

Estimation of reference evapotranspiration from climatic data

Abstract

This study investigated the capability of M5 Model Tree (M5MT) to predict reference evapotranspiration (ET_0). M5MT was trained and tested with climatic data from eight weather stations located in coastal areas of Iran for the years 2000-2008. It was validated with climatic data from seven California Irrigation Management Information System (CIMIS) weather stations for the year 2015. Four different data combinations were utilized to train, test, and validate the M5MT model. These were: daily mean air temperature, wind speed, relative humidity, and solar radiation (configuration 1); daily mean air temperature and solar radiation (configuration 2); daily mean air temperature and relative humidity (configuration 3); and daily maximum, minimum, and mean air temperature, and extraterrestrial radiation (configuration 4). The Penman-Monteith (PM) equation was used as a standard method to provide target ET_0 values. Mean absolute error (MAE), root mean square error (RMSE), and the coefficient of determination (R_2) were used to evaluate the performance of the M5MT models developed with different input configurations. Results indicated that M5MT was able to successfully estimate ET_0 . Configuration 1 provided the most accurate results. Configuration 2 showed to have the variables that have a greater influence on ET_0 than configuration 3. Configuration 4 performed the worst. MAE of ET_0 estimates from M5MT₁ was respectively 29%, 55%, and 91% lower than that of M5MT₂, M5MT₃, and M5MT₄, when the model is validated in California. Also, RMSE from M5MT₁ was 29%, 59%, and 125% smaller than that of M5MT₂, M5MT₃, and M5MT₄, respectively.

Volume 1 Issue 1 - 2017

Margaret Lum,¹ Sayed M Bateni,¹ Jalal Shiri,²
Ali Keshavarzi,³¹Department of Civil and Environmental Engineering and Water

Resources Research Center, University of Hawaii at Manoa, USA

²Department of Water Engineering, University of Tabriz, Iran³Department of Soil Science, University of Tehran, Iran

Correspondence: Sayed M Bateni, Department of Civil and Environmental Engineering and Water Resources Research Center, University of Hawaii at Manoa, Hawaii, USA, Tel 808-956-4249, Fax 808-956-5014, Email smbateni@hawaii.edu

Received: May 19, 2017 | **Published:** July 27, 2017

Introduction

Evapotranspiration (ET) is an important component of the hydrologic cycle, which significantly influences crop water requirement and water resource management.¹ Accurate ET estimates enable the proper determination of water budgeting and allocation, and thus improves water use efficiency of irrigation systems. In situ methods are often used to measure ET in a controlled crop area, but they are costly, labor intensive, and only provide localized estimates.^{2,3} To avoid the high costs, empirical, artificial intelligence, and physical models have been developed to estimate reference ET (ET_0).⁴ ET_0 is the combined process of evaporation and transpiration from a theoretical grass surface with an assumed height of 0.12 meter, a surface resistance of 70 s/m, and a surface albedo of 0.23.⁵ The Penman-Monteith (PM) equation has been accepted as a standard approach to estimate ET_0 . However, this method requires many climatic variables that are typically unavailable.⁶ Due to the drawbacks of in situ methods and the PM equation, Artificial Intelligence (AI)-based approaches have been used to estimate ET_0 .⁷ utilized Artificial Neural Network (ANN) to approximate ET_0 in the arid, semi-arid, and sub-humid regions of Inner Mongolia. In comparison with Multiple Linear Regressions (MLR), ANN showed more accurate estimates.⁴ estimated daily ET_0 in Northern Spain using Gene Expression Programming (GEP) and compared its performance with those of the Adaptive Neuro-Fuzzy Inference System (ANFIS), Hargreaves-Semani, and Priestley-Taylor models. Results indicated that GEP provided the most accurate estimates followed by ANFIS.⁸ used ANN to predict ET_0 in arid and semi-arid areas of northwest China. ANN was found to estimate ET_0 more accurately than MLRs, Priestley-Taylor, Hargreaves-Semani, and Penman-Monteith (PM) equations.⁹ predicted ET_0 in northern, mid,

