VOL. 13, NO. 12, JUNE 2018 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

ASSESSING THE INFLUENCE OF DECISION TREE FOREST PARAMETERS ON PREDICTION OF MONTHLY RIVER DISCHARGE

Ali H. Al-Aboodi, Husham T. Ibrahim and Sarmad A. Abbas
Department of Civil Engineering, College of Engineering, University of Basrah, Basrah, Iraq
E-Mail: alialaboodi90@gmail.com

ABSTRACT

The main objective of this paper is to determine the optimum values of the main parameters of Decision Tree Forest (DTF) (Maximum Trees in Decision Tree Forest (MTDTF) , Minimum Size Node to Split (MSNS) , and Maximum tree levels (MTL)), and assessing of their effect on predicted average monthly discharge of Euphrates River, in Thi Qar province, southern Iraq. Four popular statistical parameters were used as evaluation criteria for evaluating DTF models performance: Coefficient of Correlation (R), Root Mean Squared Error (RMSE), Maximum Error (ME), and Mean Absolute Percentage Error (MAPE). Five-fold cross-validation is applied in this research to evaluate the performance of the DTF models. In the first stage of this study, the observed data volume involved in DTF models was equal to 240 months; in this stage the results illustrated that the optimum values of both MTDTF, and MSNS were equal to 125 and 32 respectively, while the modification of the third parameter MTL does not show any effect on the performance of DTF, it was observed that the effect of this parameter on statistical parameters in the form of a straight line. In the second stage of this study, the observed data volume was increased to 480 months ; thus leading to increase the optimum value of MTDTF to 225, and decrease the optimum value of MSNS to 2, while the results does not show any sensitivity to the parameter MTL.

Keywords: decision tree forest, Euphrates river, maximum trees in decision tree forest, minimum size node to split, maximum tree levels.

INTRODUCTION

Discharge through a river is complex nonlinear hydrological process that shows a high degree of spatial and temporal variation. The prediction of discharge and estimation the parameters of this process as accurate value plays an effective role in any decision-making process related to water availability, such as irrigation projects management, management of water resources, construction and management of hydropower plants, construction of water pumping stations and management of urban water supply systems, and many more.

Three main types of forecasting models; (1) conceptual models (2) physically based models and (3) data-driven models [1]. The latter model is quite different from the first two models, involves non-derivative mathematical equations from physical processes in watersheds but from creating empirical equations through analysis of observed time series data. The best model is a model that gives results close to reality with the use of lower number of parameters and low degree of complexity. The models are used primarily to predict system behavior and understand different hydrological processes.

The model consists of different parameters that specify the properties of the model. Conceptual model describes all components of hydrological processes. Semi empirical equations are used in this model and the model parameters are evaluated not only from field data but also through calibration [2]. Model calibration requires changing the values of model input parameters in an attempt to match the observed data with calculated data within certain acceptable criteria. Modifications to model parameters, stresses and boundaries will be limited within reasonable ranges based on available information. The trial and error calibration procedure has several disadvantages, it remains uncertain whether the calibrated model parameter values are the optimal value based on the evaluation criteria, it requires accurate knowledge of the model parameters and the permissible range, and the multiplicity of parameters in the model leads to the dispersion of focus on the most important parameters, which will take a long time to carry out the task. Physically based model is an ideal mathematical representation of the real phenomenon. It uses state variables that can be measured and are functions of time and space. It does not require extensive hydrological and meteorological data to be calibrated, but the assessment of a large number of parameters describing the physical characteristics is required [3]. In this model a large amount of data such as surface water depth, topography, river bed roughness, river cross section dimensions etc. are required. Black-box models or data-driven models take information from observed data without considering the properties and processes of the hydrological system. It's a mathematical equation derived from input and output data without the need to know the hydrological processes. These models are valid within certain limits. There are several types of data driven models, such as unit hydrograph model, statistically based model use correlation method to find the relationship between input and output (which include: Autoregressive moving average (ARMA) models, Autoregressive integrated moving average (ARIMA) models, and linear regression), and machine learning (ML) models [4].

VOL. 13, NO. 12, JUNE 2018 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

In particular, flow prediction was accepted with artificial neural network (ANN) models as a good alternative to prediction with hydrodynamic models [5, 6]. ANN has been applied extensively in water resources modeling, such as flood forecasting, rainfall-run off modeling, evaporation estimation, water quality modeling, groundwater modeling, and stage-discharge relationship [7-17]. Fuzzy Inference System (FIS) and Adaptive Neuro Fuzzy Inference System (ANFIS) are other types of data-driven models, which have been widely applied in water resources modeling. These models have been successfully used to obtain accurate results for flood forecasting and rainfall-runoff modeling [18-22].

Decision Tree Forest (DTF) is a one of data driven techniques, it is a collection of decision trees, where each prediction is combined to produce the comprehensive prediction of the forest. A DTF and a tree Boost technique are similar in terms of that a large number of trees are grown. A DTF grows parallel and independent trees and do not intersect until their growth is complete. Three main parameters (Maximum Trees in Decision Tree Forest (MTDTF), Minimum Size Node to Split (MSNS), and Maximum tree levels (MTL)) control the accuracy of DTF results and the mechanism of prediction. The main object of this research is determining the optimum values of these parameters and assessing of their effect on predicted average monthly discharge of Euphrates River, in Thi Qar province, south of Iraq.

STUDY AREA AND DATA COLLECTION

Euphrates River is the longest river in West Asia and one of the most important rivers, Iraq was called Mesopotamia because of two longest rivers, the Tigris and the Euphrates River. Euphrates River emerges from the Turkish territory then passes through the Syrian territories and then enters Iraq to confluence the Tigris River in Basrah Province; southern Iraq, forming Shatt Al-Arab River, which empties into the Arabian Gulf. Rainfall and melting snow contribute to the most of Euphrates River water. The greatest flow of the Euphrates River occurs during the months April through May. 36% of the total annual discharge of the Euphrates occurs in these two months [23].Thi Qar Province lies on the banks of the Euphrates River, 370 km southeast of Baghdad. Thi Qar has a dry desert climate. The summer is hot and dry, with average high temperatures up to 40 ° C while the winter is mild. Rainfall is occurring during the period from November to April and an average of 100 mm per year. Thi Qar shares internal borders with the provinces of Qadissiya, Muthanna, Wassit, Missan, and Basrah as shown in Figure-1. The averages monthly discharges of Euphrates River in Thi Qar City for the period (1975-2015) are presented in this research. The observed discharges were obtained from the Iraqi Ministry of Water Resources. Table-1 provides summary statistics on the observed discharge. The difference between the maximum and minimum discharge value is a large extent (301 m^3 / sec), the values of the data set are widely scattered. The lack of symmetry of a distribution is called skewness. Skewness is a measure of the asymmetry of the probability distribution of a random variable about its mean. Negative skew indicates that the tail on the left side of the probabilistic density function is longer than the right side. In the normal distribution, the excess kurtosis is equal to zero. When the excess kurtosis is below than zero, this value indicates the tails are lighter than the normal distribution. An excess kurtosis value of 1 and above or -1 and below indicates a significant deviation from normality.

Table-1. Summary statistics of the raw data.

