R_{t-1})$
3. $Q_{t+1} = f(Q_t, Q_{t-1}, Q_{t-2}, R_t, R_{t-1}, R_{t-2})$
4. $Q_{t+1} = f(Q_t, Q_{t-1}, Q_{t-2}, Q_{t-3}, R_t, R_{t-1}, R_{t-2}, R_{t-3})$

where Q_{t+1} is the discharge at time t+1, Q_t is the discharge at time t, Q_{t-1} is the discharge at time t-1, Q_{t-2} is the discharge at time t-2, Q_{t-3} is the discharge at time t-3, R_t is the rainfall at time t, R_{t-1} is the rainfall at time t-1, R_{t-2} is the rainfall at time t-2, R_{t-3} is the rainfall at time t-3.

Table 1 shows result of this experiment and depending on the correlation of coefficient, model 3 is adapted to examine prediction problem.

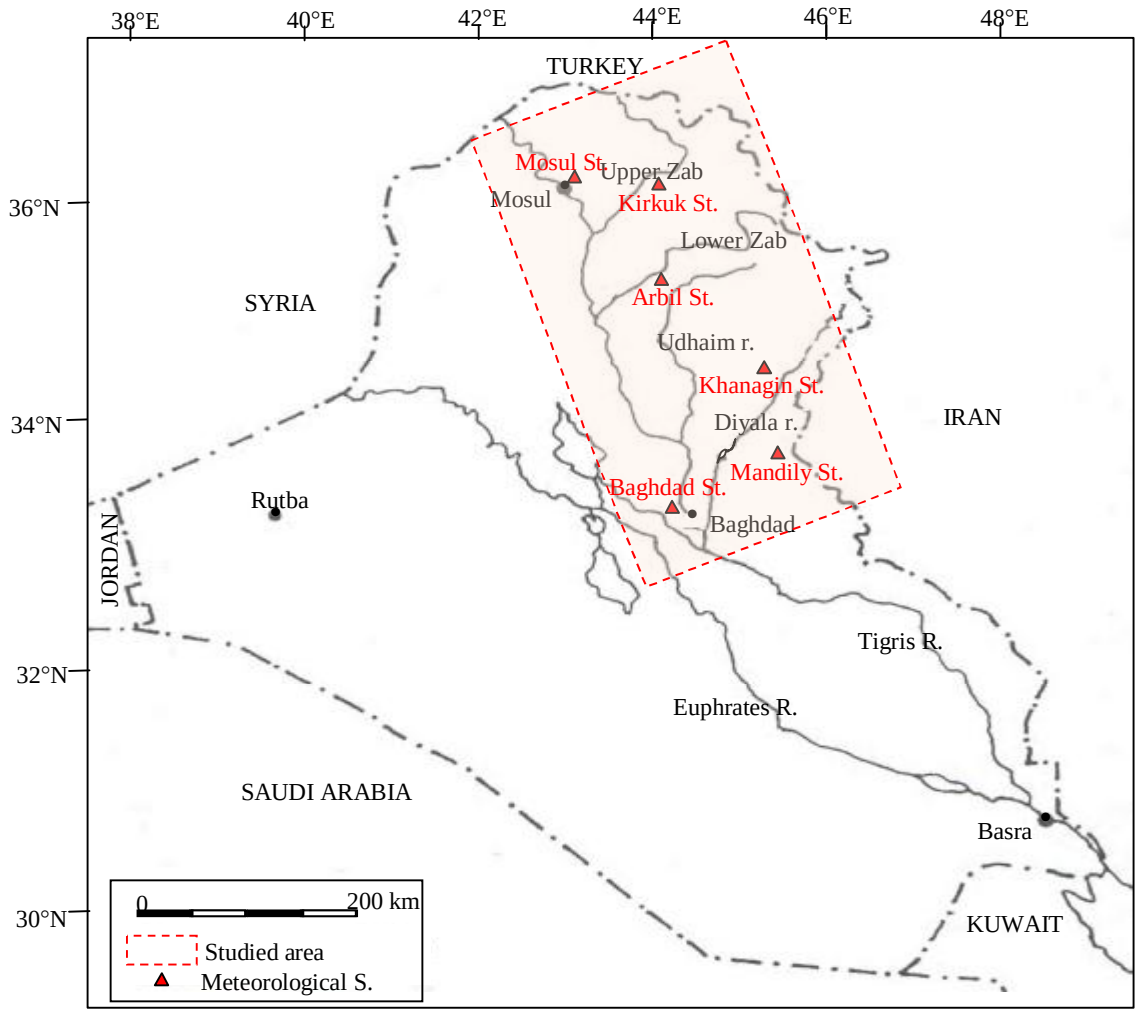

Fig. 2 Location map of the study area

.

Table 1: Effect of lag t on the accuracy of M5 models

Model No.	Evaluation on test split		
	Correlation Coefficient R^2	Number of rules	Time taken to build model (s)
1	0.7206	6	0.81
2	0.7264	5	0.64
3	0.7440	2	0.69
4	0.7158	1	1.08

After select the model which give minimum error, the selected model (model 4) are used to predict Tigris River at Baghdad one and two months ahead, Table 2. Performances of the models are examined by means of the following indices:

1. The coefficient of efficiency (Nash and Sutcliff, 1970)

$$CE = 1 - \frac{\sum_1^n (Q_i - \hat{Q}_i)^2}{\sum_1^n (Q_i - \bar{Q})^2}$$

where Q_i is the measured discharge and $\hat{Q}_i$ is the simulated discharge, $\bar{Q}$ is the arithmetic mean of the measured Q_i, and n is the number of observations (instants).

The optimal value of CE is 1, meaning a perfect match of the model. A value of zero indicates that the model predictions are as good as that of 'no-knowledge' model continuously simulating the mean of the observed signal. Negative values indicate that the model is performing worse than this 'no-knowledge' model (Beven, 2000).

2. Standard error (SE) (Kholghi and Mosseini, 206)

$$SE = \frac{RMSE}{\bar{Q}}$$

where RMSE is the root mean squared error and is defined as

