Water and Soil Management and Modeling

Online ISSN: 2783 - 2546

1978

University of Mohaghegh Ardabili

Statistical and multi-criteria methods for preprocessing meteorological data in reference evapotranspiration

Laleh Parviz[1*], Fardin Ghanbari-Maleki[2]

[1] Associate Professor, Faculty of Agriculture, Azarbaijan Shahid Madani University, Tabriz, Iran
[2] M.Sc Student, Faculty of Agriculture, Azarbaijan Shahid Madani University, Tabriz, Iran

Abstract

The capability of reference evapotranspiration (ET_0) in water loss consideration benefits irrigation planning, agricultural studies, and water resources management. The requirement for effective meteorological data determination is noteworthy in the ET_0 estimation, in which Pearson, Kendall's tau-b correlation coefficients, the standardized Beta coefficient, stepwise regression (in statistical approach), simple additive weighting with fuzzy normalization (F-SAW), and Shannon's entropy (multi-criteria) were considered as the preprocessing methods. Different combinations of meteorological data in 11 synoptic stations in Iran were applied to test the performance of the methods. The first three statistical methods (Pearson, Kendall's tau-b correlation coefficients, and the standardized Beta coefficient) yield similar results. Root mean square error decreasing from Pearson and stepwise regression to F-SAW is 15.4% and 14.7%, respectively; therefore, the F-SAW could enhance the ET_0 simulations. The differences between the two preprocessing methods based on the MCDM approach in all 132 cases are low, except for 34 cases with better performance of F-SAW. F-SAW performance investigation among all stations on annual and monthly scales stated that Zanjan and Semnan stations, with Nash-Sutcliffe efficiency equal to 0.97 and 0.99, respectively, have better performance than the other stations. The temperature, humidity, wind speed, and solar radiation were obtained as the effective data. In the decision-making analysis with high efficiency, the procedure of assigning weights to the criteria has an important role, and the high performance of F-SAW can be linked to the normalization structure. In the different climates of stations, the performance of each dataset is distinctive; therefore, an efficient preprocessing method can upgrade the adequacy of ET_0 estimation.

Keywords: Preprocessing; Reference evapotranspiration (ET_0); Fuzzy normalization; Combinations.

Article Type: Research Article

Academic Editor: Zainab Hazbavi

*Corresponding Author, E-mail: lparviz@azaruniv.ac.ir

Citation: Parviz, L., & Fardin Ghanbari, M. (2025). Comparison of methods' efficiency based on statistical and multi-criteria decision-making approaches as the preprocessing technique of reference evapotranspiration. Water and Soil Management and Modeling, 5 (Special Issue: Newfound Approaches to Water and Soil Management and Modeling), 252–269.

doi: 10.22098/mmws.2025.18259.1669

Received: 06 September 2025, Received in revised form: 23 September 2025, Accepted: 03 October 2025, Published online: 01 November 2025

Water and Soil Management and Modeling, Year 2025, Vol. 5, Special Issue, pp. 252-269.

Publisher: University of Mohaghegh Ardabili

1. Introduction

Evapotranspiration refers to the process by which water moves from soil, water resources, and plants into the atmosphere. It is a crucial indicator for understanding the water cycle in ecosystems. Knowledge of actual evapotranspiration is valuable for assessing water availability and conducting climate change studies (Yao and Malik, 2022; Zhu et al., 2024). Despite all improvements, the measurement of actual evapotranspiration is accompanied by difficulty in some locations. In this regard, an accurate method for actual evapotranspiration estimation is linked to the reference evapotranspiration (ET_0) determination as a significant component. The ET_0 computation methods can be classified into four groups, which are based on mass transfer, temperature, radiation, and combination. The Food and Agriculture Organization of the United Nations (FAO) Penman-Monteith method is widely recognized for its high accuracy and has been validated across different climatic conditions, making it a globally accepted standard (Shu et al., 2022). Also, the FAO Penman-Monteith method can test the other ET_0 computation methods and exhibit their accuracy, such as evaluating the modified Hargreaves-Samani, Thornthwaite, and Blaney-Criddle methods with the FAO 56 Penman-Monteith model (Yadeta et al., 2020). Also, the machine learning models can be applied to estimate ET_0, for instance, ET_0 determination with the artificial neural network (ANN) and climate-based methods (Chauhan and Shrivastava 2009) and with the hybrid ANN-gray wolf optimization (Maroufpoor et al., 2020).