and southern part of Iraq using Extreme Learning Machines (ELM). Compared to the PM equation and Feed Forward Back Propagation (FFBP) models, ELM estimated ET_0 better. Recently, M5 Model Tree (M5MT) has been used in many engineering problems, and showed promising results.¹⁰⁻¹² M5MT is an extension of a regression tree and provides the user with multiple linear functions.¹⁻¹³ This approach is capable of handling high dimensional datasets and the resulting model tree is significantly smaller and more precise than regression trees.¹⁴ Moreover, the M5MT is not a black-box and provides a relationship between the independent and dependent variables.¹¹ Several studies have shown M5MT to be an effective technique to provide accurate results.¹⁵ showed M5MT is advantageous over ANN because it generated more accurate wave height estimates.¹¹ found that the performance of M5MT was comparable to ANN, but indicated that the training process of M5MT was faster than that of ANN.¹⁴ performed a comparison of M5MT and Support Vector Machines (SVM) in forecasting daily river flow. Results showed M5MT performed similar to SVM, but it is computationally less expensive.¹⁶ concluded M5MT to be better than ANN as it provided a more straightforward structure consisting of linear regression equations. The objective of this study is to estimate ET_0 from climatic data using M5MT. Four different combinations of climatic data were used in M5MT. These combinations were daily mean air temperature, wind speed, relative humidity, and solar radiation (configuration 1); daily mean air temperature and solar radiation (configuration 2); daily mean air temperature and relative humidity (configuration 3); and daily maximum, minimum, and mean air temperature, and extraterrestrial radiation (configuration 4). An assessment of which data combination has the most amount of information about ET_0 was made.

Data, methods and models

Studied sites and data: Daily climatic data as well as ET_0 estimates from the PM equation were used to train, test, and validate the M5MT model. The training dataset consisted of data from eight coastal weather stations in Iran, collected from 2000 to 2007. The testing dataset contained data from the same eight stations, but for 2008. Performance of the M5MT models was validated with seven California Irrigation Management Information System (CIMIS) weather stations in 2015. CIMIS dataset was used for models validation to evaluate their feasibility in other regions, and examine whether they are applicable in areas that they were not trained in. Figure 1 & Figure 2 show the spatial distribution of the utilized weather stations in Iran and California, respectively. The recorded data consisted of daily average relative humidity (RH_{mean}), and wind speed (Ws), daily maximum, minimum and mean air temperature (T_{max} , T_{min} , and T_{mean}), and incoming solar radiation (Rs). Table 1 lists the geographical coordinates of each weather station and the corresponding annual averages of the collected data. ET_0 is the reference evapotranspiration (mm/d), Δ is the slope of saturation vapor pressure function (kPa/°C), R_n is the net radiation (MJ/m²day), R_a is extraterrestrial radiation (mm/d), G is the soil heat flux density (MJ/m²day), γ is the psychrometric constant (kPa/°C), T_{mean} is the mean air temperature (°C), T_{max} is the daily maximum air temperature (°C), T_{min} is the daily minimum air temperature (°C), Ws is the daily mean wind speed at a height of 2 m (m/s), RH is relative humidity (%), e_s is the saturation vapor pressure (kPa), and e_a is the actual vapor pressure (kPa). The commonly used equations for the estimation of ET_0 are presented in Table 2. Based on the proposed equations in Table 2 and the study conducted by,⁴ four input combinations were used to predict ET_0 . The following data configurations were used to train, test, and validate M5MT:

Configuration 1: Ws , RH_{mean} , T_{mean} , and Rs [M5MT₁]

Configuration 2: T_{mean} and Rs [M5MT₂]

Configuration 3: T_{mean} and RH_{mean} [M5MT₃]

Configuration 4: T_{mean} , T_{max} , T_{min} and R_a [M5MT₄]

Three statistical metrics (mean absolute error (MAE), root mean square error (RMSE), and the coefficient of determination (R^2)) were used to compare performance of the four M5MT models (i.e., M5MT₁, M5MT₂, M5MT₃, and M5MT₄). These statistical metrics are given below:

$$MAE = \frac{\sum_{i=1}^n |O_i - P_i|}{n} \quad (1)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (O_i - P_i)^2}{n}} \quad (2)$$

$$R^2 = \left[\frac{\sum_{i=1}^n (O_i - \bar{O})(P_i - \bar{P})}{\sqrt{\sum_{i=1}^n (O_i - \bar{O})^2} \sqrt{\sum_{i=1}^n (P_i - \bar{P})^2}} \right]^2 \quad (3)$$

