Journal of Soft Computing in Civil Engineering 6-1 (2022) 01-28

Contents lists available at SCCE

Journal of Soft Computing in Civil Engineering

Journal homepage: www.jsoftcivil.com

Modeling of Reference Crop Evapotranspiration in Wet and Dry Climates Using Data-Mining Methods and Empirical Equations

Mohammad Sadegh Zakeri[1], Sayed-Farhad Mousavi[2], Saeed Farzin[3*], Hadi Sanikhani[4]

1. Graduated MSc., Department of Water Engineering and Hydraulic Structures, Faculty of Civil Engineering, Semnan University, Semnan, Iran
2. Professor, Department of Water Engineering and Hydraulic Structures, Faculty of Civil Engineering, Semnan University, Semnan, Iran
3. Associate Professor, Department of Water Engineering and Hydraulic Structures, Faculty of Civil Engineering, Semnan University, Semnan, Iran
4. Assistant Professor, Department of Water Engineering, Faculty of Agriculture, Kurdistan University, Sanandaj, Iran

Corresponding author: *saeed.farzin@semnan.ac.ir*

https://doi.org/10.22115/SCCE.2022.298173.1347

ARTICLE INFO

Article history:
Received: 03 August 2021
Revised: 23 December 2021
Accepted: 01 January 2022

Keywords:
Climate;
Reference crop evapotranspiration;
Data-mining methods;
Uncertainty.

ABSTRACT

In the present study, performance of data-mining methods in modeling and estimating reference crop evapotranspiration (ET_o) is investigated. To this end, different machine learning, including Artificial Neural Network (ANN), M5 tree, Multivariate Adaptive Regression Splines (MARS), Least Square Support Vector Machine (LS-SVM), and Random Forest (RF) are employed by considering different criteria including impacts of climate (eight synoptic stations in humid and dry climates), accuracy, uncertainty and computation time. Furthermore, to show the application of data-mining methods, their results are compared with some empirical equations, that indicated the superiority of data- mining methods. In the humid climate, it was demonstrated that M5 tree model is the best if only accuracy criterion is considered, and MARS is a better data-mining method by considering accuracy, uncertainty, and computation time criteria. While in the dry climate, the ANN has better results by considering accuracy and all other criteria. In the final step, for a comprehensive investigation of data-mining ability in ET_o modeling, all data in humid and dry climates are combined. Results showed the highest accuracy by MARS and ANN models.

How to cite this article: Zakeri MS, Mousavi SF, Farzin S, Sanikhani H. Modeling of reference crop evapotranspiration in wet and dry climates using data-mining methods and empirical equations. J Soft Comput Civ Eng 2022;6(1):01–28. https://doi.org/10.22115/scce.2022.298173.1347

1. Introduction

Reference crop evapotranspiration (ET_o) is a variable used in irrigation planning, water resources management, and hydrological studies [1]. Evapotranspiration is a nonlinear and complex phenomenon [2]. Hence, it is essential that robust and nonlinear methods should be used for modeling this phenomenon. In this regard, data-mining methods are a good idea for modeling ET_o. Data-mining methods have been used in many studies for solving complex and nonlinear problems. Some of the applications of data-mining methods are river flow modeling [3], reservoir operation [4], minimizing irrigation deficiencies [5], optimization of energy management [6], precipitation modeling [7,8], modeling water quality parameters [9], flood frequency analysis under climate change [10], estimating pier scour depth [11], and modeling seismic retrofit cost estimation [12].

So far, many methods, based on available meteorological parameters in different geographical and climatic conditions, have been proposed to determine the ET_o. Traore et al. [13] estimated the ET_o using an artificial neural network (ANN) in Burkina Faso. Results of the study indicated that ANN is highly capable of evaluating ET_o. Rahimikhoob et al. [14] compared the M5 decision tree model and ANN to estimate ET_o in a dry climate. This study showed that ANN estimated ET_o better than the M5 decision tree model. But M5 and ANN models calculated ET_o with reasonable accuracy, and the results were close to those of FAO 56 Penman-Monteith (PM) equation. Yassin et al. [15] estimated the ET_o using ANN and gene expression programming (GEP) in dry climates. Results showed that the eight ET_o models produced by using the ANN technique were slightly more accurate than those for the GEP technique. Caminha et al. [16] estimated the ET_o using data-mining predictor models and feature selection. Results showed that highly-accurate models could be produced by using the M5 tree algorithm and feature selection technique. Mehdizadeh [17] estimated daily ET_o using artificial intelligence. Local performance of the models showed that MARS and GEP approaches could determine daily ETo using meteorological parameters and residual ET_o data as inputs. However, MARS had the best performance in meteorological-data scenarios. Ferreira et al. [1] modeled daily ET_o with limited climatic data using the MARS algorithm and FAO 56 PM equation. MARS model showed superior performance in all scenarios. Models that used solar radiation had the best performance, followed by those that used relative humidity and wind speed. Ehteram et al. [18] employed a hybrid of support vector regression (SVR) and cuckoo search (CS) algorithm, M5, GEP, and adaptive neuro-fuzzy inference system (ANFIS) for modeling ET_o in India. Results indicated more accuracy of SVR and CS hybrid in modeling ET_o than other investigated algorithms. Wang et al. [19] examined the generalized evapotranspiration models with limited data based on GEP and RF in Guangxi, China. Results showed that RF-based ET_o models performed slightly better than GEP-based models. Fan et al. [20] estimated daily ET_o with local and external meteorological data using M5, RF, lightGBM and empirical equations of Makkink, Tabari,

Hargreaves-Samani, and Trabert in humid areas. Results showed that all three soft computing models produced better daily ET_o estimates than corresponding empirical models using the same input variables. Ferreira and Cunha [21] explored a new method for estimating daily ET_o based on hourly temperature and relative humidity using ANN, RF and CNN models. Results showed that the developed CNN models offer the best performance in all cases. Granata et al. [22] developed some artificial-intelligence-based approaches to estimate actual evapotranspiration in lagoons. Results showed that RF and K nearest neighbors (KNN) models performed better than acute respiratory distress syndrome (ARDS) algorithm and MLP models. Yamaç and Todorovic [23] employed three data-mining methods, including ANN, KNN and AdaBoost for modeling ET_o. Results indicated better accuracy of ANN and KNN. Ashrafzadeh et al. [24] used SARIMA, SVM and GMDH for modeling long term ET_o in northern Iran. Results showed that SARIMA outperformed SVM and GMDH. Zhang et al. [25] employed four different ANN methods for estimating ET_o in Henan province, China. Results indicated that ANN methods can successfully estimate the ET_o in Henan province. Niaghi et al. [26] used four data-mining methods including GEP, MLR, RF and SVR for modeling ET_o. Results showed good accuracy of these methods. Feng and Tian [27] modeled the ET_o by using KNN method. Results showed good precision of this method.

According to the authors' best knowledge, different data-mining methods have been used for modeling reference crop evapotranspiration (ET_o). However, in these studies, the critical issues such as impacts of climate on the performance of data-mining methods, uncertainty, and computation time are not considered. Therefore, in the present study, different data-mining methods including ANN, M5 decision tree, LS-SVM, MARS, and RF are employed for modeling ET_o by considering the impact of climate, uncertainty, computation time and accuracy. In the present study, the uncertainty will be considered by evaluating coefficient of variation of evaluation criteria for each algorithm in several random runs. For considering the impact of climate on the performance of data-mining methods, different meteorological stations in two climates will be considered for modeling ET_o. Finally, the best data-mining method for each climate will be presented based on the accuracy, uncertainty, and computation time.

The rest of the present study is as follows: Section 2 presents the methodology of the present study, including introducing the study area, data used, investigated methods, evaluation criteria, limitation of the present study, and model ranking. Section 3 presents the results of sensitivity analysis and outcomes of empirical equations and data-mining methods. Section 4 offers the discussion about the obtained results. Section 5 presents the conclusion and novelty of the present study. Figure 1 shows the workflow of the present study.

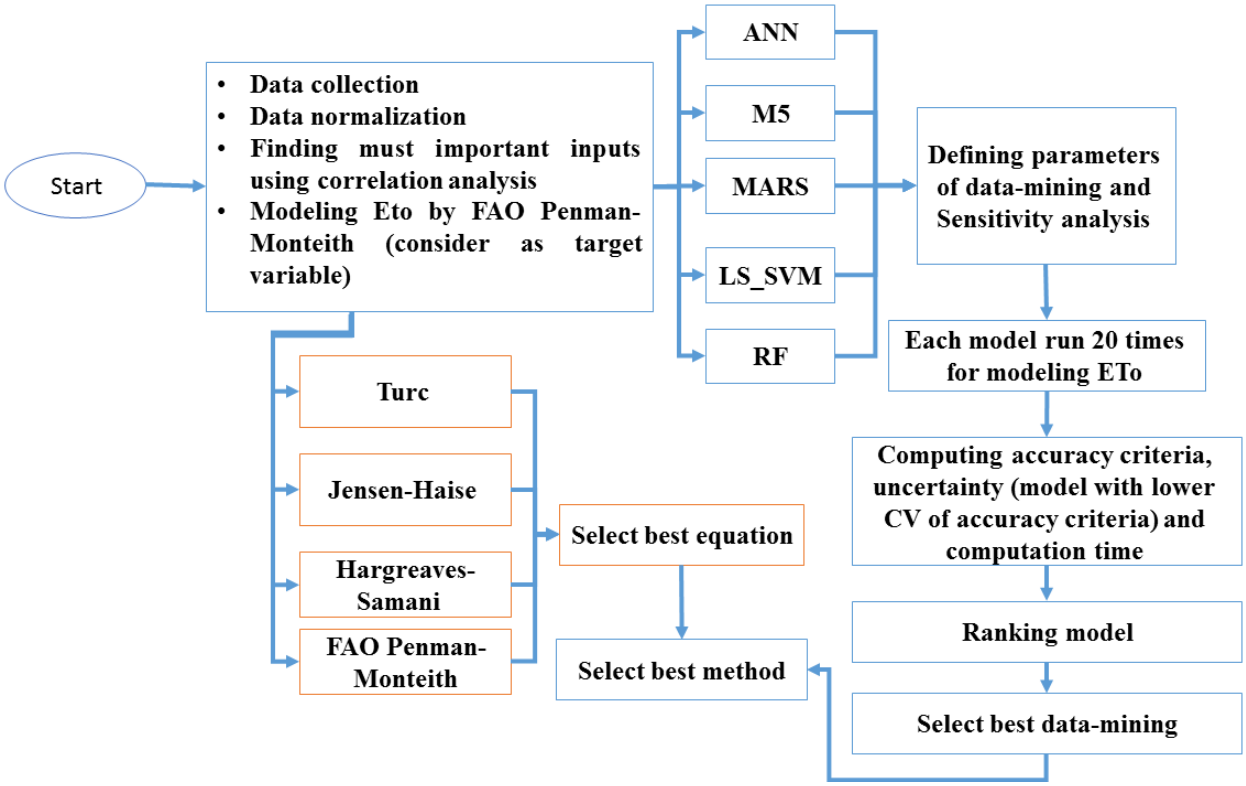

Fig. 1. The workflow of present study for modeling ET_o.