$$RMSE = \sqrt{\frac{1}{n}\sum_i^n (Q_i - \hat{Q}_i)^2}$$

The better the fit, the closer SE to zero.

The similarity between predicted and measured values is investigated by using correlation coefficient R^2 based on this equation:

$$R = \frac{\sum_{i=1}^n (Q_i - \bar{Q})(\hat{Q}_i - \bar{Q}_p)}{\sqrt{\sum_{i=1}^n (Q_i - \bar{Q})^2 (\hat{Q}_i - \bar{Q}_p)^2}}$$

where $\bar{Q}_p$ is the average of predicted values. The best value for R is one, meaning perfect match between simulated and measured values. Vieire et al. (1981) suggested that R^2=0.64 as an acceptable criterion in assessing the estimation process.

3-Results and discussion

Table 2 summarized the error statistics of the experiment and Fig.2 and 3 show a comparison between the measured and simulated monthly average discharges for one and two months ahead prediction. From these figures, it can be seen that low values and some of the high values are predicted with highly accurate but most of high peaks are underestimated. This is because the available data is collected on the monthly basis. If one collects a daily data about the interested variables, the accuracy of the model may increase considerably. This problems is in not solved in Iraq soon due to the fact that there is not a truly plan to monitor the hydrological variables and this is the mean reason to select the old data in this research. Generally, the accuracy of the models is satisfactory. The error statistics support this statement. The final model tree build in this study for month ahead prediction is as follows:

Qt <= 850 : LM1 (127/24.858%)
Qt > 850 : LM2 (128/61.311%)

The following two regression models are generated

LM num: 1

Q_{t+1} = 10.2716 * R_{t-2}+ 3.8783 * R_{t-1}+ 0.7778 * R_t - 0.2115 * Q_{t-2}- 0.0187 * Q_{t-1}+ 0.8792 * Q_t+306.7153

LM num: 2

Q_{t+1} = 0.6647 * R_{t-2}+ 12.0316 * R_t - 0.3757 * Q_{t-2} - 0.0186 * Q_{t-1} + 0.7122 * Q_t + 973.6833