Where n is the number of data points, O_i and P_i are the i th estimated ET_0 values respectively from the PM and M5MT models, and \bar{O} and

\bar{P} are the mean predicted ET_0 values from the respective models. The R^2 signifies the percentage of data that conforms to the regression line at a 45-degree angle. If all points coincide with the regression line, the variation between the variables can be explained by a linear relationship and R^2 would result in an optimal value of one. MAE describes the average of a set of absolute errors with an optimal value of zero. Since only the magnitude of the error is considered, non-negative values are obtained with no upper bound. RMSE is a measure of difference between the observed and simulated values. The greater concentration of data around the 1:1 line, the lower the value of RMSE becomes. RMSE does not have an upper bound and its optimal value is zero.



Figure 1 Location of coastal weather stations in Iran.



Figure 2 Location of CIMIS weather stations in California.

M5 Model Tree (M5MT): M5MT is an improvement of a regression tree, which replaces specific numerical values with linear regression functions relating input variables to corresponding output variables.^{12,13} Two different stages are involved to generate a final model tree. The first stage divides the input space into different regions that correspond to nodes within a tree-like structure. The standard deviation of each region is calculated and corresponds to the amount of error for each of the nodes created. Next, the expected error reduction is calculated for every value propagating to a specific node. The calculated error uses the following formula known as the standard deviation reduction (SDR):¹

Table 1 Geographical location of stations and annual averages of climatic data

Country	Station	Location			Climatic Parameters					
		altitude (m)	Latitude (°)	Longitude (°)	Tmax (°C)	Tmin (°C)	Tmean (°C)	Rs (MJ/m2d)	Ws (m/s)	Rhmean (%)
Iran	Abadan	6.6	30.2	48.2	34	18.9	26.3	19.4	3.2	64.7
	Ahwaz	22.5	31.2	48.4	34	19.4	26.6	18.5	2.4	65.6
	Bandar-e-Abbas	9.8	27.1	56.2	32	23.4	27.5	17.4	3.7	78.6
	Bandar-e-Lenge	22.7	26.3	54.2	33	21.9	27.2	19.4	3.7	73.2
	Bushehr	9	28.6	50.5	30	20.6	25.3	18.3	3.5	75.8
	Gorgan	13.3	36.5	54.1	23	12.9	18.2	15.8	2.6	63.6
	Rasht	-8.6	37.1	49.4	21	12.3	16.6	15.8	1.6	76.5
	Sari	23	36.3	53	23	13.5	18	16.3	2.2	75.6
California, USA	Atascadero	269.8	35.5	-121	24	5.8	14.1	16.5	1.2	64.2
	Delano	91.4	35.8	-119	27	9.8	17.6	18.6	1.4	55.6
	Gilroy	56.4	37	-122	24	7.5	14.8	17.4	2.2	67.2
	Arleta	298.7	34.3	-118	26	11.6	18.3	18.6	1.6	49.7
	Gerber South	75	40	-122	25	9.9	17.2	18.3	2.3	59.8
	Woodland	25	38.7	-122	25	9.8	17.1	17.5	2.2	54.8
	Diamond Springs	624.8	38.6	-121	22	10.5	16.1	17.7	1.7	49.9

Table 2 Different ET₀ equations

Study	ET ₀ Equation
Penman-Monteith et al. [2]	$ET_0 = \frac{0.408\Delta(R_n - G) + \gamma \frac{900}{T_{mean} + 273} W_s (e_s - e_a)}{\Delta + \gamma(1 + 0.34W_s)}$
Makkink [20]	$ET_0 = 0.61 \frac{\Delta R_s}{(\Delta + \gamma)\lambda} - 0.12$
Romanenko [21]	$ET_0 = 0.0018(T_{mean} + 25)^2 (100 - RH)$
Hargreaves & Samani [22]	$ET_0 = 0.0023 \frac{R_a}{\lambda} (T_{mean} + 17.8) \sqrt{T_{max} - T_{min}}$