2. Methods

2.1. Study area

In this study, two provinces of Mazandaran and Semnan, in the north of Iran, were considered to calculate the ET_0. The Mazandaran province, with an area of about 24,000 km^2, lies between 35° 46' to 36° 58' north latitude, and 50° 21' to 54° 08' east longitude. The natural conditions of Mazandaran province represent two significant areas of Alborz mountains and coastal plains. Semnan province covers 5.8% of Iran, with an area of 97491 km^2. This province lies between 34° 13' to 37° 20' north latitude and 51° 51' to 57° 3' east longitude. Its border provinces are Mazandaran and Golestan in the north, Isfahan in the south, Khorasan in the east, and Tehran in the west. Fig. 2 shows the geographical location of the two provinces of Mazandaran and Semnan. Table 1 shows the synoptic stations of Mazandaran and Semnan provinces.

Table 1
Synoptic stations of Semnan and Mazandaran provinces.

Station	Altitude (m amsl)	Longitude E	Latitude N	Station	Altitude (m amsl)	Longitude E	Latitude N
Semnan	1130	53.32	35.34	Sari	23	53.00	36.33
Shahrood	1380	54.57	36.25	Dasht-e-naz	12	53.11	36.37
Garmsar	850	52.25	35.20	Ghaemshahr	15	52.46	36.27
Damghan	1170	54.61	35.44	Babolsar	-21	52.39	36.43

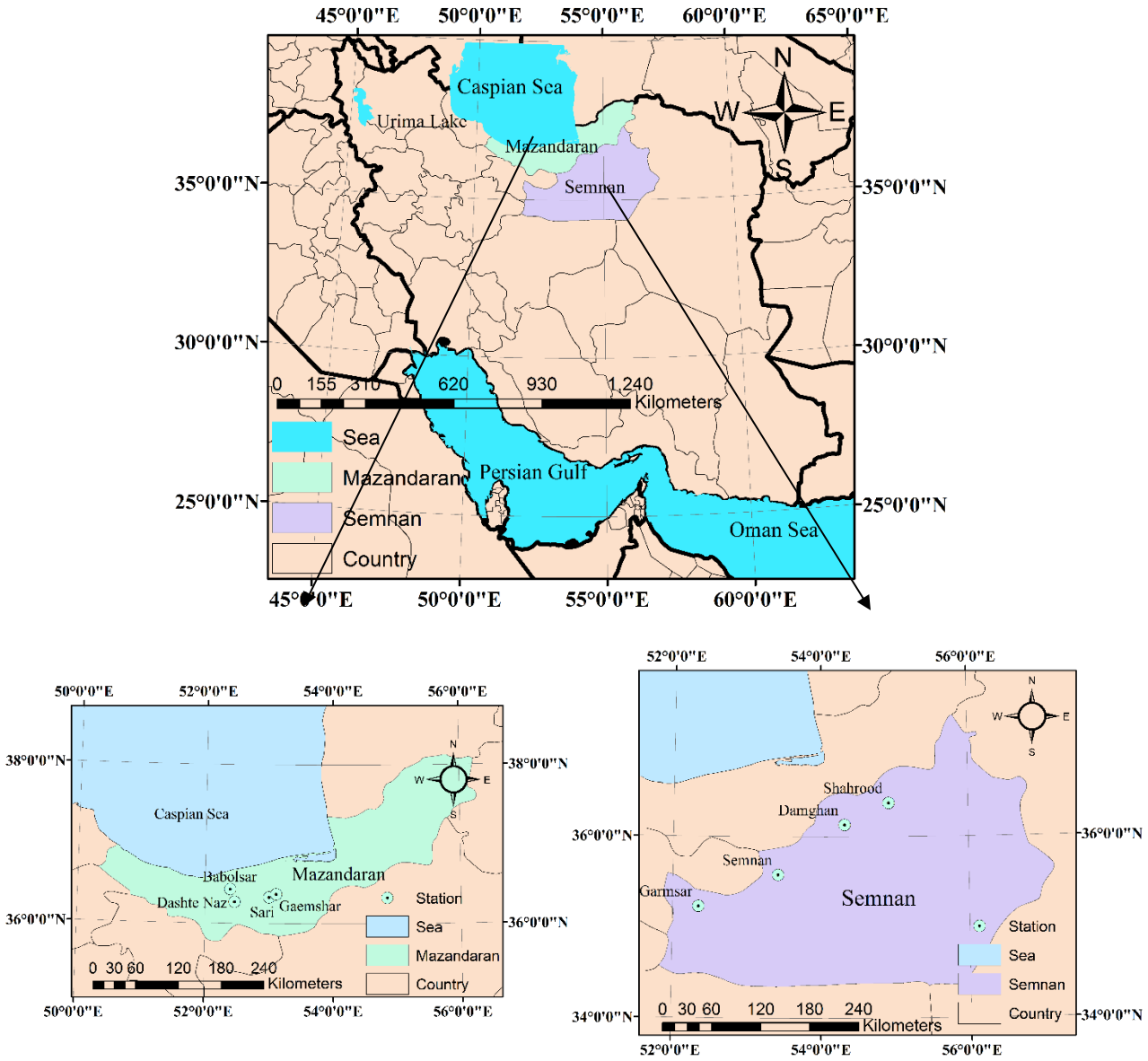

Fig. 2. The geographical location of Mazandaran and Semnan provinces.

2.2. Data used

In the present study, different parameters including minimum absolute temperature (Tmin-abs (°C)), minimum temperature (Tmin (°C)), maximum absolute temperature (Tmax-abs (°C)), maximum temperature (Tmax (°C)), mean temperature (Tmean (°C)), minimum relative humidity (Hmin (%)), maximum relative humidity (Hmax (%)), mean relative humidity (Hmean (%)), wind direction (W-d (deg)), wind speed (W-s (m/s)), and sunshine hours are considered as inputs for modeling and estimating reference crop evapotranspiration (ET_o). The statistical criteria and number of samples of target data in different stations are presented in Table 2. These data are provided by Water Resources Management Company, Tehran, Iran.

Table 2
Statistical criteria of inputs and target data in different stations.

Station	Mean (mm/day)	Min (mm/day)	Max (mm/day)	Std dev (mm/day)	Number of samples
Semnan	3.89	0.85	7.90	2.18	192
Shahrood	3.86	0.77	8.23	2.24	276
Garmsar	3.44	0.76	7.18	1.92	288
Damghan	3.17	0.67	6.89	1.91	276
Sari	2.54	0.71	5.28	1.37	204
Dasht-e-naz	2.65	0.85	5.52	1.39	144
Ghaemshahr	2.48	0.71	4.96	1.31	192
Babolsar	2.48	0.67	5.00	1.28	204

Many researchers have modeled various phenomena using data-mining methods [7,28,29]. In this study, the ET_o is modeled using intelligent and empirical methods. In this regard, 70% of data is considered for training period, and 30% of data is used for the testing period. Also, random calibration method is used for training and testing machine learning algorithms [8, 23, 24].

Data normalization before entering them into a model is one of the essential steps in using data-mining methods. When the range of model changes is high, normalization will significantly help the model to have better and faster training. When the data is normalized, the accuracy and speed of the network increases. The following equation describes how to normalize the data [30]:

$$X_n = 0.1 + 0.8\frac{X_i - X_{\min}}{X_{\max} - X_{\min}} \quad (1)$$

where, X_n is normalized value of X_i input, and X_{max} and X_{min} are maximum and minimum data values, respectively.

2.3. Artificial neural network (ANN) model

The ANN consists of three layers: Input, output, and hidden layers, between the input and output layers. The ANN may be expressed as a network of interconnected neurons [30]. The underlying unit in the ANN is a neuron or node. The nerve cells are connected by synapses, which each synapse has a weight factor. Artificial neural networks are nonlinear models and use a structure that links the inputs and outputs of each system to represent complex nonlinear processes. The structure of each ANN is expressed as (i, j, k), where i represents the number of nodes in the input layer, j represents the number of layers in the hidden layer, and k represents the number of layers in the output layer [31]. The target value in ANN is calculated as follows:

$$\hat{Y}(x) = \sum_{i=1}^{n} \alpha_i f\left(\alpha_i\right), \alpha_i = \sum_{j=1}^{q} w_{ij} x_j + \beta_j \quad (2)$$

where, α_i and w_{ij} are weights of the network, β_j is bias of the network, f is a transfer function, x_j is j^{th} input, n denotes the number of neurons in the hidden layer, and q is number of inputs. The number hidden layer and number of its neuron are considered equal to one and five, respectively that are similar to the study of [32]. For more information about ANN, please see [11].

2.4. M5 decision tree model

The M5 decision tree model was introduced by Quinlan in 1992, based on a binary decision tree that also has linear regression functions on the terminal nodes that form a link between the input and output variables [33] (Fig. 3). The M5 tree model is one of the most common tree models in which the multidimensional parameter space is subdivided into subspaces and substrates, and a linear regression model is created for each subspace in the leaf [34]. This model focuses on quantitative data, which increases the importance of the model compared to other models [35]. Standard deviation selects the best feature for splitting the dataset into each node [36]. The M5 model is obtained by using standard deviation reduction calculated as follows:

$$SDR = sd(E) - \sum_i \frac{|E_i|}{|E|} sd(E_i) \tag{3}$$

In this standard deviation equation, E is a set of samples that reach the node, and Ei is a subset of input data to the parent node. These steps are completed until the proper tree structure is formed. In that way, the tree is pruned in the back step to deal with overfitting [34].

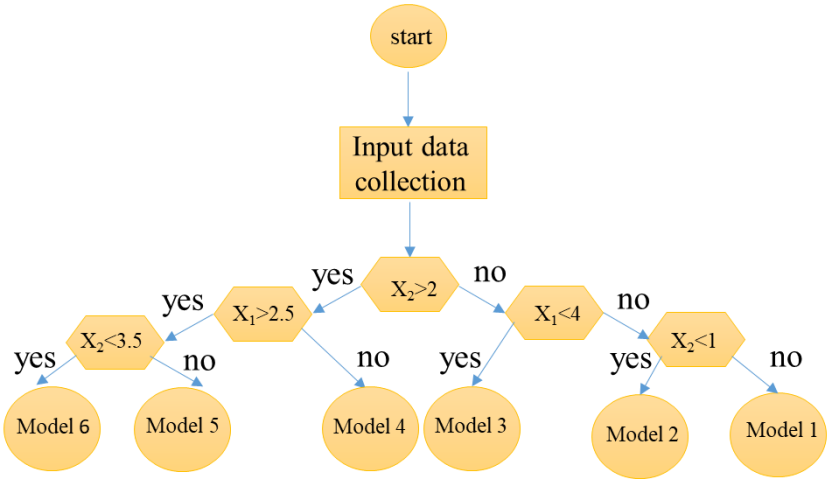

Fig. 3. Schematic structure of the M5 tree model.