Statistics parameter	Value
Average (m^3/sec)	151.45
Standard Deviation (m^3/sec)	64. 36
Skew	-0.092
Excess Kurtosis	-0.64
Median (m^3/sec)	157.58
Minimum (m^3/sec)	9
Maximum(m^3/sec)	310
1st Quartile (m^3/sec)	96.36
3rd Quartile (m^3/sec)	197.65

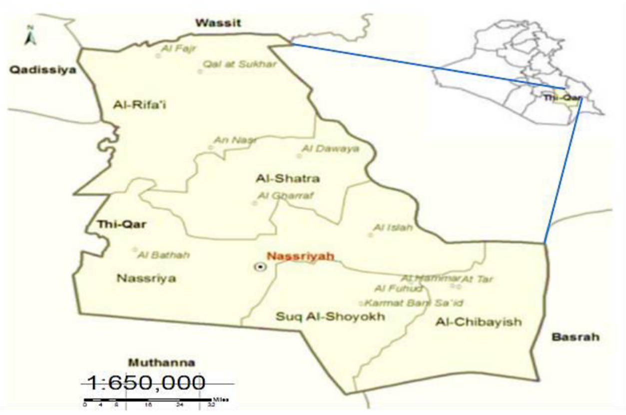

Figure-1. Location of study area in reference to map of Iraq [24].

Decision Tree Forest (DTF)

DTF algorithm was first presented by Leo Breiman in 2001 [25]. A DTF model is similar to a Treeboost model [26]. A large number of trees are using in both models, but the main difference in the mechanism of learning process between these two models is that in the Treeboost model, trees are grown in series so that the output data is supplied from tree to next tree. In contrast, a DTF model is a group of independent trees that are grown in parallel. Three parameters must be selected to adjust the behavior of DTF models; MTDTF, MSNS, and MTL. The main object of this research is determining the optimal values of these parameters and their effect on predicted monthly discharge of Euphrates River. The model containing large numbers of trees in the forest gives high precision results. When determining the value of MSNS, a

VOL. 13, NO. 12, JUNE 2018 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

node in a tree in the forest will not be split if it is less than that number of rows in it. The last parameter is MTL; means the maximum number of levels (depth) in each tree can be grown in the forest, some researchers suggest that it is better to grow very large trees, so the maximum tree level should be set to a large level. The outline of the algorithm used to construct a DTF as shown in the following points:

- Select a random sample of (N) observations from data set, this process is named "bagging". Some observations are selected again and others are not selected. 2/3 of the rows are selected during sampling process. The remaining rows are called "out of bag (OOB)". A new random selection of rows is made for each constructed tree.
- The selected rows from step (i) is used for constructing a decision tree as shown in Figure. 2. The tree does not prune when it is built to the maximum size level. A subset of the total set of predictor variables is selected to be considered as possible splitters for each node. A new random selection of variables is achieved for each split. Some predictors (perhaps the best one) will not be considered for each split, but a predictor excluded from one split may be used for another split in the same tree.
- Repeat steps i and ii for many times for constructing a thick forest of trees.
- For evaluating a row, run the row through each tree in the forest and record the predicted value (that is, the terminal node). The average score of trees is determined for regression analysis.

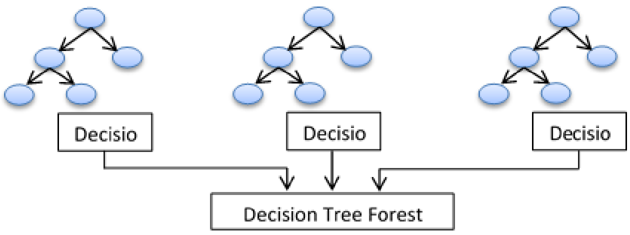

Figure-2. Basics of DTF architecture.

METHODOLOGY

The averages monthly discharges of Euphrates River were used in this research during the period (1975-2015). Data of average monthly discharge released from Euphrates River at three antecedent time steps (Q (t-1), Q (t-2), and Q (t-3)) were considered as input to the DTF models. The total data is extended over a period of 40 years (480 months); these data are divided into two equal periods, each period representing 20 years (240 months), in order to estimate the effect of the data volume on the model parameters. Accordingly, six different DTF models with their data volume (Table 2) were proposed and their performance compared to determine the best model. For data volume (240 months), model No. 1 represents the average monthly discharge at time (t) is Q (t) as a function of average monthly discharge corresponding to one time step lag (t-1) (Table 2). Likewise, Q (t) = f (Q (t-1), Q (t-2)) represents the monthly average discharge at time t, being a function of the average monthly discharge release at (t–1) and (t–2). The third model was proposed by considering the integrated effect of the release values up to three antecedent time steps. Same models above were used for volume data (480 months). The DTF simulations, and analysis of the results, were performed using DTREG (Predictive Modeling Software) [27].

Several evaluation criteria have been applied in this paper to evaluate model performance. Appropriate evaluation criteria for any mathematical model are important when using a multi-criteria analysis to validate the performance of a model. In this research, the following four popular statistical parameters were used as evaluation criteria for evaluating DTF models performance: coefficient of correlation (R), root mean squared error (RMSE), maximum error (ME) (the maximum difference between the observed value and the predicted value), and mean absolute percentage error (MAPE). The mathematical equations of (R, RMSE, and MAPE) are shown in equation (1), (2), and (3) respectively. The evaluation criteria describe the degree of accuracy of the developed model.

$$R = \frac{\sum_{i=1}^{N}(x_i-\bar{x})^2(y_i-\bar{y})^2}{\sqrt{\sum_{i=1}^{N}(x_i-\bar{x})^2\sum_{i=1}^{N}(y_i-\bar{y})^2}} \tag{1}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(y_i-x_i)^2}{N}} \tag{2}$$

$$MAPE = \frac{100}{N} \times \sum_{i=1}^{N}\left|\frac{x_i-y_i}{x_i}\right| \tag{3}$$

The observed data are subdivided into two parts; training and testing. The training data are presented to the network during training phase, according to the training error, the architectural of network and its parameters are modified. The remaining data (i.e., testing data) are used to check the performance of the trained black box model; these data are not used previously in the training phase. Using different percentage values of training and testing dataset may lead to different results and may be to different conclusions. To overcome this problem and to reach the same conclusions, cross validation method is presented in this paper. The popular v-fold cross-validation, which provides a good trade-off between model under-fitting and over-fitting, has been used to assess the performance of candidate models [28]. The observed datasets (average monthly discharge recorded from 1975 to 2015) were divided randomly into v equal size subgroups. During the modeling process for each DTF model, one of the partitions was used for training phase, while another was used for validation phase. The modeling process was repeated v times, and then the evaluation criteria were averaged to check the final performance. According to previous studies, using a v of 5, 10, and 20, it can produce very similar errors, which are often slightly difference and not significant [28, 29]. Therefore, five-fold cross-validation is applied in this research to evaluate the performance of the DTF models. There are three significant parameters for constructing a DTF model (MTDTF, MSNS, and MTL).

Table-2. Different inputs combination with data volume of DTF models.