The model tree build for two months ahead prediction is:

R_t <= 0.25 :
| R_{t-1} <= 0.85 : LM1 (90/15.646%)
| R_{t-1} > 0.85 :
| | R_{t-1} <= 2.55 : LM2 (6/12.477%)
| | R_{t-1} > 2.55 :
| | | Q_{t-2} <= 1690 : LM3 (5/23.747%)
| | | Q_{t-2} > 1690 : LM4 (12/8.908%)
R_t > 0.25 :
| Q_t <= 848.5 :
| | Q_t <= 378.5 :
| | | Q_{t-2} <= 351.5 : LM5 (10/14.857%)
| | | Q_{t-2} > 351.5 : LM6 (7/11.418%)
| | Q_t > 378.5 : LM7 (36/52.541%)
| Q_t > 848.5 : LM8 (89/83.565%)

The following eight regression models are generated

LM num: 1

Q_{t+2} = 2.2503 * R_{t-1} + 1712.1008 * R_t - 0.1015 * Q_{t-2} - 0.0861 * Q_{t-1}+ 0.2849 * Q_t + 504.1909

LM num: 2

Q_{t+2} = 8.3446 * R_{t-1}+ 1.2288 * R_t - 0.1757 * Q_{t-2} - 0.0336 * Q_{t-1}+ 0.2136 * Q_t + 626.2154

LM num: 3

Q_{t+2} = 8.9571 * R_{t-1} + 1.2288 * R_t - 0.2115 * Q_{t-2} - 0.0336 * Q_{t-1}+ 0.2163 * Q_t + 818.5974

LM num: 4

Q_{t+2} = 6.9659 * R_{t-1}+ 1.2288 * R_t - 0.1935 * Q_{t-2} - 0.0336 * Q_{t-1}+ 0.2387 * Q_t + 680.9323

LM num: 5

Q_{t+2} = 1.6716 * R_{t-2} + 0.8712 * R_{t-1} - 1.2166 * R_t - 0.5913 * Q_{t-2} - 0.092 * Q_{t-1} + 0.9565 * Q_t + 897.0746

LM num: 6

Q_{t+2} =1.6716 * R_{t-2} + 0.8712 * R_{t-1} - 1.2166 * R_t - 0.6023 * Q_{t-2} - 0.092 * Q_{t-1} + 0.9565 * Q_t+ 858.5818

LM num: 7

Q_{t+2} =1.6716 * R_{t-2} + 0.8712 * R_{t-1} - 0.3901 * R_t - 0.8623 * Q_{t-2} - 0.092 * Q_{t-1} + 0.6832 * Q_t + 1412.4231

LM num: 8

Q_{t+2} = 1.093 * R_{t-2} + 0.8712 * R_{t-1}+ 7.3272 * R_t - 0.613 * Q_{t-2} - 0.284 * Q_{t-1}+ 0.4351 * Q_t + 2276.7514

Table 2 Error statistics of the experiment

Prediction horizon	Training			Verification		
	R^2	CE	SE	R^2	CE	SE
t+1	0.74	0.8	0.25	0.73	0.75	0.30
t+2	0.79	0.68	0.47	0.65	0.53	0.56