2.5. Multivariate adaptive regression splines (MARS)

The MARS algorithm is a nonlinear and non-parametric method that Friedman introduced in 1992 (Fig. 4), whose structure is unknown before the modeling process [37]. The MARS model is a mathematical model whose internal function is based on a scattered polynomial and a piece known as the basis function (B) or splines. The k-node constrains the spacing, and the internal connections are applied at different time intervals from the input features. The MARS basis function is expressed as follows [38]:

$$B = \begin{cases} (k-x)^q & if \quad x < k \\ 0 & Otherwise \end{cases} \tag{4}$$

$$B = \begin{cases} (k-x)^q & if \quad x > k \\ 0 & Otherwise \end{cases} \tag{5}$$

where, q> 0 is the power that determines the polynomial function of the sub-piece. If q = 1, the splines are linear. If we want to obtain Y with M functions, the MARS model can be obtained by:

$$\hat{Y} = f_m(X) = C_0 + \sum_{M=1}^{N} C_m B_m(X) \tag{6}$$

where, Y is a prediction made by the model and C_0, C_m and B_m (X) are the constant, the basis function coefficient obtained by the least-squares method, and the basis function obtained by multiplying two or more functions, respectively, and M is the number of sentences in the final model. The modeling process in MARS is performed in forward and backward phases. In the first phase, important features are selected, while in the second phase, unnecessary samples are removed to prevent overfitting and enhance the model's accuracy. The unnecessary samples are removed by generalized cross-validation (GCV) as follows:

$$GCV = \frac{\sum_{i=1}^{N}\left(Y_i - \hat{Y}_i\right)^2}{\left(1 - \frac{M + p * \frac{(M-1)}{2}}{N}\right)^2} \tag{7}$$

where, p is penalty parameter. For more information please see [12,39]. Figure 4 shows the MARS model with q=1 and one feature.

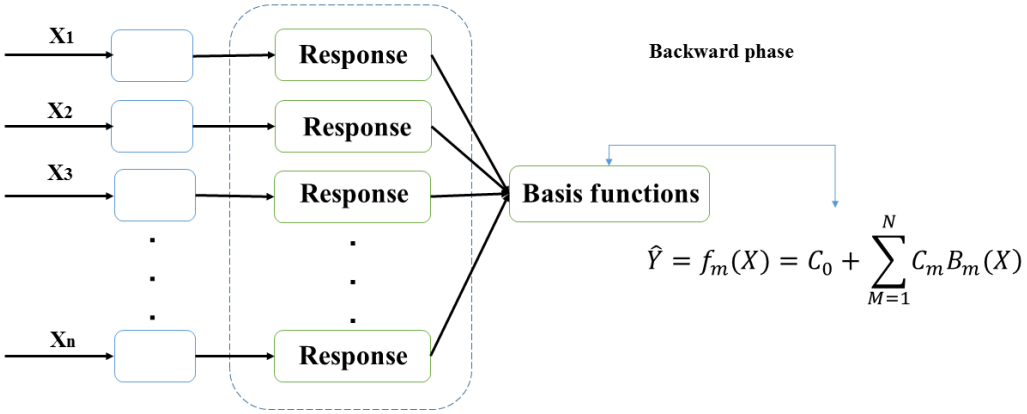

Fig. 4. Schematic structure of the MARS model.

2.6. Least square support vector machine (LS-SVM)

Support vector machines are efficient learning systems based on bounded optimization theory, which employs the structural error minimization inductive principle that results in a general optimal solution presented by Cortes and Vapnik in 1995 (Fig. 5)[40]. LS-SVM is a productive tool for tackling nonlinear issues, classification, and function estimation. The following regression model is used in the LS-SVM model to estimate various problems [19]:

$$Y(X_i) = W^T.\Phi(X_i) + b \tag{8}$$

In Eq.(8), $\Phi(X_i)$ are called nonlinear diagrams of the inputs in the feature space with high dimensions, and b and w are regression functions and weights of the dimensions of the same calculated property using objective function minimization according to the following equation:

$$\min_{w,e,b} j(w,e) = \frac{1}{2} w^T w + \frac{\gamma}{2} \sum_{i=1}^{N} e_i^2 \quad (9)$$

with the following restrictions:

$$y_i = w^T \Phi(x_i) + b + e_i \quad , \mathrm{I} = 1,\ 2,\ 3 \ldots \mathrm{N} \quad (10)$$

in Eqs. (9, 10), e_i is the error of training data, and γ is the penalty parameter and is called gamma. Large gamma values lead to more contribution of the error term in the objective function. Finally, the estimation function of the LS-SVM model is defined as follows:

$$y(x) = \sum_{i=1}^{N} a_i K(x_i, x_j) + b \quad (11)$$

where, $K(x_i, x_j)$ is kernel function described as a function of internal multiplication in the feature space. According to the following equation:

$$K(x_i, x_j) = \exp\left(-\frac{\left\| x_i - x_j \right\|^2}{2\sigma^2} \right) \quad (12)$$

where, σ or sigma is kernel width.

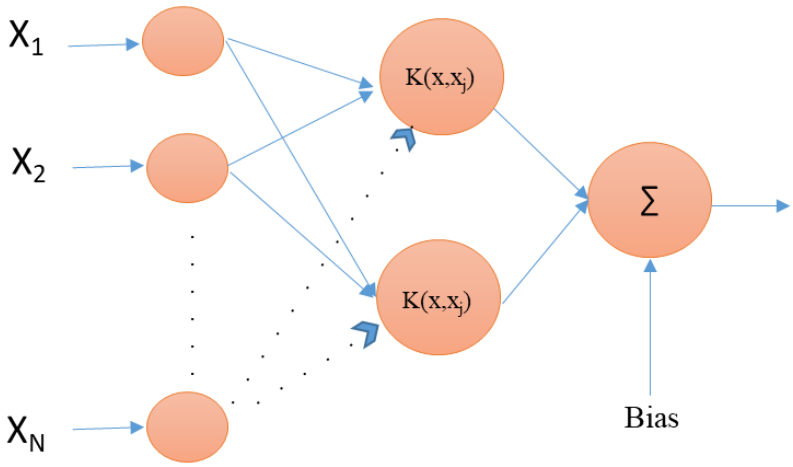

Fig. 5. Schematic structure of the LS-SVM model.

The gamma and sigma are essential parameters of LS-SVM that have essential influence on its efficiency.

2.7. Random forest (RF) model

The random forest algorithm, first introduced by Breiman in 2001[41], is a powerful and robust learning algorithm used for classification, regression analysis, and unsupervised learning goals [42]. In the RF algorithm, the user defines three parameters: Number of trees, minimum size in each terminal state or node size, and the number of variables to predict each tree [43] (Fig. 6). First, K random samples are generated by bootstrapping method. Then, for each sample, one decision tree is fitted. After that, the final results of RF are the average of the results of K trees. The final results in RF are estimated as follows [12,44]:

$$\hat{Y} = \frac{1}{K}\sum_{i=1}^{K} f_i(\Theta_i) \tag{13}$$

where, Θ_i is i[th] random sample and f_i is i[th] decision tree model.

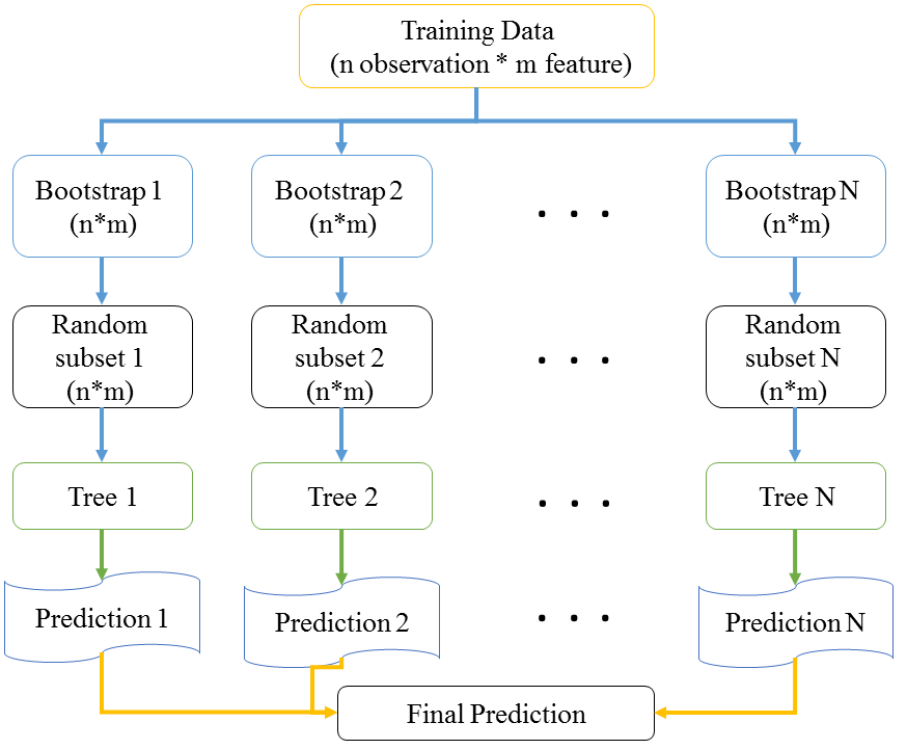

Fig. 6. Schematic structure of the random forest (RF) algorithm.

2.8. Empirical models for evaluating ET_o

A common approach for calculating crop ET consists of calculating ET_o and multiplying it by a crop coefficient.

Turc [45] developed an equation for calculating daily potential evapotranspiration as a function of air temperature, relative humidity, and solar radiation. The Turc method depends on the relative humidity of the air. If the relative humidity is greater than 50%, then:

$$ET_0 = 0.31\frac{T}{T+15}(Rs + 2.09) \tag{14}$$

If the relative humidity is less than 50%:

$$ET_0 = 0.31\frac{T}{T+15}(Rs + 2.09)(1+\frac{50-RH}{70}) \tag{15}$$

where, ET_o is daily reference crop evapotranspiration (mm/d), R_s is solar radiation (MJ/m^2.d), T is mean daily air temperature (°C), and RH is average daily relative humidity (%).

Jensen and Haise [46] developed an equation to predict potential evapotranspiration by combining the effect of temperature and solar radiation.

$$ET_0 = (0.0252T_m + 0.078)R_s \quad (16)$$

ET_o is in mm/d, T_m is mean daily temperature (°C), and *Rs* is short-wavelength incoming solar radiation to the earth's surface (MJ/m^2.d).

Hargreaves-Samani model is based on the maximum, minimum and average temperatures and radiation [47]:

$$ET_0 = 0.0023R_a(T_{mean} + 17.8)\sqrt{T_{\max} - T_{\min}} \quad (17)$$

where, ET_o is in mm/d, T_{max}, $T_{min,}$ and T_{mean} are maximum, minimum and average daily temperatures, respectively (°C), and R_a is extraterrestrial radiation (mm/d).

The United Nations Food and Agriculture Organization (FAO) has adopted the Penman-Monteith method in its Irrigation and Drainage Paper No. 56. Known as FAO 56 PM, this method is a global reference model for calculating reference crop evapotranspiration based on meteorological data [48]. It works well in different locations if the required data are available. It even works well in regions with limited data. Temperature, relative humidity, wind speed and solar radiation data are necessary for the FAO 56 PM method.

This model is derived from the following eqation [49]:

$$ET_0 = \frac{0.408\Delta(R_n - G) + \gamma(\frac{900}{T + 273})U_2(e_s - e_a)}{\Delta + \gamma(1 + 0.34U_2)} \quad (18)$$

where, ET_o is reference crop evapotranspiration (mm/d), T is mean daily air temperature at 2 m height (°C), U_2 is the wind speed at 2 m height (m/s), R_n is net radiation at the crop surface (MJ/m^2.d), G is soil heat flux density (MJ/m^2.d), e_s-e_a is saturation vapor pressure deficit (kPa), Δ is temperature-saturated vapor pressure curve gradient (kPa/°C), γ is psychrometric constant (kPa/°C).