Model No.	Data volume (month)	Input variable	Target variable
1	240	(Qt-1)	Q(t)
2	240	(Qt-1,Qt-2)	Q(t)
3	240	(Qt-1,Qt-2,Qt-3)	Q(t)
4	480	(Qt-1)	Q(t)
5	480	(Qt-1,Qt-2)	Q(t)
6	480	(Qt-1,Qt-2,Qt-3)	Q(t)

In general, the more trees used to construct a DTF model lead to more accurate results, in contrast; the improvement in the accuracy of the results decreases with the increase in the number of trees after a specified number of trees. At a certain point the interest in regression performance from learning extra constructed trees will be lower than the cost of calculation time for learning theses additional trees. There is another parameter that controls the size node to split; a node in a tree will not be split if it has value less the value of MSNS, this is a suitable process to reduce tree growing. The third important parameter is MTL, which controls the maximum limits (depth) in which each tree can be grown in the forest. In order to evaluate the parameters of the DTF models and to study their effect on the values of average monthly discharges as well as the calculation of the optimal values of these parameters, the following process was achieved, initially selecting two parameters such as (MSNS, and MTL) to be fixed as a default value and adjusting the value of parameter (MTDTF) then run the model and check the evaluation criteria to find the optimal value of this parameter and assessing its effect on predicted discharge. The same process was repeated to another parameter (MSNS) to find its optimal value and to evaluate its effect on the discharge values, but here the optimal value of the parameter (MTDTF) obtained from the first process is identified and not changed. Likewise, same process is repeated for third parameter (MTL), but with a relative difference is to fix the values of the previous parameters and change the value of the current parameter. These above processes are applied on two stages, the first stage for data volume (240 months), while the second stage for data volume (480 months).

RESULTS AND DISCUSSIONS

This research used the DTF model to predict averages monthly discharges of Euphrates River in Thi Qar City; southern Iraq, for the period (1975-2015). Six different DTF models were proposed and their performance compared to determine the optimum values of the main parameters of DTF (MTDTF, MSNS, and MTL), and assessing of their effect on predicted values. For data volume (240 months), the effect of MTDTF on the performances of developed DTF models (Model No. 1, 2, & 3) was done, by fixed the default values for parameters (MSNS, and MTL) of 10 and 50 respectively, and changing the value of MTDTF from 25 to 250 as shown in Figures (3 & 4).The results show that the optimum value of MTDTF was equaled to 125, the model No. 3 was better than the other two models (1&2) with (R, RMSE,

ME and MAPE) of 0.860, 34.148, 103.508 and 26.073, respectively. The results also show that the input combination variables (Qt-1, Qt-2) increases the model's performance of model No.2 by reducing RMSE, ME, and MAPE of 7.37%, 9.47% and 8.33% respectively, and increases R by 2.76% compared with model No.1. There is no significant improvement in the results of model No. 3 when adding the variable (Qt-3) to the input variables (Qt-1, Qt-2) as can be seen from the results of evaluation criteria of model No.3. It was also noticed through the Figures (3&4) that increasing the number of trees in the forest does not necessarily lead to improvement of results, there is a gradient in the improvement of the performances up to the value of MTDTF equaled to 125 after which there was a decrease in the performances of the DTF models (1, 2, & 3) at the value of MTDTF equaled to 150 and then the effect of MTDTF on the evaluation criteria is continued as semi-straight line, where the model performances were not affected by increasing the value of MTDTF. At a certain point the interest in regression performance from learning extra constructed trees will be lower than the cost of calculation time for learning these additional trees.

The same process was repeated to the second parameter (MSNS); the values of the parameters (MTDTF, and MTL) has been fixed on 125 and 50 respectively, and changing the value of MSNS from 2 to 44 as illustrated in Figures (5&6). It was observed that by reviewing all studied cases and evaluation criteria, there was no significant improvement in the efficiency of the model No.3 compared to the model No.2. The optimum value of MSNS was equal to 32. The performances of model No. 3 was better than model No.1 and very slightly better than model No.2. The statistical evaluation criteria ((R, RMSE, ME and MAPE) of model No.3 were equaled to (0.873,

ARPN Journal of Engineering and Applied Sciences

32.433, 99.109, and 24.846) respectively. It was obvious that the modification of parameter (MSNS) from default value to (32) lead to significant improvement in the efficiency of model No.3. The percentage ratio of R is increased by 1.5% and the percentage ratios of (RMSE, ME and MAPE) are decreased by (5.0%, 4.25%, and 4.71%) respectively. Also can clearly notice that there is deterioration in the efficiency of the models after the value of 32 and then the effect of MSNS on the evaluation criteria is continued as semi-straight line, there is no improvement in the performance of models after this value and increasing the value of this parameter is useless.

Similarly, the same process was repeated to the third parameter (MTL), but with a relative difference, the previous parameter values are set to optimal values and changing the current parameter value from 25 to 250 as shown in Figures (7&8). The results illustrated that the performances of model No. 3 is better than model No.1 and very slightly better than model No.2. There is no significant difference in evaluation criteria between model No.2 and model No.3. The statistical evaluation criteria (R, RMSE, ME and MAPE) of model No.3 are equaled to (0.873, 32.433, 99.109, and 24.846) respectively. There is no any effect of MTL on the evaluation criteria. The modification of this parameter does not lead to any improvement in the results, and the effect of this parameter in the form of a straight line.

The second stage of this research is applied DTF models on data volume (480 months). The data size was multiplied to study the effect of data volume on model parameters. The effect of MTDTF on the performances of developed DTF models (Model No. 4, 5, & 6) was presented in Figures. (9 & 10). The optimum value of MTDTF is equal to 225. The performances of model No.6 is better than model No.4 and very slightly better than model No.5. The evaluation criteria (R, RMSE, ME and MAPE) of model No.6 are equaled to (0.942, 22.421, 75.415, and 15.721) respectively. It became clear that the increasing of data volume would lead to improve of model performance, where the regression performance from learning additional data caused to improve the efficiency of DTF models. The percentage ratio of evaluation criteria (R, RMSE, ME and MAPE) for model No.6 are improved by (7.90%, 30.87%, 23.91%, and 36.73%) respectively compared with performances of model No.3. There is a flat gradient in the improvement of the performances except the maximum error up to the number of maximum trees equaled to 225 after which there was a decrease in the performances of the DTF models (4, 5, & 6). The optimum value of MSNS is equal to (2) as shown in Figures. (11&12). The increasing of MSNS value led to deterioration in the efficiency of the DTF models after the value of 2. Clearly, the size of the data has greatly affected the value of this parameter. There is an inverse correlation between the value of this parameter and the size of the data involved in the regression. It was observed, the dramatically change in the value of MTL does not lead to any amelioration in the evaluation criteria as presented in Figures (13&14), where its effect is similar to the straight line along the phase of the parameter change. Tables (3&4) show the optimum evaluation criteria for the main parameters of DTF models with observed data volume equaled to (240 months) and (480 months) respectively.

Table-3. Optimum evaluation criteria of DTF models with observed data volume equaled to (240 months).

Parameter	Optimum Value	Model 1				Model 2				Model 3			
		R	RMSE	ME	MAPE	R	RMSE	ME	MAPE	R	RMSE	ME	MAPE
MTDTF	125	0.833728	37.02102	114.3383	28.4428	0.857437	34.2882	104.161	26.1962	0.859778	34.14784	103.50816	26.07326
MSNS	32	0.855875	34.18037	112.225	26.54579	0.870136	32.6286	100.3691	25.22972	0.873191	32.43379	99.10958	24.84579
MTL	25-250	0.855875	34.18037	112.225	26.54579	0.870136	32.6286	100.3691	25.22972	0.873191	32.43379	99.10958	24.84579

Table-4. Optimum evaluation criteria of DTF models with observed data volume equaled to (480 months).