$$SDR = sd(T) - \sum_{i=1}^A \frac{|T_i|}{|T|} sd(T_i) \tag{4}$$

Where T is the set of values that reach a node, Ti is the subset of values that have the ith outcome of a potential set, A is the final amount of values in set T, and sd is the standard deviation. This dividing process results in the algorithm to perform iterations, which generates subsequent nodes that will exhibit a reduction in standard deviation from the previous nodes. The algorithm will continue to iterate,

considering all possible splits, and ends when the least expected error is attained.¹⁷The conclusion of the first stage leaves the model tree to have large structure, which initiates pruning of the overgrown model tree (i.e., the second stage of M5MT).¹² Pruning will occur if the estimated error of nodes branched below a specific node is greater.¹⁻¹⁸ Linear regression equations will replace the pruned nodes, resulting in a more simplified and accurate model tree.¹⁵⁻¹⁹

Results and Discussion

Building M5 Model Tree: This study used WEKA (Waikato Environment for Knowledge Analysis), which is a data mining

software to estimate ET_0 . It consists of a wide variety of machine learning algorithms including M5MT. The WEKA interface provided different testing options (i.e., percentage split, train-test, and cross validation) to assist in the modeling process. Among the three aforementioned options, the Train-Test method was selected because of its better performance (Table 3).

Table 3 Performance of M5MT for different testing options

Methods	MAE (mm/d)	RMSE (mm/d)	R2
Percentage Split	0.255	0.3422	0.9904
Train-Test	0.2328	0.3189	0.9914
Cross Validation	0.2396	0.3337	0.9906

Performance of M5MT models in Iran (training and testing stages): Table 4 shows MAE, RMSE, and R_2 of ET_0 estimates from the four M5MT models for training and testing stages. The training process resulted in MAE, RMSE, and R_2 values ranging between 0.33–0.76mm/d, 0.47–1.03mm/d, and 0.81–0.96, respectively. The testing stage showed similar values that ranged between 0.38–0.77mm/d (MAE), 0.55–1.06mm/d (RMSE), and 0.80–0.95 (R_2). It was observed that the presence or absence of certain input parameters influenced the performance of the models. Comparing the models with two input variables, M5MT₃ (whose inputs were Tmean and RHmean) had higher MAE (0.76mm/d) and RMSE (1.0mm/d) than M5MT₂ (whose inputs were Tmean and R_s) in the training stage.

Table 4 Statistical metrics of M5MT models for training and testing stages

	Training: Iran (2000-2007)				Testing: Iran (2008)			
	M5MT1	M5MT2	M5MT3	M5MT4	M5MT1	M5MT2	M5MT3	M5MT4
MAE (mm/d)	0.33	0.62	0.76	0.53	0.38	0.64	0.77	0.56
RMSE (mm/d)	0.47	0.85	1.03	0.73	0.55	0.92	1.06	0.82
R2	0.96	0.87	0.81	0.91	0.95	0.85	0.8	0.88

Solar radiation tends to have a greater effect on ET_0 , as replacing mean relative humidity by solar radiation increased accuracy in the training phase and decreased MAE and RMSE by 23% and 21%, respectively. This is in agreement with the results from the testing stage, in which a respective 20% and 15% decrease in MAE and RMSE was observed when solar radiation was used in lieu of mean relative humidity. Assessing the performance of the M5MT models with four input variables during the training stage, M5MT₄ (whose inputs were T_{mean} , T_{max} , T_{min} and R_s) had larger MAE (0.53mm/d), RMSE (0.73mm/d), lower R_2 (0.91) values than M5MT₁ (whose inputs were W_2 , RH_{mean} , T_{mean} and R_s). This is consistent with the testing phase because MAE and RMSE are decreased by 47% and 49% by using M5MT₁ instead of M5MT₄. Figure 3A & Figure 3B show estimated ET_0 values from the four M5MT models versus PM ET_0 estimates for training and testing stages, respectively. The concentration of data around the 1:1 line in Figure 3A & Figure 3B reflects the low RMSE values obtained in both training and testing stages. In general, the small MAE and RMSE and high R_2 suggest that M5MT can accurately estimate ET_0 . Overall, the results in Figure 3A & Figure 3B and Table 4 indicate that the combination M5MT₁ provided the most accurate ET_0 estimates during the training and testing stages. In the training stage, MAE (RMSE) of ET_0 estimates from M5MT₁ were respectively 88% (81%), 130% (119%) and 61% (55%) lower than those of M5MT₂, M5MT₃ and M5MT₄. A similar tendency was seen in the testing stage with MAE (RMSE) values of M5MT₁ were 68% (67%), 103% (93%), and 47% (49%) lower than those of M5MT₂, M5MT₃ and M5MT₄, respectively.