This study selected 70% of the data as training data and 30% as testing data. Temperature, relative humidity, solar radiation, wind speed, and sunshine hours were used as model inputs, and the FAO 56 PM method is used as the output.

2.9. Evaluation criteria

The performance of data-mining approaches are compared based on the coefficient of determination (R^2), mean absolute error (MAE), root mean squared error (RMSE), and mean square error (MSE). The equations for calculating these criteria are given as [50–52]:

$$R^2 = \frac{n\left(\sum x_k y_k\right) - \left(\sum x_k\right)\left(\sum y_k\right)}{\sqrt{\left[n\sum x^2{}_k - \left(\sum x_k\right)^2\right]\left[n\sum y^2{}_k - \left(\sum y_k\right)^2\right]}} \quad (19)$$

$$MSE = \frac{1}{N}\sum_{i=1}^{n}\left(X_k - Y_k\right)^2 \quad (20)$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{n}\left(X_k - Y_k\right)^2} \tag{21}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{n}\left|X_k - Y_k\right| \tag{22}$$

In Eqs. (19-22), X_K is the observed value, Y_K is the estimated value, and N is number of data.

2.10. Model ranking

In this study, models and methods were ranked according to the presented method by [52], and by considering computation time and measurement accuracy. The lower the computation time and the higher the measurement accuracy, the better the model.

3. Results

In this study, mean monthly meteorological parameters including temperature-based variables, humidity-based variables, sunshine hours, wind direction, and wind velocity are used in the classical and modern models of ANN, M5, MARS, LS-SVM, and RF. In the following section, sensitivity analysis, results of applying the abovementioned data-mining methods in estimation of ET_o are reported for selected stations in Mazandaran and Semnan provinces.

3.1. Sensitivity analysis of data

The sensitivity analysis of input variables is done using correlation analysis in Semnan province (Fig. 7). In this method, the Pearson correlation between inputs and target (ET_o) variables are estimated. If the Pearson correlation is positive, it means that by increasing input variable, ET_o will increase. However, if Pearson correlation is negative, it means that by increasing the input variable, the ET_o will decrease. According to Fig. 7, by increasing the temperature-based variables, wind direction, wind speed, and sunshine hours, the ET_o increases, while increasing the humidity-based parameters decreases the ET_o.

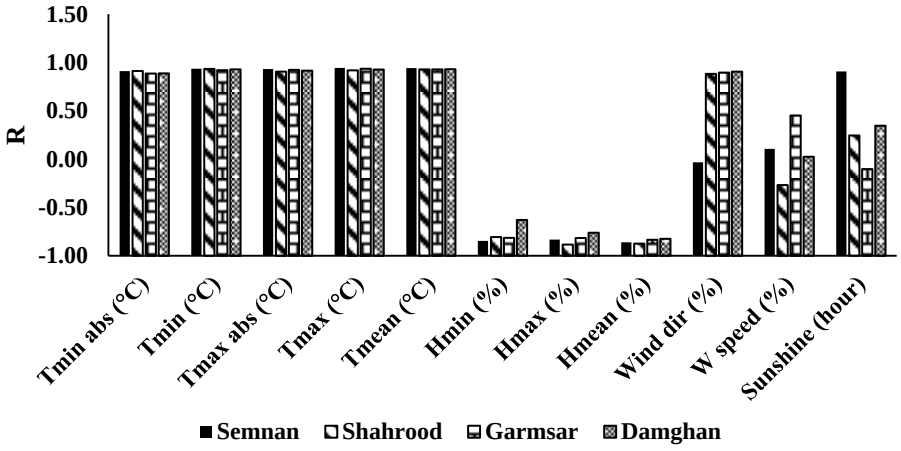

Fig. 7. Sensitivity analysis of input data for Semnan province.

3.2. Sensitivity analysis of machine-learning methods

Figure 8 shows MSE values for different number of trees for testing RF in the investigated stations. According to Fig. 8, RF in Semnan, Shahrood, Garmsar, and Damghan has better results with 100, 400, 200, and 500 trees, respectively. These values for Sari and Ghaemshahr are equal to 100 and for Dasht- e-naz and Babolsar are equal to 350. According to the results of this figure, MSE value is specific for each station. This issue is probably due to the impact of climate typeon the results of data-mining metods.

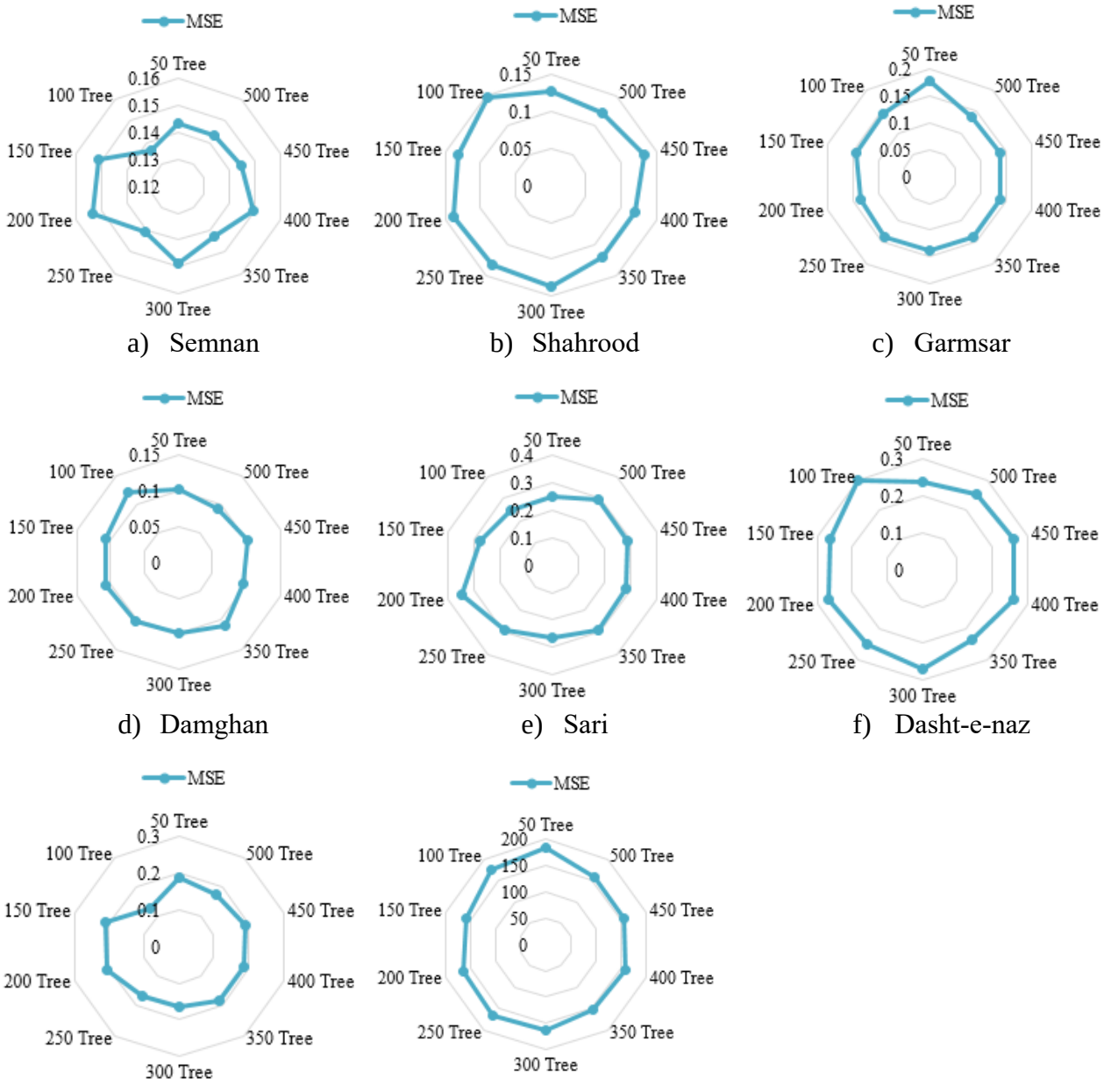

Fig. 8. Sensitivity analysis of RF for selecting the number of trees in: a) Semnan, b) Shahrood, c) Garmsar, d) Damghan, e) Sari, f) Dasht-e-naz, g) Ghaemshahr, and h) Babolsar.

Figure 9 demonstrates the contour plot of MSE for different values of gamma and sigma. It is seen that the best values of (gamma, sigma) for Semnan, Shahrood, Garmsar, and Damghan are equal to (8, 3), (10, 4), (8.5, 2.25), and (6.5, 2.75). The mentioned values for Sari, Daht-e-naz, Ghaemshahr, and Babolsar are estimated as (10, 5.5), (10, 6.75), (8.75, 5), and (4.75, 3).

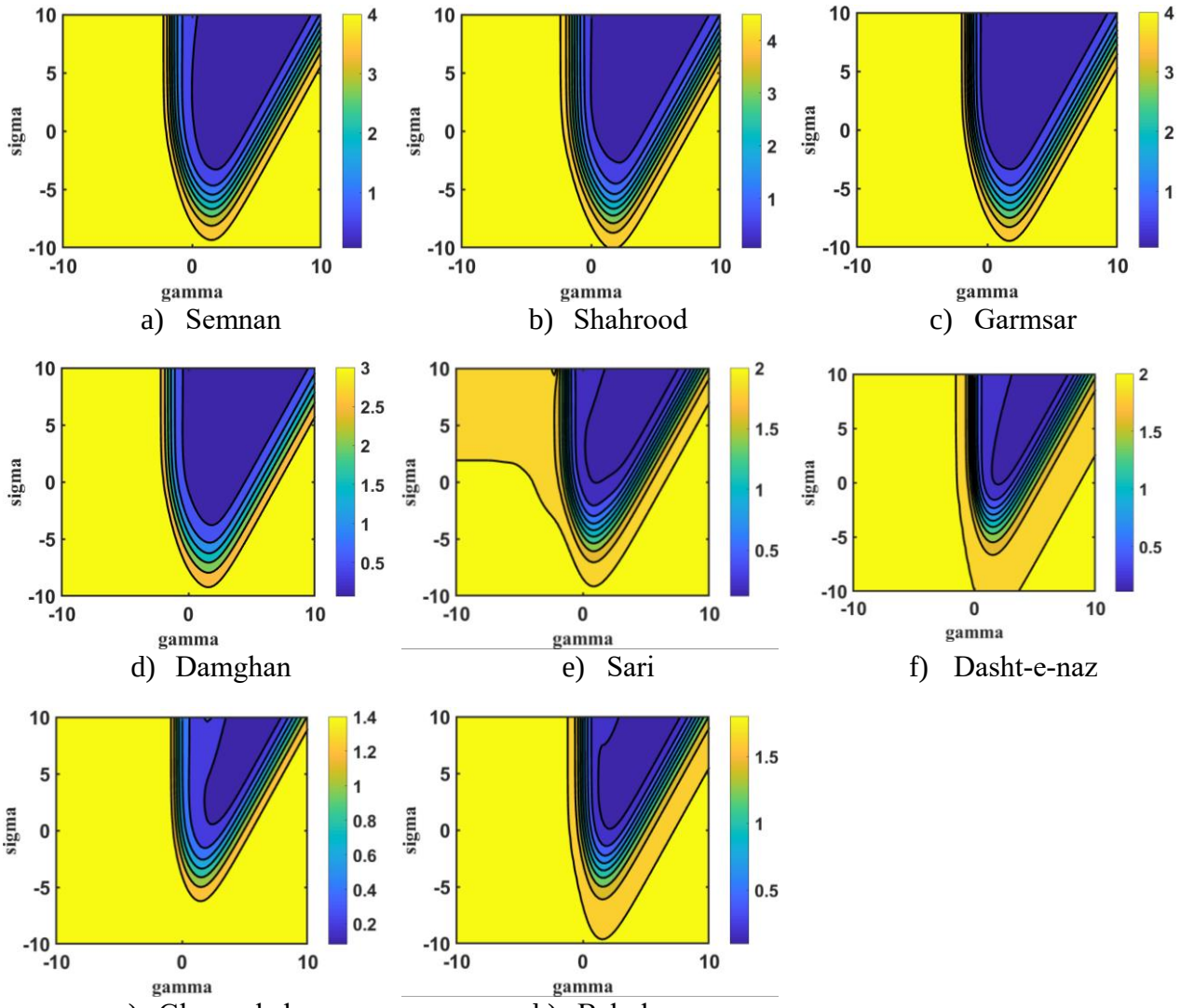

Fig. 9. Sensitivity analysis of LSSVM for selecting gamma and sigma in a) Semnan, b) Shahrood, c) Garmsar, d) Damghan, e) Sari, f) Dasht-e-naz, g) Ghaemshahr and h) Babolsar.