Parameter	Optimum value	Model 4				Model 5				Model 6			
		R	RMSE	ME	MAPE	R	RMSE	ME	MAPE	R	RMSE	ME	MAPE
MTDTF	225	0.902606	28.80004	80.19167	20.01582	0.936142	23.07936	75.85243	16.47459	0.94171	22.42086	75.415154	15.72126
MSNS	2	0.908992	27.71613	72.01568	16.79860	0.974989	15.1426	52.34844	11.31388	0.979323	13.88869	49.144643	10.18740
MTL	25-250	0.908992	27.71613	72.01568	16.79860	0.974989	15.1426	52.34844	11.31388	0.979323	13.88869	49.144643	10.18740

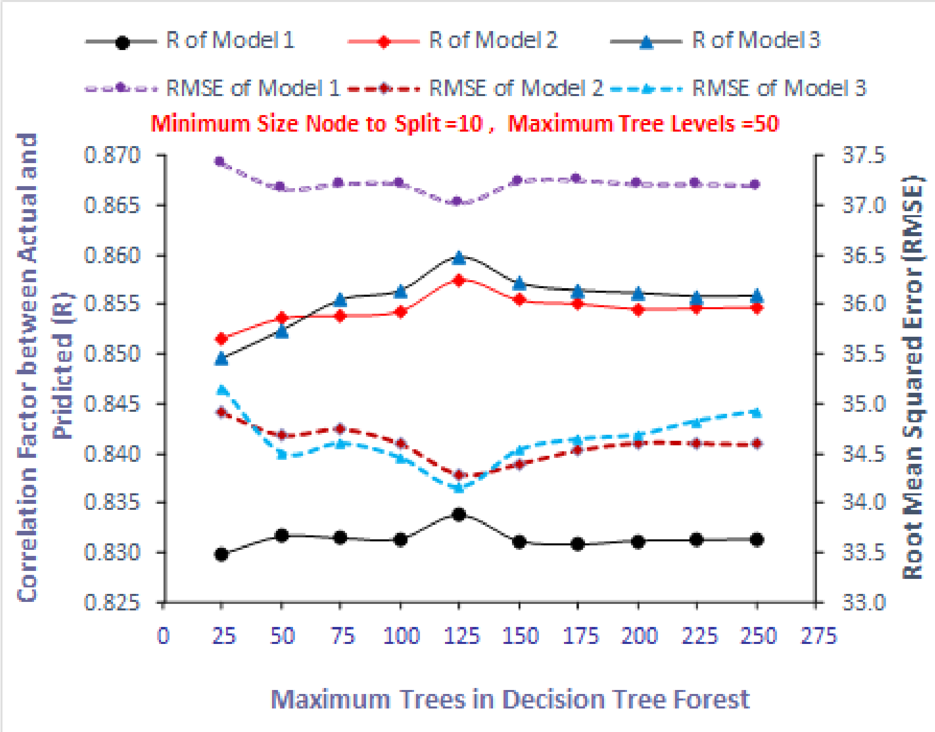

Figure-3. Effect of maximum trees in decision tree forest on the values of R and RMSE for model No. (1, 2, and 3).

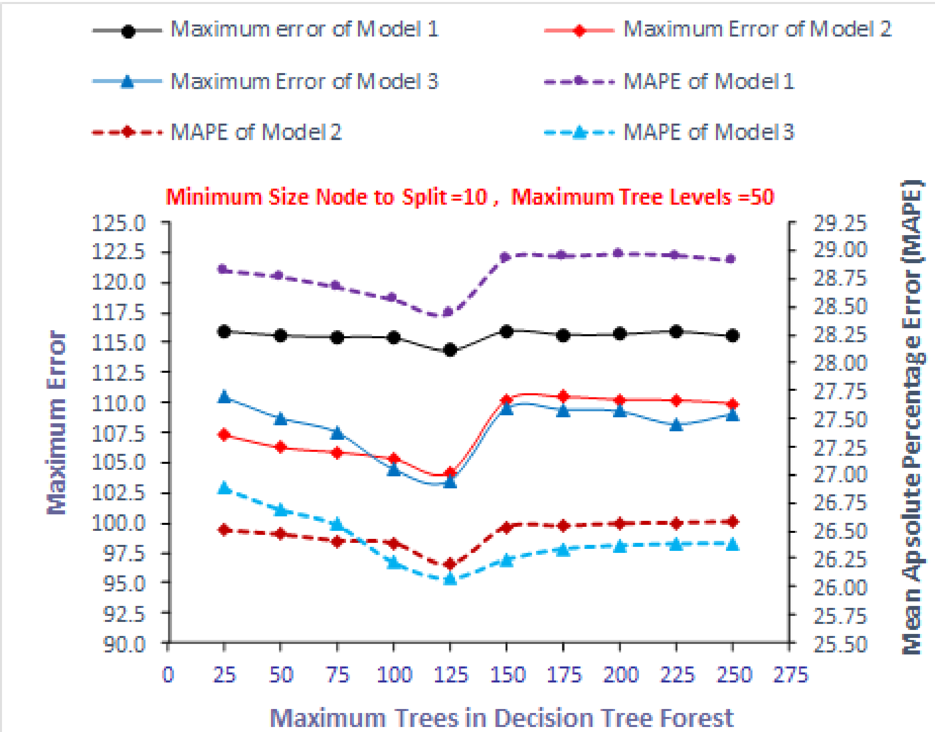

Figure-4. Effect of maximum trees in decision tree forest on the values of ME and MAPE for model No. (1, 2, and 3).

ARPN Journal of Engineering and Applied Sciences

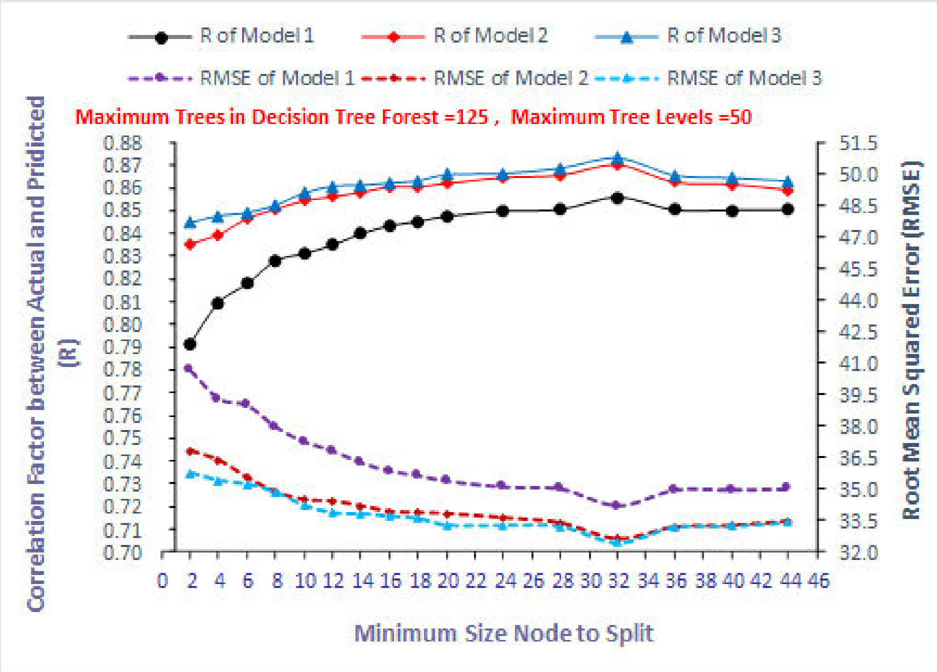

Figure-5. Effect of minimum size node to split on the values of R and RMSE for model No. (1, 2, and 3).