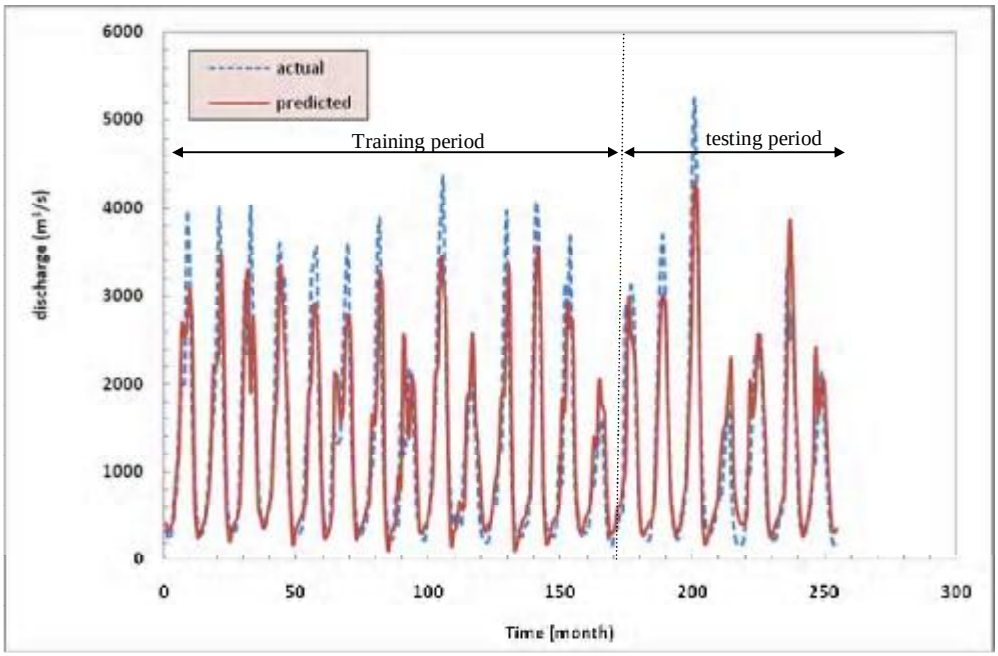

Fig.3 Comparison between measured and simulated discharges for one month ahead prediction

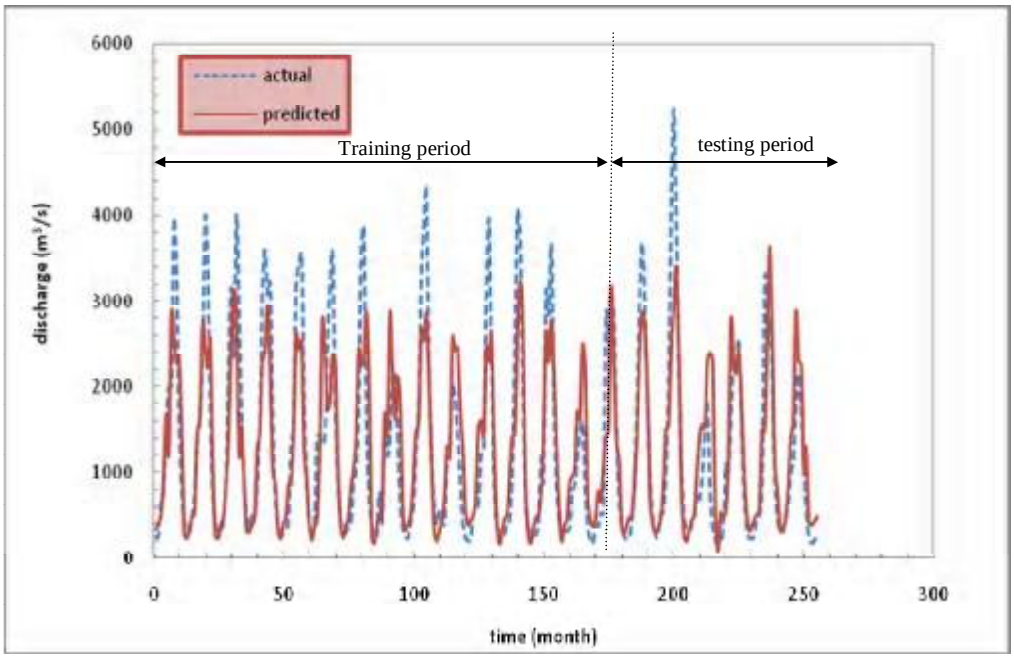

Fig.4 Comparison between measured and simulated discharges for two months ahead prediction

4-Conclusions

M5 model tree is an efficient tool to emulate rainfall-runoff relationship. The results of M5 model tree is 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 hyprid model with physically models to improve situation of that worse managed resource.

5-References

Al-Ansari, N. (1979) Principle of Hydrology, Baghdad, 340p. (in Arabic)

Beven, K. (2000) Rainfall- Runoff Modeling, the Primer, Wiley Chichester, UK.