3.3. Estimation of ET_o by data-mining models

3.3.1. Mazandaran province

The ET_o estimation was performed by using M5, MARS, RF, LS-SVM, and ANN models for Sari station. As it is seen in Table 3, for all datasets, among the models, the MARS model has the highest mean coefficient of determination (0.9678) and coefficient of variation (0.0093). The MARS model also has the lowest MSE and RMSE. The recorded errors were 0.1149 and 0.3366,

respectively, and the coefficients of variation are 0.2397 and 0.1252. The M5 model has the lowest mean coefficient of determination (0.8548), coefficient of variation (0.0591), computation time (0.8272 s), lowest MAE (0.0478), and coefficient of variation (0.7894). The RF model has the highest error values. Mean values of MAE, MSE and RMSE are 0.0881, 0.1817, and 0.4435, respectively, and coefficients of variation are 0.4093, 0.3166, and 0.758, respectively. This model recorded a high computation time (1176.482 s).

Table 3
Statistics of data-mining models for Sari station.

	R^2	CV of R^2	MAE (mm/day)	CV of MAE	MSE $(mm/day)^2$	CV of MSE	RMSE (mm/day)	CV of RMSE	Time (s)
M5	0.8548	0.0591	0.0478	0.7894	0.2352	0.3275	0.4795	0.1597	0.8272
MARS	0.9678	0.0093	0.0821	1.475	0.1149	0.2397	0.3366	0.1252	1.0377
LS-SVM	0.9657	0.0082	0.0318	0.1153	0.1238	0.7991	0.3504	0.1832	37.6515
ANN	0.9406	0.0232	0.0525	0.7993	0.1989	0.1955	0.4429	0.0964	1.3575
RF	0.9589	0.0068	0.0881	0.4093	0.1817	0.3166	0.4435	0.0758	1176.482

Examination of the results of Babolsar station showed that among the models, the LS-SVM model has the highest coefficient of determination (0.9732), coefficient of variation (0.0072), the lowest error rate between the models (0.0382, 0.1071, and 0.3259, respectively) and coefficient of variation (0.7991, 0.1832, and 0.0943, respectively). The M5 model recorded the lowest mean coefficient of determination (0.8599), coefficient of variation (0.0962), and computation time (0.7864 s). Also, this model has the highest MAE (0.0885) and coefficient of variation (0.5125). The RF model has the highest MSE and RMSE values (0.2974 and 0.5357, respectively), and the coefficients of variation are 1.0948, 0.2939, and 0.1299, respectively. It also recorded a high average computation time of 1027.479 seconds.

In the Dasht-e-Naz station, among the models, the LS-SVM model has the highest average coefficient of determination (0.9665) and a coefficient of variation of 0.0058. The MARS model has the lowest MAE, MSE, and RMSE (0.0551, 0.1334, and 0.3610, respectively) and coefficient of variation (0.3069, 0.1624, and 0.5986, respectively). The M5 model has the lowest mean coefficient of determination (0.8853), coefficient of variation (0.0494), and computation time of 0.6631 seconds. The RF model has the highest MSE and RMSE values (0.3003 and 0.5457), coefficient of variation (0.867 and 0.962, respectively), and high average computation time (787.5662 s).

Analysis of the results for the Ghaemshahr station showed that among the models, the MARS model has the highest average coefficient of determination (0.9761) as well as coefficient of

variation (0.0074). The MARS model also has the lowest error values of MAE, MSE, and RMSE (0.0551, 0.1334, and 0.361) and coefficients of variation of (0.5986, 1962, and 1041, respectively). The M5 model has the lowest mean coefficient of determination (0.8969), coefficient of variation (0.0525), and less computation time (0.7579 s). The RF model had the highest error values of MAE, MSE, and RMSE (0.0671, 0.2189, and 0.4659, respectively), and the coefficients of variation are 0.6879, 1948, and 0.0984, respectively. It also recorded a high average computation time (1051.348 s).

According to the results, at all the stations, the M5 has the lowest R^2. It also has the least computation time, which can be advantageous for this model, especially when time matters to us.

Given the above results, the RF model has a good and acceptable coefficient of determination, but because of its high error rate and the high computation time, it is not recommended in the humid climate of Mazandaran province. Rankings of the models [53] are shown in Table 4.

In terms of computation time, M5 is ranked first with a score of 4, MARS is rated 2nd with a score of 8, and ANN, LSSVM, and RF models are next.

Table 4
Ranking of the smart models in terms of accuracy and computation time.

Station	Accuracy					Computation time				
	M5	**MARS**	**LSSVM**	**ANN**	**RF**	**M5**	**MARS**	**LSSVM**	**ANN**	**RF**
Babolsar	5	2	1	3	4	1	2	4	3	5
Dasht-e-Naz	5	1	2	3	4	1	2	4	3	5
Ghaemshahr	5	1	2	3	4	1	2	4	3	5
Sari	5	1	2	3	4	1	2	4	3	5
Total	20	5	7	12	16	4	8	16	12	20

In terms of accuracy, MARS is ranked 1st with a score of 5, and LSSVM, ANN, RF, and M5 models rank second to fifth. In terms of time and accuracy, MARS model is ranked 1st with a score of 13, and LSSVM model is 2nd with a score of 23, ANN and M5 models have the 3rd place with a score of 24, and RF model is ranked 4th with a score of 36. Figures 10 to 13 show the computed and observed MARS and M5 models in the Mazandaran climate.

Finally, to provide a comprehensive model in the humid climate, the data of 4 synoptic stations of Mazandaran province were implemented together. Results showed that MARS model with R^2, MAE, MSE, and RMSE values of 0.9637, 0.0267, 0.1266, and 0.3558, respectively, was the best model.

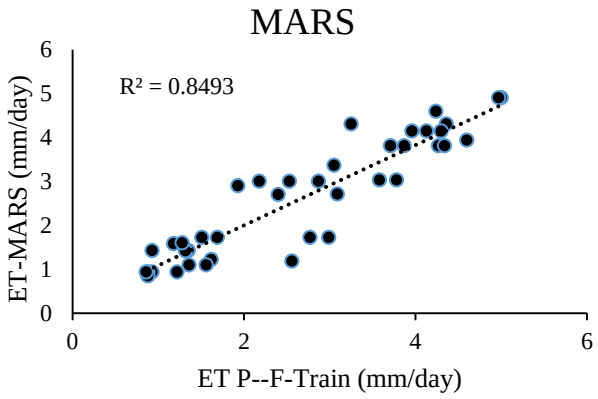

Fig. 10. Computed and observed values of ET_o for the MARS model in Mazandaran climate for training period.

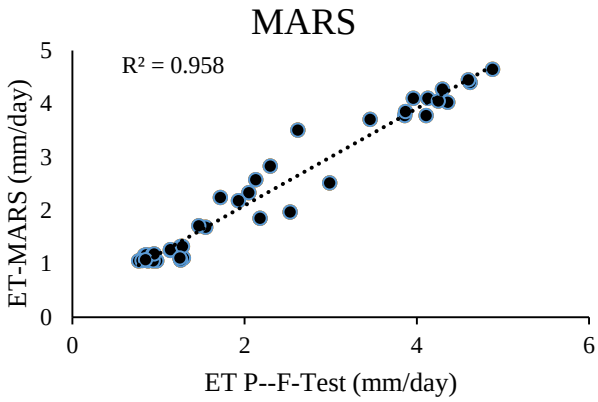

Fig. 11. Computed and observed values of ET_o for the MARS model in Mazandaran climate for the testing period.

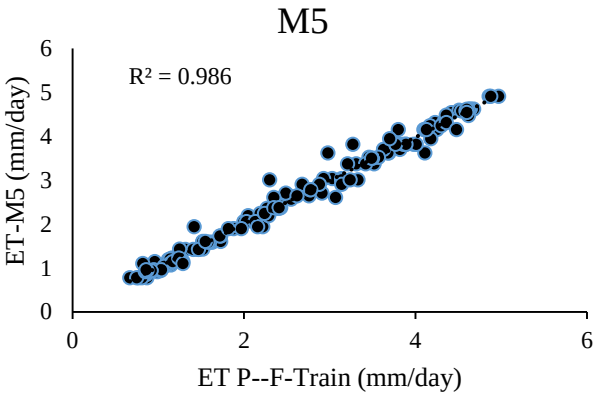

Fig. 12. Computed and observed values of ET_o for the M5 model in Mazandaran climate for the training period.

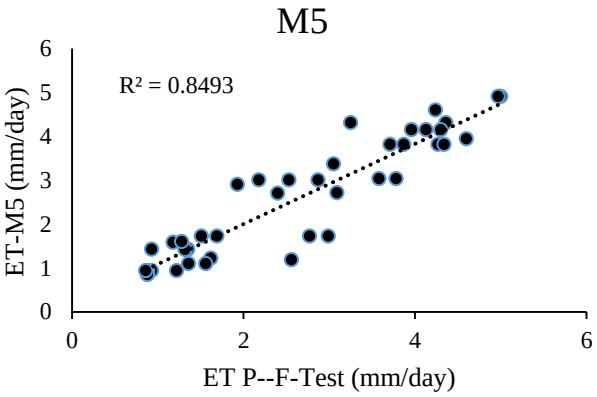

Fig. 13. Computed and observed values of ET_o for the M5 model in Mazandaran climate for the testing period.

3.3.2. Semnan province stations

Table 5 presents the results of data-mining models for Semnan station and all datasets. The LS-SVM model recorded the highest coefficient of determination (0.9896) and coefficient of variation (0.002). Also, this model has the highest MAE, MSE, and RMSE values (0.1036, 0.3288, and 0.5706, respectively), and the coefficients of variation are 1.3585, 0.2041, and 0.1059. The M5 model recorded the lowest coefficient of determination (0.9453) and coefficient of variation (0.0224). Also, the least computation time (1.024 s) is for M5. The ANN model has the lowest error rates of MAE, MSE, and RMSE (0.0393, 0.0987, and 0.03125, respectively), and the coefficients of variation are 0.9562, 0.02106, and 0.1117, respectively.