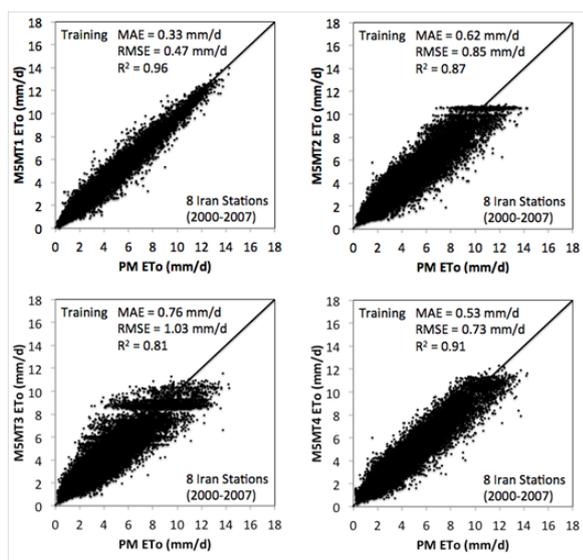


Figure 3 (A) Estimated ET_0 values from different M5MT models versus PM ET_0 estimates for the training step.

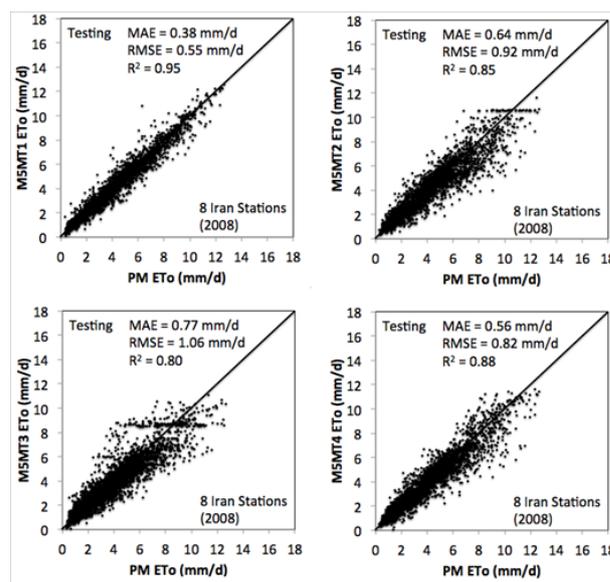


Figure 3 (B) The same as Figure 3b, but for testing stage.

Performance of M5MT models in California (validation stage):

To validate the robustness of the M5MT models, they were applied to seven CIMIS weather stations in California. It should be noted that the CIMIS data was not used to train the M5MT models. The statistical metrics of the M5MT models were given in Table 5. MAE, RMSE, and R_2 values ranged between 0.65-1.24mm/d, 0.80-1.80mm/d, and 0.68-0.90, respectively Table 5. Figure 4 illustrates plots of ET_0 estimates from the four M5MT models versus PM ET_0 estimates at the seven CIMIS stations. Performance of the models can be ranked as follows: M5MT₁, M5MT₂, M5MT₃, and M5MT₄. Similar to the training and testing phases, M5MT₁ outperformed the other models when tested at the California stations. Although no CIMIS data was used to train the M5MT₁ model, it performed well when applied to the stations in California. This implies that the M5MT₁ can provide accurate results in other regions. Figure 4 & Figure 5 showed that the M5MT₁ model with MAE, RMSE, and R_2 values of 0.65mm/d, 0.80mm/d, and 0.90, respectively, can be selected as the best M5MT model for ET_0 estimation. With only two input parameters (i.e., M5MT₂ and M5MT₃), M5MT₂ showed to provide better results than M5MT₃. This implies that a combination of R_s and T_{mean} contributes more significantly towards the estimation of ET_0 than a combination of RH_{mean} and T_{mean} . Figure 5 indicates time series of ET_0 estimates from M5MT₁ and PM models at four CIMIS stations (i.e., Delano, Gerber South, Woodland, and Diamond Springs). As shown, the estimated ET_0 values from M5MT₁ agree well with those of the PM equation. Remarkably, ET_0 estimates from M5MT₁ captured the fluctuations of the PM ET_0 values. Compared to M5MT₁, MAE values of M5MT₂, M5MT₃, and M5MT₄ were respectively 29%, 55%, and 91% larger. Also, RMSE values from M5MT₂, M5MT₃, and M5MT₄ were respectively 29%, 59%, and 125% greater than that of M5MT₁.