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

Iraqi Ministries of Environment, Water Resources, Municipalities and Public Works (2006) Overview of present conditions and current use of the water in the water resources marshland area, Book1, Italy-Iraq, 146p.

Kholghi, M. and Hosseini, S. M. (2006) Estimation of aquifer transmissivity using kriging, artificial neural network, and neuro-fuzzy models, *J. of Spatial Hydrology*, **6**(6):68-81

Nash, J. E., Sutcliff, J. V. (1970) River flow forecasting through conceptual models. *J. Hydrol*. 10, 282-290.

Nazemi, A. Hossein P. N., Mohammed R., Akbarzadeh, T., and Seyed M. (2003) Evolutionary neural network modeling for describing rainfall-runoff process. *Hydrology Days*, 224-235.

Quinlan, J. R. (1992) Learning with continuous classes, Proc. A192, 5th Australian Join Conference on Artificial Intelligence, Adams and Sterling (eds.). World Scientific, Singapore, pp 343-348.

Shaw, M. E. (1999) Hydrology in Practice, Stanly Thornes Ltd, Glos, UK, 569 p.

Solomatine, D. P. & Dulal, K. N. (2003) Model tree as alternative to neural network in rainfall-runoff modeling, *Hydrological Sci. J.* **48**(3), 399-411.

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.

Vernieuwe, H. Georgieva, O., Baets, B., Pauwels, V., Verhoest, N. and Francois, T. (2005) Comparison of data-driven Takagi-Sugeno models of rainfall-discharge dynamics. *J. of Hydro.*, 302, 171-186.

Vieira, S. R., Neilsen, D. R., Biggar, J. W. (1980) Spatial variability of field measured infiltration rate. *Soil Sci. Soc. Am. J.*, 45, 1040-1048.

Whigham, P. A. & Crapper, P. E. (2001) Modeling rainfall-runoff relationship using genetic programming, *Mathematical and Computer Modeling* 33, 707-721.

Witten, I. H. & Frank, E. (2001) Data Mining, Morgan Kaufmann, San Francisco.

نمذجة عملية تحول الأمطار - السيول باستخدام تقنية النماذج الشجرية

علاء محسن عطية و حسين بدر غالب

قسم علم الأرض، كلية العلوم، جامعة البصرة، البصرة، العراق

الملخص

يتحرى البحث الحالي عن إمكانية استخدام أحدى تقنيات النمذجة الشجرية Decision trees والمعروفة باسم M5 وهي أحدى تقنيات الذكاء الصناعي Artificial intelligence وتعليم الآلات Machine learning لمحاكاة عملية تحول الأمطار - السيول لحوض تصريف نهر دجلة في محافظة بغداد وسط العراق. لغرض إنشاء النموذج المطلوب استخدم البرنامج المعروف باسم Weka . بنيت اربعة نماذج اولاً لمعرفة تأثير التصاريف ومعدل الامطار السابقة على مقدار التنبؤ وفي اختيار المدخلات الفاعلة في بناء النموذج المطلوب. اختبرت إمكانية النموذج المبني باستخدام هذه الطريقة من خلال التنبؤ بتصريف نهر دجلة في منطقة بغداد لشهر وشهرين لاحقيين. بينت النتائج الدقة العالية لهذه التقنية لمحاكاة قيم التصاريف الواطئة وبعض التصاريف العالية ، في حين كانت معظم قيم التصاريف العالية المتنبئ بها تحت القيم الحقيقية underestimated ، واعزي السبب إلى نوعية المعلومات المتوفرة في بناء النموذج والتي كانت معدل التصاريف الشهرية للنهر ومعدل مجموع الأمطار الشهري الساقط على الحوض. خرجت الدارسة باستنتاج مفاده إمكانية استخدام مثل هذه النماذج لوحدها آو من خلال دمجها مع النماذج المعتمدة على القوانين الفيزيائية Physically based models مثل برنامج HEC-HMS لإدارة الموارد المائية في العراق بعد البدء ببرنامج تحريات هيدرولوجي واسع لهذا المصدر الوطني المهم للغاية.