Table 5
Statistics of data mining models for Semnan station.

	R^2	CV of R^2	MAE (mm/day)	CV of MAE	MSE $(mm/day)^2$	CV of MSE	RMSE (mm/day)	CV of RMSE	TIME (s)
M5	0.9453	0.0224	0.043	0.6775	0.2075	0.2266	0.4529	0.1137	1.024
MARS	0.9857	0.0049	0.0407	0.6961	0.1031	0.3179	0.3174	0.1625	1.5812
LS-SVM	0.9896	0.002	0.1036	1.3585	0.3288	0.2041	0.5706	0.1059	55.6474
ANN	0.9892	0.0012	0.0393	0.9562	0.0987	0.2106	0.3125	0.1117	1.4096
RF	0.982	0.0038	0.0462	0.6243	0.13	0.1963	0.3589	0.1013	1435.75

For Damghan station, the LS-SVM model has the highest average coefficient of determination (0.9928) and coefficient of variation (0.0016). Also, this model has the highest MAE, MSE, and RMSE (0.1092, 0.3868, and 0.6161, respectively), and the coefficients of variation are 0.7157, 0.2756, and 0.1405, respectively. The M5 tree model recorded the lowest average coefficient of determination (0.9713) and coefficient of variation (0.0087). Also, the least computation time for the M5 tree was 0.7864 seconds. The ANN model has the least MAE, MSE, and RMSE values (0.0331, 0.1016, and 0.3151, respectively), and their coefficients of variation are 0.6997, 0.322 and 0.1607.

Examination of the results for Garmsar station indicates that among the models, the LS-SVM model has the highest average coefficient of determination (0.9923) and a coefficient of variation of 0.002. Also, this model has the highest MAE, MSE and RMSE values (0.0854, 0.3742, and 0.6099, respectively), and coefficients of variation (0.7615, 0.1642, and 0.0822, respectively). The M5 tree model recorded the lowest mean coefficient of determination (0.9653) and the coefficient of variation as 0.0043. Also, the lowest amount of computation time for M5 tree model is 0.9459 seconds. The MARS model has the least MAE, MSE, and RMSE values (0.0327, 0.0112, and 0.3198, respectively) and coefficients of variation (0.6628, 774, and 0.3251, respectively).

For Shahrood station, the LS-SVM model has the highest average coefficient of determination (0.9925) and coefficient of variation (0.0022). Also, this model has the highest MAE, MSE, and RMSE values (0.1122, 0.92994, and 0.5451, respectively) and coefficients of variation (0.4365, 0.1824, and 0.0918, respectively). The M5 tree model recorded the least coefficient of determination (0.9678) and coefficient of variation (0.0108). This model also has the least computation time (0.8991 s). The ANN model has the least MAE, MSE, and RMSE values (0.0328, 0.0756, and 0.2734, respectively), and the coefficients of variation are 0.6943, 0.2212, and 0.113, respectively.

Also, the abovementioned data-mining models (M5 tree, MARS, LS-SVM, ANN, and RF) are ranked for the four stations in the Semnan province. The results are presented in Table 6 and Table 7.

Table 6
Prioritizing data mining models in Semnan stations in terms of time.

	M5	MARS	LSSVM	ANN	RF
Damghan	1	2	4	3	5
Garmsar	1	3	4	2	5
Semnan	1	3	4	2	5
Shahrood	1	2	4	3	5
Total	4	10	16	10	20

Table 7
Prioritizing data mining models in Semnan stations in terms of accuracy.

	M5	MARS	LSSVM	ANN	RF
Damghan	5	2	3	1	4
Garmsar	5	1	3	2	4
Semnan	5	2	3	1	4
Shahrood	4	2	3	1	5
Total	19	7	12	5	17

In terms of time, according to Table 6, the M5 tree model, with a score of 4, ranked 1st, and the RF model, with a score of 20, ranked last.

In terms of accuracy, according to Table 7, the ANN model, with a score of 5, ranked 1st, and the MARS model, with a score of 7, ranked 2nd.

Taking the time and accuracy into account, the ANN model, with a total score of 15, had the least score and is chosen as the top model. The MARS, M5, LSSVM, and RF models rank next.

Figures 14-17 show the computed and observed values of the ET_0 by ANN and M5 models in the Semnan province.

Finally, to provide a comprehensive model in the arid climate, the data of 4 synoptic stations of Semnan province were combined with the best model (ANN model). The results are presented in Table 8.

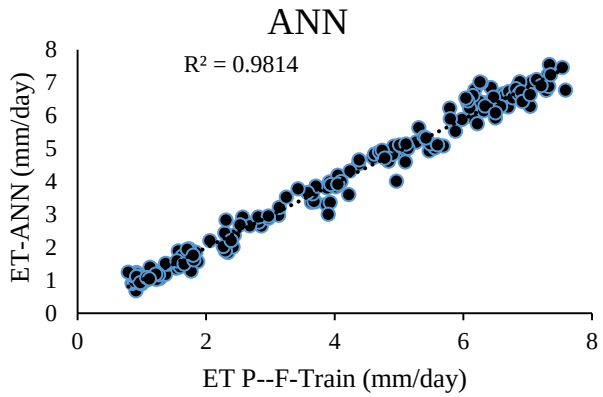

Fig. 14. Computed and observed values of ET_o data by ANN model in Semnan climate for the training period.

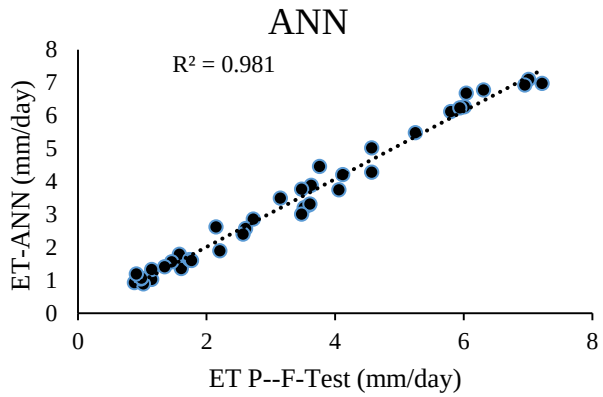

Fig. 15. Computed and observed values of ET_o data by ANN tree model in Semnan climate for the testing period.

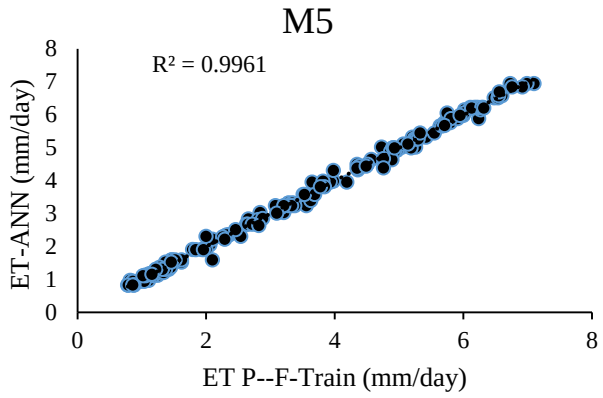

Fig. 16. Computed and observed values of ET_o data by M5 model in Semnan climate for the training period.

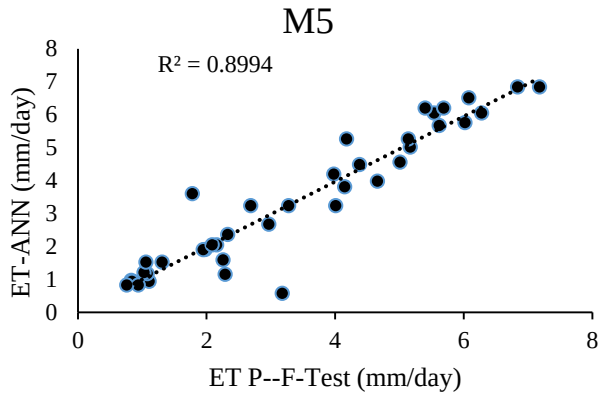

Fig. 17. Computed and observed values of ET_o data by M5 tree model in Semnan climate for the testing period.

Table 8
Comprehensive ET0 estimation model in arid climate.

	R^2	CV of R^2	MAE (mm/day)	CV of MAE	MSE $(mm/day)^2$	CV of MSE	RMSE (mm/day)	CV of RMSE	TIME (s)
ANN model in Semnan climate	0.985	0.0028	0.0238	0.807	0.1292	0.1782	0.3582	0.0872	2.9515

3.4. Results of empirical models in estimating ET_o

3.4.1. Meteorological stations of Mazandaran province

In this section, the results are presented for all datasets. The Jensen-Haise method has the highest determination coefficient (0.9843) for Sari station (Table 9). It also has the lowest error values of MAE, MSE, and RMSE (0.8764, 1.3582, and 1.1654, respectively). The Turc method has the

least determination coefficient (0.9089) and the highest error values of MAE, MSE, and RMSE (2.0398, 5.523, and 2.3501, respectively).

Table 9
Statistics of the empirical models for Sari station.

	R^2	MAE (mm/day)	MSE $(mm/day)^2$	RMSE (mm/day)
Hargreaves-Samani	0.9571	1.3049	2.6099	1.6155
Turc	0.9089	2.0398	5.523	2.3501
Jensen-Haise	0.9843	0.8764	1.3582	1.1654

At Babolsar station, the Jensen-Haise method has the highest determination coefficient (0.9897). The Hargreaves-Samani method has the lowest MAE, MSE, and RMSE values of 0.8334, 1.005, and 1.0025, respectively. The Turc method has the lowest determination coefficient (0.8907), and the highest MAE, MSE, and RMSE values are 2.947, 5.9462, and 2.4385, respectively.

At Dasht-e-Naz station, the Jensen-Haise method has the highest determination coefficient (0.9818). It also has the lowest MAE, MSE, and RMSE values of 0.8295, 1.2084, and 1.0993, respectively. The Turc method has the lowest determination coefficient (0.887), and the highest error values (2.2157, 6.503, and 2.5501, respectively).

At Gaemshahr station, the Jensen-Haise method has the highest determination coefficient (0.9702). It also has the lowest error values of 0.8054, 1.1236, and 1.06. The Turc method has the highest MAE, MSE, and RMSE values (2.0434, 5.6115 and 2.3689, respectively). The Hargreaves-Samani method has the lowest determination coefficient (0.7121).

Based on the above results, the Jensen-Haise method has the highest determination coefficient (R^2) and the lowest error values at all stations. The Turc method has the lowest determination coefficient, except for the Ghaemshahr station. It also recorded the highest error values at all stations. Thus, it can be concluded that the Jensen-Haise method is the best method for estimating ET_o in the humid climate of Mazandaran. The Hargreaves-Samani method ranks second, and the Jensen-Haise method ranks third.

3.4.2. Meteorological stations of Semnan province

At the Semnan station (Table 10), the Jensen Haise method has the highest determination coefficient (0.9429). It also has the lowest error rate of 1.2959, 3.089, and 1.7576. The highest error rates were 2.8954, 11.3384, and 3.3673, respectively, and the lowest determination coefficient was 0.6695 for the Turc method.

Table 10
Statistics of the empirical models for Semnan station.