Table 5 Statistical metrics of M5MT models for validation stage

Validation: california (2015)				
	M5MT1	M5MT2	M5MT3	M5MT4
MAE (mm/d)	0.65	0.84	1.01	1.24
RMSE (mm/d)	0.8	1.03	1.27	1.8
R2	0.9	0.93	0.74	0.68

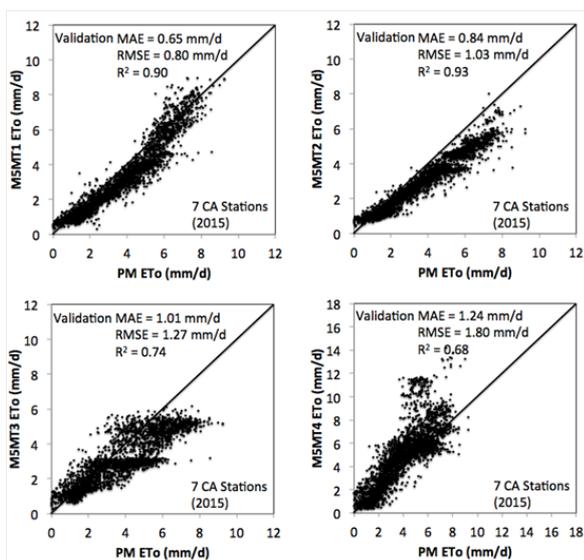


Figure 4 The same as Figure 3a, but for validation stage.

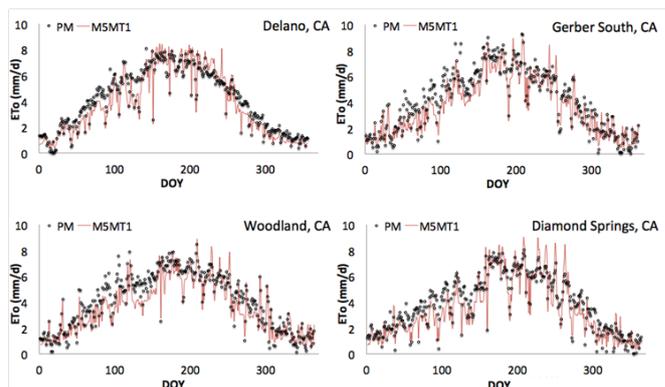


Figure 5 Time series of ET_0 estimates from M5MT1 and PM at four CIMIS stations for 2015.

Conclusion

This study examined the ability of M5 Model Tree (M5MT) to estimate reference evapotranspiration (ET_0) from climatic data. Four combinations of data were used in the M5MT model. These combinations were daily mean air temperature, wind speed, relative humidity, and solar radiation (configuration 1); daily mean air temperature and solar radiation (configuration 2); daily mean air temperature and relative humidity (configuration 3); and daily maximum, minimum, and mean air temperature, and extraterrestrial radiation (configuration 4). The objective was to determine which data combination had the most significant amount of information on ET_0 . The results indicated that M5MT can estimate ET_0 accurately. Data combination 1 generated the most accurate ET_0 estimates and thus consisted of variables that have the most amount of information on ET_0 compared to data combinations 2, 3, and 4. Comparing M5MT models with two input variables, configuration 2 resulted in more accurate ET_0 estimates than configuration 3. This suggests the greater importance of solar radiation and air temperature in comparison to relative humidity and air temperature to estimate ET_0 .

Acknowledgement

None.

Conflict of interest

None.