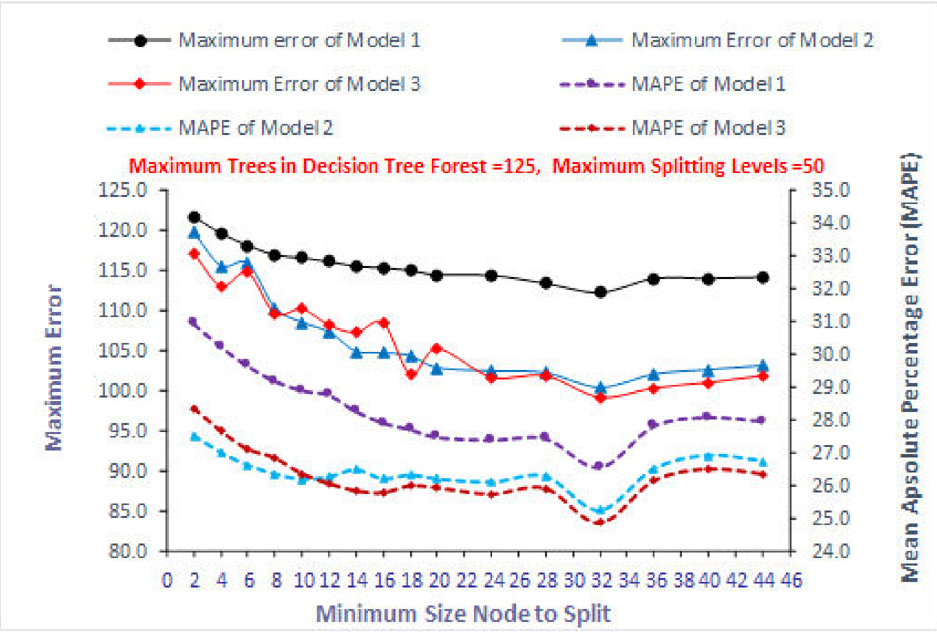

Figure-6. Effect of minimum size node to split on the values of ME and MAPE for model No. (1, 2, and 3).

VOL. 13, NO. 12, JUNE 2018 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

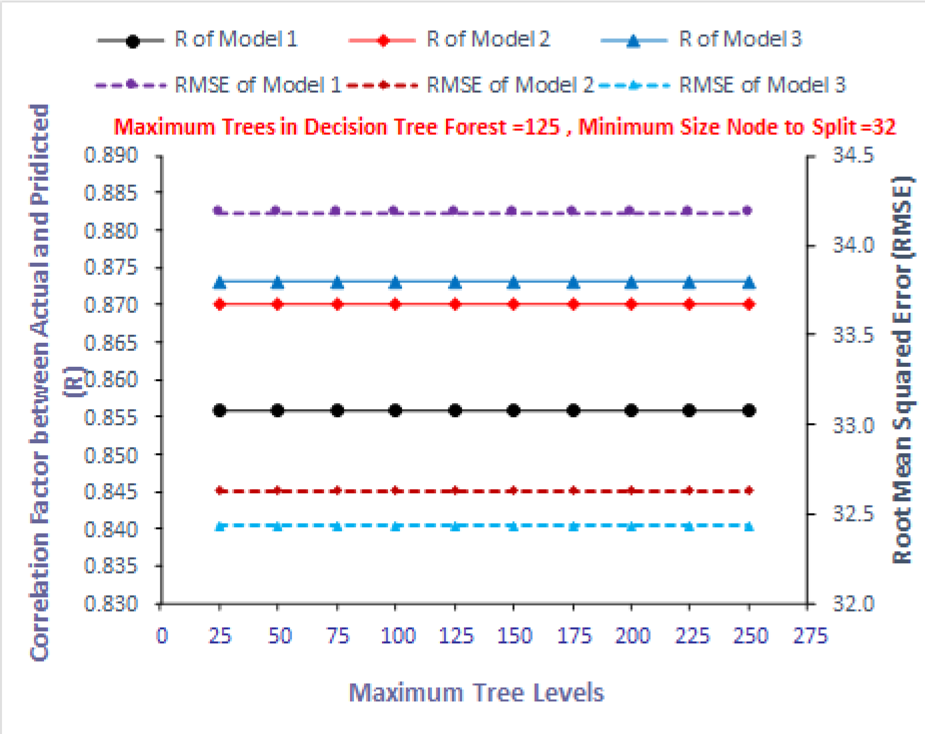

Figure-7. Effect of maximum tree levels on the values of R and RMSE for model No. (1, 2, and 3).

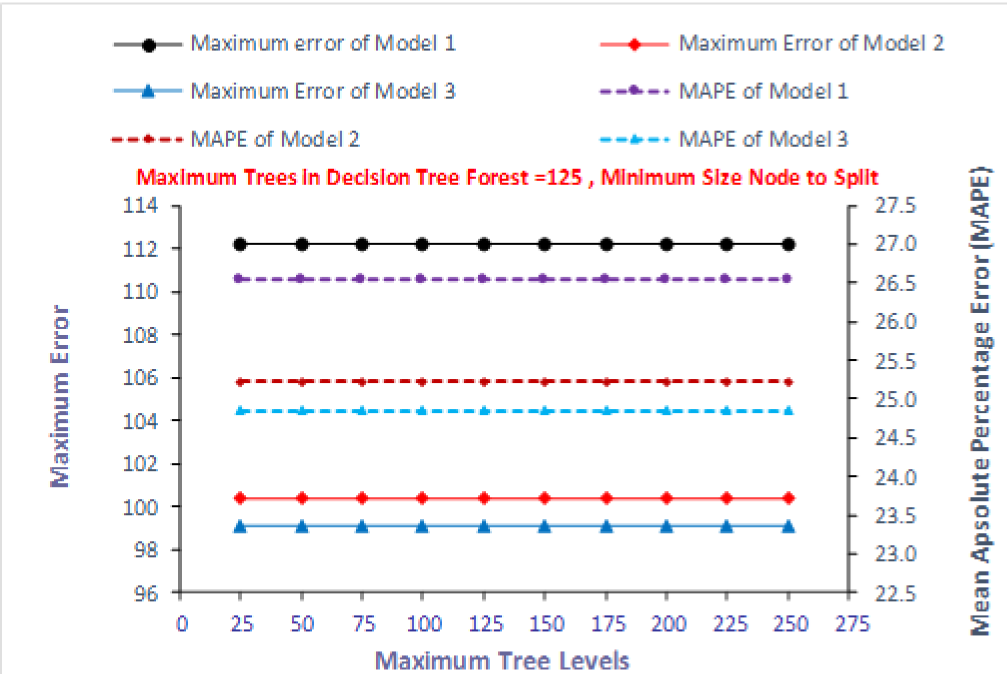

Figure-8. Effect of maximum tree levels on the values of ME and MAPE for model No. (1, 2, and 3).