	R^2	MAE (mm/day)	MSE $(mm/day)^2$	RMSE (mm/day)
Hargreaves-Samani	0.8613	1.9572	5.5159	2.3486
Turc	0.6695	2.8954	11.3384	3.3673
Jensen-Haise	0.9429	1.2995	3.089	1.7576

For Damghan synoptic station, The Jensen-Haise method has the highest determination coefficient (0.968). It also has the lowest error values (0.8396, 1.1506, and 1.0727, respectively). The highest error values are 3.3673, 15.0858, and 3.8841, and the lowest determination coefficient is 0.907.

The Hargreaves-Samani, Turc, and Jensen-Haise empirical methods are also investigated for Garmsar synoptic station. According to the results, the Jensen-Haise method has the highest determination coefficient (0.9347). It also has the lowest error values (1.1027, 2.2322, and 1.4941, respectively). The highest error values are 3.2794, 14.6375, and 3.8259, respectively, and the lowest determination coefficient is 0.8955.

Finally, for the Shahrood synoptic station, the Hargreaves-Samani method has the highest determination coefficient (0.9662). The Jensen-Haise method has the lowest error values (1.4531, 3.7023, and 1.9241, respectively). The highest error values are 2.8087, 10.1492, and 3.1858, respectively. The lowest determination coefficient is 0.8793.

Based on the above results, the Jensen-Haise method has obtained the highest determination coefficient and the lowest error values in all stations, except the Shahrood station (with very little difference from the Hargreaves-Samani method). The Turc method has the lowest determination coefficient among all the methods in all the stations. It also recorded the highest values of MAE, MSE, and RMSE values. According to the above results, it can be concluded that the Jensen-Haise method is the best method for estimating ET_o in the dry climate of Semnan province. The Hargreaves-Samani method ranked second, and the Turc ranked third.

A reasonable conclusion from this research is that the Jensen-Haise method is chosen as the superior method in both climates due to its high determination coefficient and low error values. In general, the results of this study, compared to other researches elsewhere, show high strength and ability of the proposed models in estimating the reference crop evapotranspiration.

4. Discussion

The critical difficulties in modeling with data-mining were the quality of inputs and target data, selecting machine-learning parameters, and calibration with deficient data. To this end, the pre-processing of input data was done by normalizing them. The parameters of data-mining methods were selected by sensitivity analysis as well as the experience of the authors. Random calibration method was used for training and testing data-mining methods to enhance accuracy. Also, by sensitivity analysis, most necessary inputs were selected for modeling ET_o. One of the main challenges that the present study faced was external disturbances, modeling errors, and uncertainties. To overcome these cases, each model was run 20 times, and inputs in each run were generated by the bootstrapping method. Then, in each run, the accuracy criteria were calculated and their coefficient of variation in all runs were estimated and reported. Indeed, the lower coefficient of variation indicates lower uncertainty, and by the multiple running of data-mining methods and using bootstrapping method, the impacts of uncertainty were reduced.

The good accuracy of ANN in modeling ET_o is reporte in [54–56] studies. According to the mentioned results, MARS and ANN have better results in humid and dry climates. Different parameters such as accuracy criteria, coefficient of variation of accuracy criteria, and computation times are considered, too. The reason for recommending ANN can be for its data processing in multilayers, using appropriate activation function and backpropagation as learning the algorithm, and consequently good accuracy and less computation time of this algorithm. While, other algorithms such as LSSVM and RF are not recommended, for significantly more computation times, despite their competitive accuracy. The more computation tim of LSSVM is for calculating kernel functions and trial and error for selecting their parameters.

Moreover, high computation time of RF is due to generating several random samples and fitting one decision tree to each sample. The excellent accuracy of MARS algorithm could be due to using the divide and conquer strategy in this algorithm. In MARS, the input sets are divided into multi subsets, and one spline regression is fitted to each subset. This ability helps MARS to consider the nonlinear relations between inputs and outputs with good accuracy. It is worth mentioning that different results of data-mining methods in the two different climates could be due to different statistical characteristics of inputs and target data in different stations. For example, the variation of ET_o in dry climate is more than humid climate. This leads to more accurate modeling of ET_o in humid climate than in dry climate. For better accuracy of data-mining methods compared to empirical equations, it can be said that data-mining methods use black-box approaches and data of investigated regions, and process the inputs and outputs data which leads to modeling with more accuracy.

The competitive results of MARS are similar to the studies by Mehdizaeh (2018) and Shan et al. (2020). On the other hand, the present study results indicated better accuracy of data-mining methods than empirical equations. This issue is reported by many studies such as Mehdizadeh et al. (2017) and Martin et al (2021).

5. Conclusions

Reference crop evapotranspiration (ET_o) is a variable used in irrigation planning, water resources management, and hydrological studies. Its other application is to estimate crop water requirement in large irrigation areas. This study used five data-mining methods: ANN, M5 decision tree, MARS, LS-SVM, and RF algorithms. Also, valid and applied empirical methods of Turc, Jensen-Haise, and Hargreaves-Samani are applied in this study, too. The well-known FAO Penman-Monteith method was used as a base for comparing the other three empirical models. A total of 8 synoptic stations in the humid and dry climates (Mazandaran and Semnan provinces, respectively) were considered. Mazandaran synoptic stations included Sari, Ghaemshahr, Babolsar, and Dasht-e Naz, and Semnan synoptic stations included Semnan, Shahrood, Damghan, and Garmsar. Results of this study revealed that in the humid climate of Mazandaran, the Jensen- Haise method was the best empirical model. The MARS model ranked first among the data-mining methods, and LS-SVM, ANN, RF, and M5 tree models ranked second to fifth. If the computation time is the criterion, the M5 tree model ranks first, and the MARS, ANN, LS-SVM, and RF models rank second to fifth. But if both accuracy and computation time are

considered, the MARS model ranks first, and ANN, LS-SVM, M5 tree, and RF models are in the second to fifth place. Finally, among all the models, the MARS model and the Jensen-Haise method in the humid climate of Mazandaran province are selected as the top choice.

In the dry climate of Semnan province, the Jensen-Haise method was the best method among the empirical equations. Among the data-mining models, the ANN comes first, and MARS, LS-SVM, RF, and M5 tree rank second to fifth in terms of accuracy. The M5 tree model ranks first, and MARS, ANN, LS-SVM, and RF models rank second to fifth if the computation time is considered. However if accuracy and computation time are considered, the ANN model ranks first. Finally, the ANN model and Jensen-Haise method were selected as the best model in the dry climate of Semna province. The better accuracy of MARS was for using the divide and conquer strategy in this algorithm and fitting one spline in each subset of the original dataset. Also, the excellent accuracy of ANN was for processing data in multilayers. Different results of modeling ET_o in each climate are due to different statistical characteristics of inputs and target times series.

Data-mining methods have high potential in solving various civil engineering problems if there is a suitable length and quality of data, choosing an excellent pre-processing method and calibration. However, it is necessary to consider the accuracy, uncertainty, and computation time in selecting these algorithms. Furthermore, it is suggested that the accuracy of MARS is enhanced by selecting parameters of MARS using sensitivity analysis or optimization algorithms in the future studies.

Funding

This research received no funding.

Conflicts of interest

The authors declare no conflict of interest.

Authors contribution statement

SF, SFM: Conceptualization; SZ: Data curation; SZ, SF, SFM, HS: Formal analysis; SZ, SF, SFM, HS: Investigation; SZ, SF, SFM, HS: Methodology; SZ, SF, SFM, HS: Project administration; SZ: Resources; SZ: Software; SF, SFM, HS: Supervision; SZ, SF, SFM, HS: Validation; SZ, SF, SFM, HS: Visualization; SZ: Roles/Writing – original draft; SF, SFM, HS: Writing – review & editing.

References

[1] Ferreira LB, França F, Oliveira RA De, Inácio E, Filho F. Estimation of reference evapotranspiration in Brazil with limited meteorological data using ANN and SVM ; a new approach. J Hydrol 2019. https://doi.org/10.1016/j.jhydrol.2019.03.028.

[2] Wu L, Zhou H, Ma X, Fan J, Zhang F. Daily reference evapotranspiration prediction based on hybridized extreme learning machine model with bio-inspired optimization algorithms: Application in contrasting climates of China. J Hydrol 2019;577:123960. https://doi.org/10.1016/j.jhydrol.2019.123960.

[3] Azad A, Farzin S, Kashi H, Sanikhani H, Karami H, Kisi O. Prediction of river flow using hybrid neuro-fuzzy models. Arab J Geosci 2018;11:718. https://doi.org/10.1007/s12517-018-4079-0.

[4] Mohammadi M, Farzin S, Mousavi S-F, Karami H. Investigation of a New Hybrid Optimization Algorithm Performance in the Optimal Operation of Multi-Reservoir Benchmark Systems. Water Resour Manag 2019;33:4767–82. https://doi.org/10.1007/s11269-019-02393-7.

[5] Valikhan-Anaraki M, Mousavi S-F, Farzin S, Karami H, Ehteram M, Kisi O, et al. Development of a Novel Hybrid Optimization Algorithm for Minimizing Irrigation Deficiencies. Sustainability 2019;11:2337. https://doi.org/10.3390/su11082337.

[6] Karami H, Ehteram M, Mousavi S-F, Farzin S, Kisi O, El-Shafie A. Optimization of energy management and conversion in the water systems based on evolutionary algorithms. Neural Comput Appl 2019;31:5951–64. https://doi.org/10.1007/s00521-018-3412-6.

[7] Azad A, Manoochehri M, Kashi H, Farzin S, Karami H, Nourani V, et al. Comparative evaluation of intelligent algorithms to improve adaptive neuro-fuzzy inference system performance in precipitation modelling. J Hydrol 2019;571:214–24. https://doi.org/10.1016/j.jhydrol.2019.01.062.

[8] Azad A, Farzin S, Sanikhani H, Karami H, Kisi O, Singh VP. Approaches for Optimizing the Performance of Adaptive Neuro-Fuzzy Inference System and Least-Squares Support Vector Machine in Precipitation Modeling. J Hydrol Eng 2021;26:04021010. https://doi.org/10.1061/(ASCE)HE.1943-5584.0002069.

[9] Farzin S, Nabizadeh Chianeh F, Valikhan Anaraki M, Mahmoudian F. Introducing a framework for modeling of drug electrochemical removal from wastewater based on data mining algorithms, scatter interpolation method, and multi criteria decision analysis (DID). J Clean Prod 2020;266:122075. https://doi.org/10.1016/j.jclepro.2020.122075.

[10] Anaraki MV, Farzin S, Mousavi S-F, Karami H. Uncertainty Analysis of Climate Change Impacts on Flood Frequency by Using Hybrid Machine Learning Methods. Water Resour Manag 2021;35:199–223. https://doi.org/10.1007/s11269-020-02719-w.

[11] Siahkali MZ, Ghaderi A, Bahrpeyma A, Rashki M. Estimating Pier Scour Depth : Comparison of Empirical Formulations. J AI Data Min 2021;9:109–28. https://doi.org/10.22044/jadm.2020.10085.2147.

[12] Safaeian Hamzehkolaei N, Alizamir M. Performance evaluation of machine learning algorithms for seismic retrofit cost estimation using structural parameters. J Soft Comput Civ Eng 2021;5:32–57. https://doi.org/10.22115/SCCE.2021.284630.1312.

[13] Traore S, Wang Y-M, Kerh T. Artificial neural network for modeling reference evapotranspiration complex process in Sudano-Sahelian zone. Agric Water Manag 2010;97:707–14. https://doi.org/10.1016/j.agwat.2010.01.002.