References

1. Pal M, Deswal S. M5 model tree based modeling of reference evapotranspiration. *Hydrological processes*. 2009;23(10):1437–1443.
2. Gavilan P, Berengena J, Allen RG. Measuring versus estimating net radiation and soil heat flux: impact on Penman–Monteith reference ET estimates in semiarid regions. *Agricultural Water Management*. 2007;89(3):275–286.
3. Verstraeten W, Veroustraete F, Feyen J. Assessment of Evapotranspiration and Soil Moisture Content Across Different Scales of Observation. *Sensors*. 2008;8(1):70–117.
4. Shiri J, Kisi O, Landaras G, et al. Daily reference evapotranspiration modeling by using genetic programming approach in the Basque Country (Northern Spain). *Journal of Hydrology*. 2012;414(2012):302–316.
5. Allen R, Pereira LS, Raes D, et al. Crop evapotranspiration-Guidelines for computing crop evapotranspiration-FAO Irrigation and Drainage Paper 56. *FAO - Food and Agriculture Organization of the United Nations Rome*. 1998;1998: 1–15.

6. Yassin M, Alazba AA, Mattar M. Artificial neural networks versus gene expression programming for estimating reference evapotranspiration in arid climate. *Agricultural Water Management*. 2012;163:110–124.
7. Dai X, Shi H, Li Y, et al. Artificial neural network models for estimating regional reference evapotranspiration based on climate factors. *Hydrological Processes*. 2009;23(2):442–450.
8. Huo Z, Feng S, Kang S, et al. Artificial neural network models for reference evapotranspiration in an arid area of northwest China. *Journal of Arid Environments*. 2012;82:81–90.
9. Abdullah SS, Malek MA, Abdullah NS, et al. Extreme Learning Machines: A new approach for prediction of reference evapotranspiration. *Journal of Hydrology*. 2015;527(2015):184–195.
10. Solomatine DP, Dulal KN. Model trees as an alternative to neural networks in rainfall-runoff modeling. *Hydrological Sciences Journal*. 2003;48(3):399–411.
11. Solomatine DP, Xue Y. M5 model trees compared to neural networks: application to flood forecasting in the upper reach of the Huai River in China. *Journal of Hydrologic Engineering*. 2005;9(6):491–501.
12. Bhattacharya B, Solomatine DP. Machine learning in sedimentation modeling. *Neural Network*. 2005;19(2):208–214.
13. Witten IH, Frank E. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, California, USA; 2005. p. 1–558.
14. Sattari MT, Apaydin H, Ozturk F, et al. M5 Model Tree Application in Daily River Flow Forecasting in Sohu Stream, Turkey. *Water Resources*. 2013;40(3):233–242.
15. Etemad Shahidi A, Mahjoobi J. Comparison between M5 model tree and neural networks for prediction of significant wave height in Lake Superior. *Ocean Engineering*. 2009;36(2009):1175–1181.
16. Alipour A, Yarahmad J, Mahdavi M. Comparative Study of M5 Model Tree and Artificial Neural Network in Estimating Reference Evapotranspiration Using MODIS Products. *Journal of Climatology*. 2014;2014(2014):1–11.
17. Rahimikhoob A, Asadi M, Mashal M. A Comparison Between Conventional and M5 Model Tree Methods for Converting Pan Evaporation to Reference Evapotranspiration for Semi-Arid Region. *Water Resource Management*. 2013;27(14):4815–4826.
18. Atiaa AM, Ghalib HB. Rainfall-Runoff modeling by using M5 model trees technique: an example of Tigris catchment area in Baghdad, Middle of Iraq. *Marsh Bulletin*. 2008;3(2):125–135.
19. Wang Y, Witten IH. Induction of model trees for predicting continuous classes. *Proceedings of the Ninth European Conference on Machine Learning*. 1997;1–10.
20. Makkink GF. Testing the Penman formula by means of lysimeters. *Journal of the Institution of Water Engineering*. 1957;11(3):277–288.
21. Romanenko VA. Computation of the autumn soil moisture using a universal relationship for a large area. *Proceedings of Ukrainian Hydrometeorological Research Institute*. 1961;3:12–25.
22. Hargreaves GH, Samani ZA. Reference crop evapotranspiration from temperature. *Applied Engineering in Agriculture*. 1985;1(2):96–99.