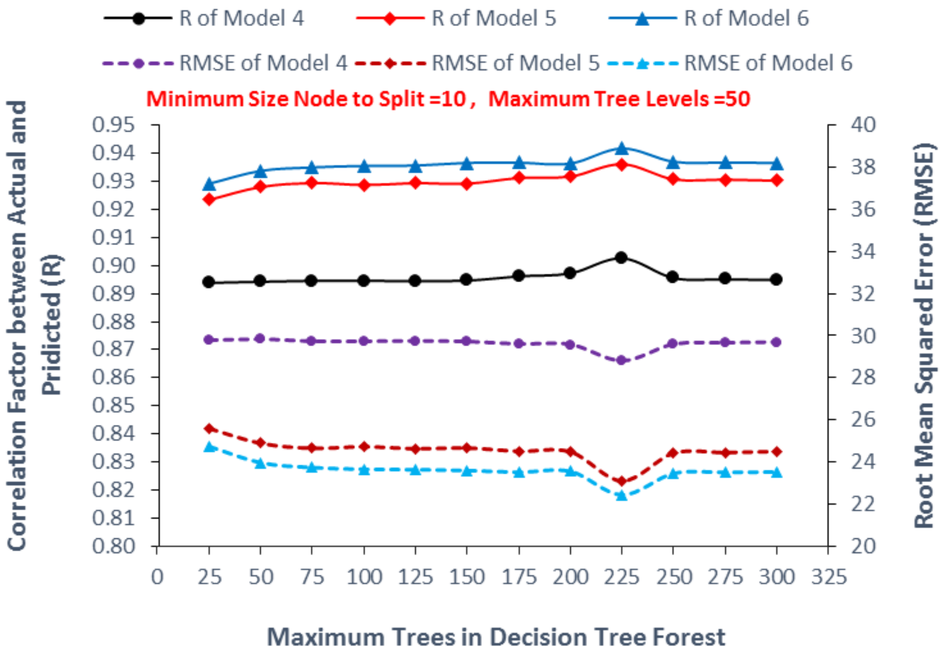

Figure-9. Effect of maximum trees in decision tree forest on the values of R and RMSE for model No. (4, 5, and 6).

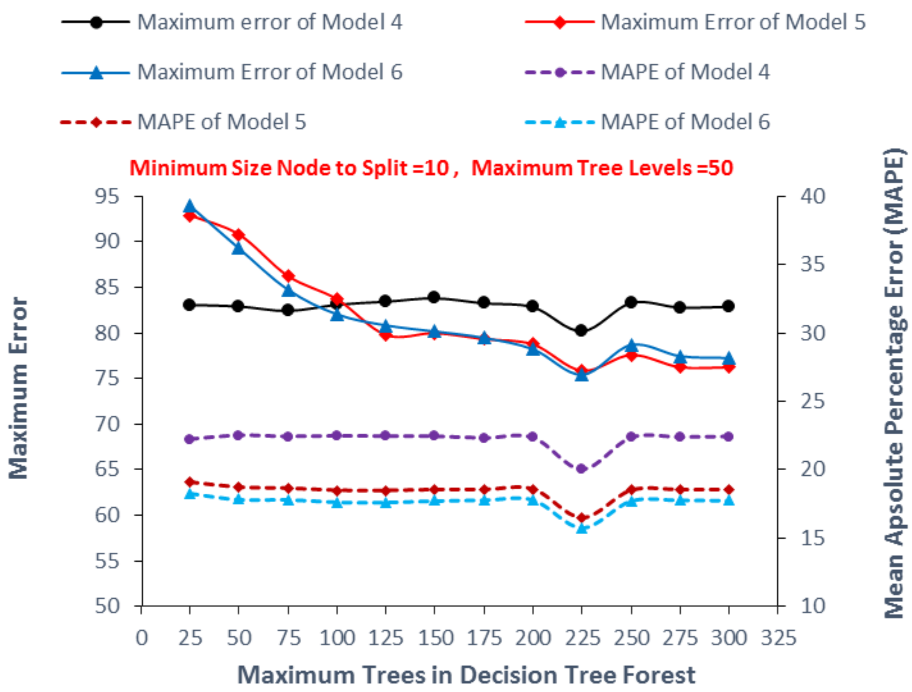

Figure-10. Effect of maximum trees in decision tree forest on the values of ME and MAPE for model No. (4, 5, and 6).

VOL. 13, NO. 12, JUNE 2018 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

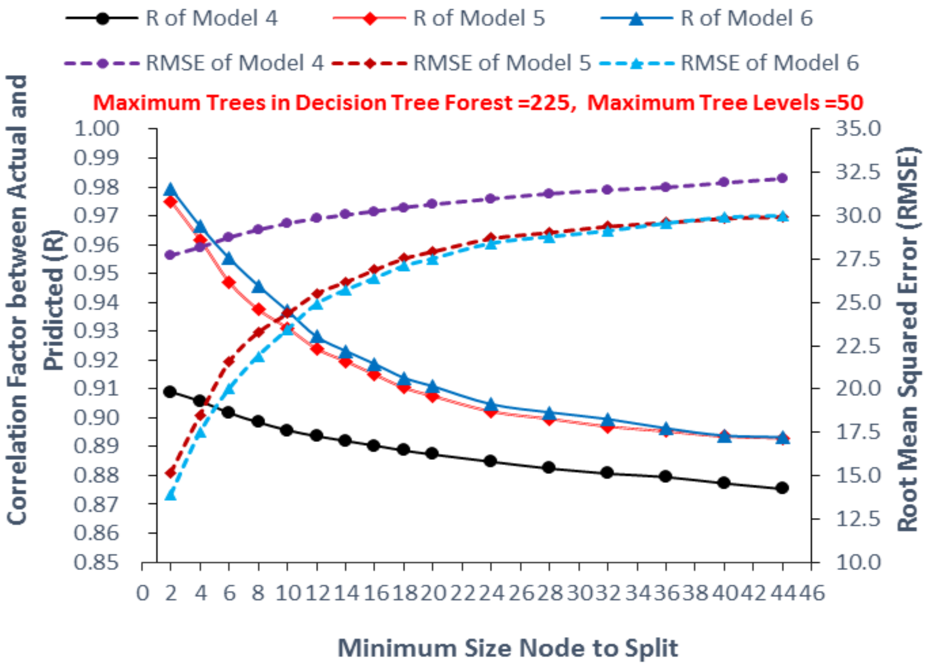

Figure-11. Effect of minimum size node to split on the values of R and RMSE for model No. (4, 5, and 6).