[14] Rahimikhoob A, Behbahani MR, Fakheri J. An evaluation of four reference evapotranspiration models in a subtropical climate. Water Resour Manag 2012;26:2867–81.

[15] Yassin MA, Alazba AA, Mattar MA. Artificial neural networks versus gene expression programming for estimating reference evapotranspiration in arid climate. Agric Water Manag 2016;163:110–24. https://doi.org/10.1016/j.agwat.2015.09.009.

[16] Caminha HD, Da Silva TC, Da Rocha AR, Lima SCRV. Estimating reference evapotranspiration using data mining prediction models and feature selection. ICEIS 2017 - Proc 19th Int Conf Enterp Inf Syst 2017;1:272–9. https://doi.org/10.5220/0006327202720279.

[17] Mehdizadeh S. Estimation of daily reference evapotranspiration (ETo) using artificial intelligence methods: Offering a new approach for lagged ETo data-based modeling. J Hydrol 2018. https://doi.org/10.1016/j.jhydrol.2018.02.060.

[18] Ehteram M, Singh VP, Ferdowsi A, Mousavi SF, Farzin S, Karami H, et al. An improved model based on the support vector machine and cuckoo algorithm for simulating reference evapotranspiration. PLoS One 2019;14:e0217499. https://doi.org/10.1371/journal.pone.0217499.

[19] Wang S, Lian J, Peng Y, Hu B, Chen H. Generalized reference evapotranspiration models with limited climatic data based on random forest and gene expression programming in Guangxi, China. Agric Water Manag 2019;221:220–30. https://doi.org/10.1016/j.agwat.2019.03.027.

[20] Fan J, Ma X, Wu L, Zhang F, Yu X, Zeng W. Light Gradient Boosting Machine: An efficient soft computing model for estimating daily reference evapotranspiration with local and external meteorological data. Agric Water Manag 2019;225:105758. https://doi.org/10.1016/j.agwat.2019.105758.

[21] Ferreira LB, da Cunha FF. New approach to estimate daily reference evapotranspiration based on hourly temperature and relative humidity using machine learning and deep learning. Agric Water Manag 2020;234:106113. https://doi.org/10.1016/j.agwat.2020.106113.

[22] Granata F, Gargano R, de Marinis G. Artificial intelligence based approaches to evaluate actual evapotranspiration in wetlands. Sci Total Environ 2019:135653. https://doi.org/10.1016/j.scitotenv.2019.135653.

[23] Yamaç SS, Todorovic M. Estimation of daily potato crop evapotranspiration using three different machine learning algorithms and four scenarios of available meteorological data. Agric Water Manag 2020;228:105875. https://doi.org/10.1016/j.agwat.2019.105875.

[24] Ashrafzadeh A, Kişi O, Aghelpour P, Biazar SM, Masouleh MA. Comparative Study of Time Series Models, Support Vector Machines, and GMDH in Forecasting Long-Term Evapotranspiration Rates in Northern Iran. J Irrig Drain Eng 2020;146:04020010. https://doi.org/10.1061/(ASCE)IR.1943-4774.0001471.

[25] Zhang M, Su B, Nazeer M, Bilal M, Qi P, Han G. Climatic Characteristics and Modeling Evaluation of Pan Evapotranspiration over Henan Province, China. Land 2020;9:229. https://doi.org/10.3390/land9070229.

[26] Rashid Niaghi A, Hassanijalilian O, Shiri J. Estimation of Reference Evapotranspiration Using Spatial and Temporal Machine Learning Approaches. Hydrology 2021;8:25. https://doi.org/10.3390/hydrology8010025.

[27] Feng K, Tian J. Forecasting reference evapotranspiration using data mining and limited climatic data. Eur J Remote Sens 2021;54:363–71. https://doi.org/10.1080/22797254.2020.1801355.

[28] Kadkhodazadeh M, Farzin S. A Novel LSSVM Model Integrated with GBO Algorithm to Assessment of Water Quality Parameters. Water Resour Manag 2021;35:3939–68. https://doi.org/10.1007/s11269-021-02913-4.

[29] Mohaghegh A, Valikhan Anaraki M, Farzin S. Modeling of qualitative parameters (Electrical conductivity and total dissolved solids) of Karun River at Mollasani, Ahvaz and Farsiat stations using data mining methods. Iran J Heal Environ 2020;13:101–20.

[30] Nourani V, Jabbarian Paknezhad N, Sharghi E, Khosravi A. Estimation of prediction interval in ANN-based multi-GCMs downscaling of hydro-climatologic parameters. J Hydrol 2019;579:124226. https://doi.org/10.1016/j.jhydrol.2019.124226.

[31] Antonopoulos VZ, Antonopoulos A V. Daily reference evapotranspiration estimates by artificial neural networks technique and empirical equations using limited input climate variables. Comput Electron Agric 2017;132:86–96. https://doi.org/10.1016/j.compag.2016.11.011.

[32] Valikhan Anaraki M, Mousavi S-F, Farzin S, Karami H. Introducing a Nonlinear Model Based on Hybrid Machine Learning for Modeling and Prediction of Precipitation and Comparison with SDSM Method (Cases Studies: Shahrekord, Barez, and Yasuj). Iran J Soil Water Res 2020;51:325–39.

[33] Kisi O. Pan evaporation modeling using least square support vector machine, multivariate adaptive regression splines and M5 model tree. J Hydrol 2015;528:312–20. https://doi.org/10.1016/j.jhydrol.2015.06.052.

[34] Adnan RM, Liang Z, Trajkovic S, Zounemat-Kermani M, Li B, Kisi O. Daily streamflow prediction using optimally pruned extreme learning machine. J Hydrol 2019;577:123981. https://doi.org/10.1016/j.jhydrol.2019.123981.

[35] Kisi O, Parmar KS. Application of least square support vector machine and multivariate adaptive regression spline models in long term prediction of river water pollution. J Hydrol 2016;534:104–12. https://doi.org/10.1016/j.jhydrol.2015.12.014.

[36] Rezaie-Balf M, Kim S, Fallah H, Alaghmand S. Daily river flow forecasting using ensemble empirical mode decomposition based heuristic regression models: Application on the perennial rivers in Iran and South Korea. J Hydrol 2019;572:470–85. https://doi.org/10.1016/j.jhydrol.2019.03.046.

[37] Keshtegar B, Heddam S, Kisi O, Zhu SP. Modeling total dissolved gas (TDG) concentration at Columbia river basin dams: high-order response surface method (H-RSM) vs. M5Tree, LSSVM, and MARS. Arab J Geosci 2019;12. https://doi.org/10.1007/s12517-019-4687-3.

[38] Abdulelah Z, Sudani A, Salih SQ, Yaseen ZM. Development of Multivariate Adaptive Regression Spline Integrated with Differential Evolution Model for Streamflow Simulation Computer Science Department , College of Computer Science and Information Technology , Sustainable developments in Civil Engineer. J Hydrol 2019. https://doi.org/10.1016/j.jhydrol.2019.03.004.

[39] Zhang G, Hamzehkolaei NS, Rashnoozadeh H, Band SS, Mosavi A. Reliability assessment of compressive and splitting tensile strength prediction of roller compacted concrete pavement: introducing MARS-GOA-MCS. Int J Pavement Eng 2021:1–18. https://doi.org/10.1080/10298436.2021.1990920.

[40] Cortes C, Vapnik V. Support vector machine. Mach Learn 1995;20:273–97.

[41] Breiman L. Random Forests. Mach Learn 2001;45:5–32. https://doi.org/10.1023/A:1010933404324.

[42] Forghani SJ, Pahlavan-Rad MR, Esfandiari M, Torkashvand AM. Spatial prediction of WRB soil classes in an arid floodplain using multinomial logistic regression and random forest models, south-east of Iran. Arab J Geosci 2020;13. https://doi.org/10.1007/s12517-020-05576-4.

[43] Zhang Y, Sui B, Shen H, Ouyang L. Mapping stocks of soil total nitrogen using remote sensing data : A comparison of random forest models with di ff erent predictors. Comput Electron Agric 2019;160:23–30. https://doi.org/10.1016/j.compag.2019.03.015.

[44] Crawford J, Venkataraman K, Booth J. Developing climate model ensembles : A comparative case study. J Hydrol 2019;568:160–73. https://doi.org/10.1016/j.jhydrol.2018.10.054.

[45] Turc L. Water requirements assessment of irrigation, potential evapotranspiration: simplified and updated climatic formula. Ann. Agron., vol. 12, 1961, p. 13–49.

[46] Jensen ME, Haise HR. Estimating evapotranspiration from solar radiation. Proc Am Soc Civ Eng J Irrig Drain Div 1963;89:15–41.

[47] Ahmadi H, Baaghideh M. Assessment of anomalies and effects of climate change on reference evapotranspiration and water requirement in pistachio cultivation areas in Iran. Arab J Geosci 2020;13. https://doi.org/10.1007/s12517-020-05316-8.

[48] Mossad A, Alazba AA. Simulation of temporal variation for reference evapotranspiration under arid climate. Arab J Geosci 2016;9. https://doi.org/10.1007/s12517-016-2482-y.

[49] Dinpashoh Y, Babamiri O. Trends in reference crop evapotranspiration in Urmia Lake basin. Arab J Geosci 2020;13. https://doi.org/10.1007/s12517-020-05404-9.

[50] Farrokhi A, Farzin S, Mousavi S-F. A New Framework for Evaluation of Rainfall Temporal Variability through Principal Component Analysis, Hybrid Adaptive Neuro-Fuzzy Inference System, and Innovative Trend Analysis Methodology. Water Resour Manag 2020;34:3363–85. https://doi.org/10.1007/s11269-020-02618-0.

[51] Ghazvinian H, Karami H, Farzin S, Mousavi SF. Effect of MDF-Cover for Water Reservoir Evaporation Reduction, Experimental, and Soft Computing Approaches. J Soft Comput Civ Eng 2020;4:98–110. https://doi.org/10.22115/scce.2020.213617.1156.

[52] Farzin S, Valikhan Anaraki M. Modeling and predicting suspended sediment load under climate change conditions: a new hybridization strategy. J Water Clim Chang 2021. https://doi.org/10.2166/wcc.2021.317.

[53] Sanikhani H, Deo RC, Samui P, Kisi O, Mert C, Mirabbasi R. Survey of di ff erent data-intelligent modeling strategies for forecasting air temperature using geographic information as model predictors. Comput Electron Agric 2018;152:242–60. https://doi.org/10.1016/j.compag.2018.07.008.

[54] Sattari MT, Apaydin H, Band SS, Mosavi A, Prasad R. Comparative analysis of kernel-based versus ANN and deep learning methods in monthly reference evapotranspiration estimation. Hydrol Earth Syst Sci 2021;25:603–18. https://doi.org/10.5194/hess-25-603-2021.

[55] Sayyahi F, Farzin S, Karami H. Forecasting Daily and Monthly Reference Evapotranspiration in the Aidoghmoush Basin Using Multilayer Perceptron Coupled with Water Wave Optimization. Complexity 2021;2021:1–12. https://doi.org/10.1155/2021/6683759.

[56] Gao L, Gong D, Cui N, Lv M, Feng Y. Evaluation of bio-inspired optimization algorithms hybrid with artificial neural network for reference crop evapotranspiration estimation. Comput Electron Agric 2021;190:106466. https://doi.org/10.1016/j.compag.2021.106466.