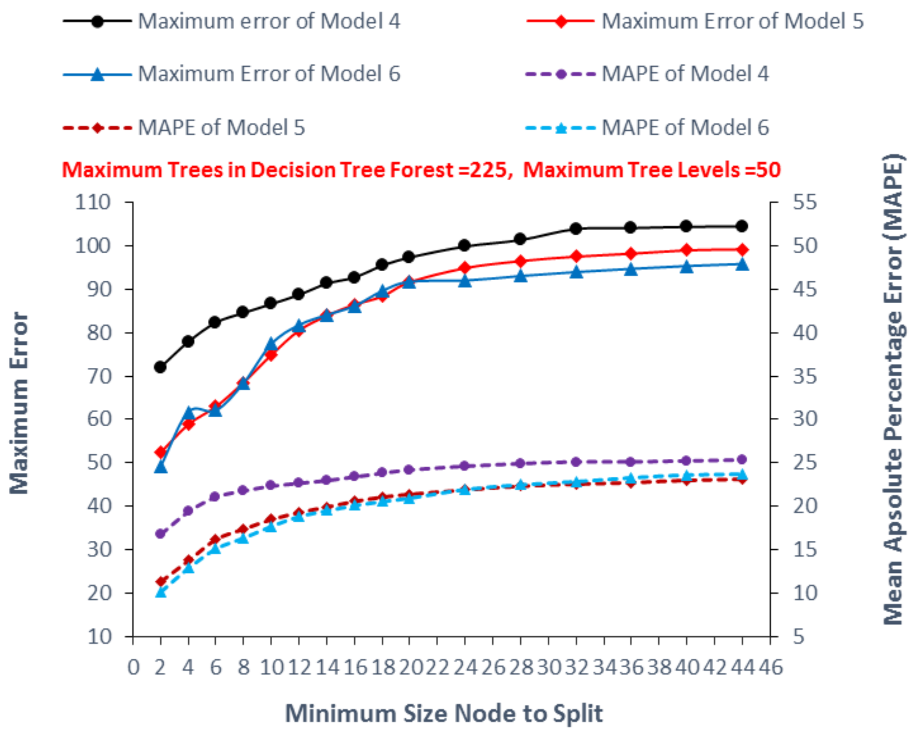

Figure-12. Effect of minimum size node to split on the values of ME and MAPE for model No. (4, 5, and 6).

ARPN Journal of Engineering and Applied Sciences

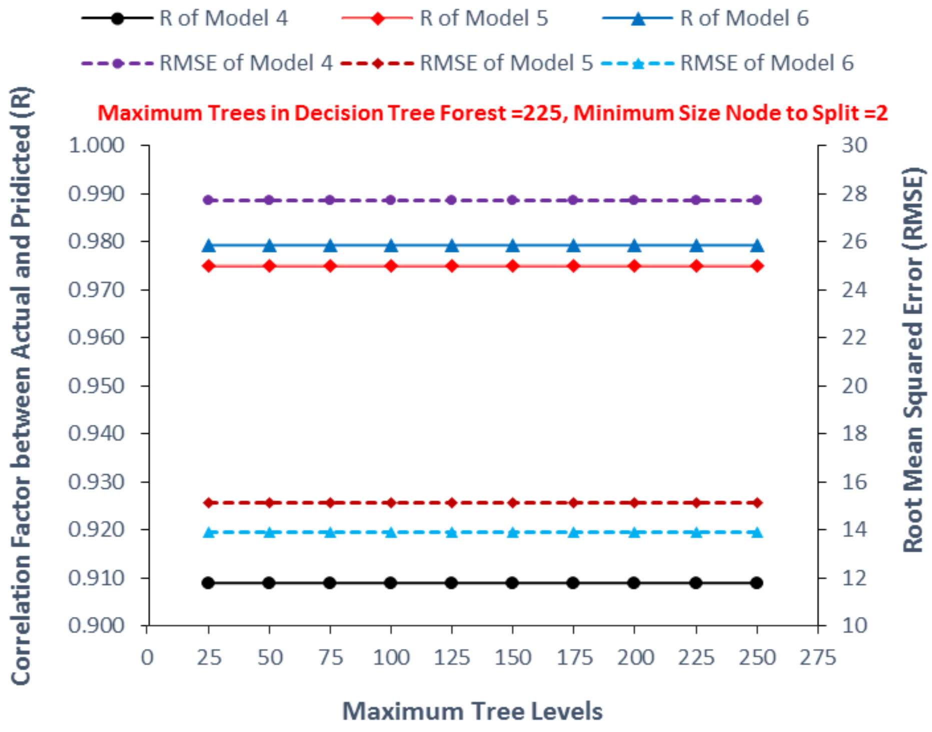

Figure-13. Effect of maximum tree levels on the values of R and RMSE for model No. (4, 5, and 6).

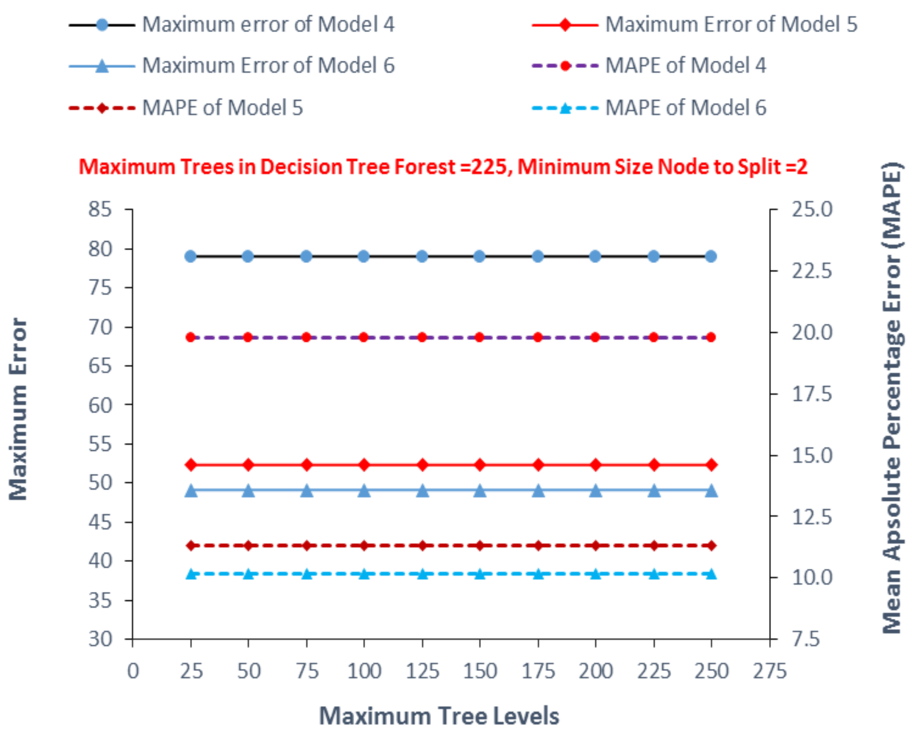

Figure-14. Effect of maximum tree levels on the values of ME and MAPE for model No. (4, 5, and 6).

CONCLUSIONS

Cross validation method is applied in this research, the popular v-fold cross-validation, which provides a good trade-off between model under-fitting and over-fitting has been used to assess the performance of candidate models. The observed datasets (average monthly discharge recorded from 1975 to 2015) were divided randomly into v equal size subgroups. During the modeling process for each DTF model, one of the partitions was used for training phase, while another was used for validation phase. Five-fold cross-validation is applied in this research to evaluate the performance of the DTF models. There are three significant parameters for constructing a DTF model (MTDTF, MSNS, and MTL). Six different DTF models based on lag time and data volume were proposed and their performance compared to determine the best model. In the first stage of this study, the observed data volume involved in DTF models is equal to (240 months); the optimum value of MTDTF is equal to (125). Model No. 3 performed better than the other two models (1&2) with (R, RMSE, ME and MAPE) values of 0.860, 34.148, 103.508 and 26.073, respectively. Increasing the number of trees in the forest does not necessarily lead to improvement of results; there is a gradient in the improvement of the performances up to the number of maximum trees equaled to 125 after which there was a decrease in the performances of the DTF models. The optimum value of MSNS is equal to (32). The modification of parameter (MSNS) from default value (10) to (32) led to significant improvement in the efficiency of DTF models, for example; the percentage ratio of R is increased by 1.5% and the percentage ratios of (RMSE, ME and MAPE) are decreased by (5.0%, 4.25%, and 4.71%) respectively for model No. 3. The effect of MSNS on the evaluation criteria after the value of (36) is continued as semi-straight line. The modification of the third parameter (MTL) does not lead to any improvement in the results, it was observed that the effect of this parameter in the form of a straight line. The second stage of this study is involved data volume (480 months) in the regression. The optimum number of maximum trees in decision tree forest is equaled to (225). The percentage ratio of evaluation criteria (R, RMSE, ME and MAPE) for model No.6 are improved by (7.90%, 30.87%, 23.91%, and 36.73%) respectively compared with performances of model No.3. The optimum value of MSNS in this stage is equal to (2). The size of the data volume has greatly affected on the value of this parameter. There is an inverse correlation between the value of this parameter and the size of the data volume involved in the modeling. Finally, the dramatically change in the value of MTL does not lead to any amelioration in the performances of DTF models.

REFERENCES

[1] Kisi Ozgur and Cigizoglu H. Kerem. 2007. Comparison of different ANN techniques in river flow prediction. Civil Engineering and Environmental Systems. 24(3): 211-231.

[2] Gayathri K Devia, Ganasri B P, Dwarakish G. S. (015. A Review on Hydrological Models, Aquatic Procedia. 4, 1001-1007.

[3] Abbott M.B., Bathrust J.C., Cunge J.A., O'Connell P.E. and Rasmussen J. 1986. An introduction to the European Hydrological System - Systeme Hydrologique Europeen. SHE. 1: History and philosophy of a physically-based, distributed modelling system. 87(1-2): 45-59.

[4] Solomatine DP and Ostfeld. 2008. A Data-driven modelling: some past experiences and new approaches. Journal of Hydroinformatics. 10(1): 3-22.

[5] ASCE. 2000a. Task Committee on Application of Artificial Neural Networks in Hydrology, Artificial Neural Networks in Hydrology, II:Hydrologic Application, J. Hydrol. Eng. 5, 124-136.

[6] ASCE. 2000b. Task Committee on Application of Artificial Neural Networks in Hydrology, Artificial Neural Networks in Hydrology. I: Preliminary Concepts, J. Hydrol. Eng. 5, 115-123.

[7] Chang F. J., Chang L. C. & Huang H. L. 2002. Real-time recurrent neural network for stream-flow forecasting. Hydrol. Processes. 16, 2577-2588.

[8] Sivakumar B., Jayawardena A. W. & Fernando T. M. K. G. 2002. River flow forecasting use of phase-space reconstruction and artificial neural networks approaches. J. Hydrol. 265, 225-245.

[9] Panagoulia S. 2006. Artificial neural networks and high and low flows in various climate regimes. J. Hydrol. Sci. 51(4): 563-587.

[10] Agarwal A., Mishra S.K., Ram S. & Singh J.K. 2006. Simulation of runoff and sediment yield using artificial neural networks. Biosystems Eng. 94 (4): 597-613.

[11] Rabindra K. Panda., Niranjan Pramanik & Biplab Bala. 2010. Simulation of river stage using artificial neural network and MIKE 11 hydrodynamic model. Journal of Computers & Geosciences. 36(6): 735-745.

[12] Chen D., Lu J. & Shen Y. 2010. Artificial neural network modeling of concentrations of nitrogen, phosphorus and dissolved oxygen in a non-point source polluted river in Zhejiang Province, southeast China. Hydrol. Process. 24, 290-299.

[13] Emiroglu ME., Bilhan O. & Kisi O. 2011. Neural networks for estimation of discharge capacity of triangular labyrinth side-weir located on a straight channel. International Journal of Expert Systems. 38(1): 867-874.

[14] Fei Z., Luo D., Li B. 2012. Simulation and Prediction for Groundwater Dynamics Based on RBF Neural Network, Journal of Water Resource and Protection. 4, 540-544.

[15] Al-Aboodi A. H. 2014. Prediction of Tigris river discharge in Baghdad city using Artificial Neural networks..Kufa journal of Engineering (K.J.E), Iraq. 5(2): 107-116.

[16] Al-Aboodi A. H., Dakheel A. A. Ibrahim, H. T. 2017a. Comparison of Data-Driven Modelling Techniques for Predicting River Flow in an Arid Region, International Journal of Applied Engineering Research. 12(11): 2647-2655.

[17] Al-Aboodi A. H., Al-Abadi A. M., Ibrahim H. T. 2017b. A Committee Machine with Intelligent Systems for Estimating Monthly Mean Reference Evapotranspiration in an Arid Region. Research Journal of Applied Sciences, Engineering and Technology. 14(10): 386-398.

[18] Hundecha Y., Bardossy A. & Theisen H.-W. 2001. Development of a fuzzy logic based rainfall–runoff model. Hydrol. Sci. J. 46(3): 363-377.

[19] Xiong L. H., Shamseldin A. Y. & O'Connor K. M. 2001. A nonlinear combination of the forecasts of rainfall–runoff models by the first order Takagi-Sugeno fuzzy system.. Jour. Hydrol. 245, 196-217.

[20] Nayak P.C., Sudheer K.P., Ramasastri K.S. 2004. Fuzzy computing based rainfall-runoff model for real time flood forecasting. Hydrological Process. 17, 3749-3762.

[21] Firat M.; Güngör M. 2007. River flow estimation using adaptive neuro fuzzy inference system.Math. Comput. Simul. 75, 87-96.

[22] Bisht D C S, Jangid A. 2011. Discharge Modelling using Adaptive Neuro - Fuzzy Inference System, International Journal of Advanced Science and Technology. 31, 99-114.

[23] Iraqi Ministries of Environment, Water Resources and Municipalities and Public Works (2006a), p. 91.

[24] JAU Thi-Qar Governorate Profile. 2013. http://www.jauiraq.org/documents/464/GP-Thi-Qar%202013.pdf, 26/03/2015.

[25] Breiman L. 2001. Random Forests. Machine Learning. 45, 5-32.

[26] Friedman J. H. 2002. Stochastic gradient boosting. Computational Statistics & Data Analysis. 38, 367-378.

[27] Phillip H. Sherrod. 2003. DTREG Predictive Modeling Software. www.dtreg.com.

[28] Hipni A., El-shafie A., Najah A., Karim O. A., Hussain A. & Mukhlisin M. 2013. Daily forecasting of dam water levels: comparing a support vector machine (SVM) model with adaptive neuro fuzzy inference system (ANFIS). Water Resour. Manage. 27(10): 3803-3823.

[29] El-Shafie A. & Noureldin A. 2010. Generalized versus nongeneralized neural network model for multi-lead inflow forecasting at Aswan High Dam. Hydrol. Earth Syst. Sci. Discuss. 7(5): 7957-7993.