Input data is one of the important factors in the ET_0 estimation (Hu et al., 2023). The need to multiply data as the input of the model is one of the problems related to the FAO 56 Penman-Monteith method, which a preprocessing technique can conquer this issue. A review of research indicated that a comprehensive investigation of preprocessing methods has not been done for dominant input data determining the potential evapotranspiration, and there is little research in this area. Also, the commonly used methods were based on statistical concepts, while the use of multi-criteria decision making can be considered a valid method.

Tabari and Talaee (2013) tested different combinations of meteorological data using a multilayer perceptron network model with high-performance input data of mean temperature, relative humidity, and wind speed at 2 m. The other research focused on linking the extreme learning machine with the optimization method as an alternative method to estimate ET_0 using the limited meteorological data. Gong et al. (2021) found that air temperature could be introduced as the dominant data, especially in the field with no radiation data. The limited number of synoptic stations and observed data can decrease the precision of estimated ET_0; therefore, the Shu et al. (2022) proposed method in a study in China was considered in the empirical and optimized machine learning models to simulate ET_0 in the field with low data. The empirical models consist of Albrecht, Hargreaves-Samani, Priestley-Taylor, Penman, and multiple linear regression, and random forest for the machine learning models, which they regionalize the empirical model parameters. The incorporation of empirical and machine learning models could obtain an accurate simulated ET_0 in the field with scarce data. In the mentioned research, only a modeling structure was applied to discover the relationship between data and ET_0 without any data investigation, while the importance of input data into the ET_0 models revealed the need for preprocessing methods to uncover the dominant climatic data as the input data for ET_0 estimation. Ghabaei Sough et al. (2010) used the stepwise regression and Gamma test approaches to evaluate the effect of the preprocessing step in daily evapotranspiration estimation. The results showed that coupling ANN with the gamma test is a useful tool in the preprocessing steps. The preprocessing method in the previous research is based on statistical concepts, but the multi-criteria decision-making approach can be a good selection in this regard. The potential of the entropy method in the Southeastern United States was examined to reveal important information about the data by Ellenburg et al. (2017). The Shannon entropy for each data was computed for daily 8 and 32-day aggregations. The different ET data from the energy budget approach and based on a complex iterative solution and a modified Penman method, were compared, and the results showed a considerable difference in the uncertainty of the methods. An optimal input combination was investigated to estimate ET_0 with a hybrid artificial intelligence model in some synoptic stations of Iran. The research tries to capture the uncertainties of variables by information entropy principles (Shannon entropy). Entropy analysis indicates the importance of each variable at each location, which has a different climate. The model of Maroufpoor et al. (2020) with inputs of minimum and maximum temperature and wind speed had the lowest error. Ahmadi et al. (2021) introduced the τ Kendall and entropy as the preprocessing methods, which showed that the impressive data are the air temperature consisting of minimum, maximum, mean temperature, and solar radiation for ET_0 estimation. The important outcome of this research is the significance of preprocessing methods for effective data determination. Kim et al. (2023) increased the efficiency of ET_0 methods without a high number of input data by using the recalibration process, which was proposed for the Hargreaves-Samani model parameters. The recalibrated Hargreaves-Samani had a better fit of estimated ET_0 by the FAO-56-Penman-Monteith method. This type of calibration is a kind of modeling process, and it does not emphasize the input data. In some studies, such as Saroughi et al. (2024), the series decomposition methods, such as wavelet transform methods, were applied as the preprocessing step, like using the variational mode decomposition method in evapotranspiration estimation in dry land (Fu et al., 2021). This approach works only on decomposed time series, for which the number is high. In this case, the dominant input data selection is necessary to decrease the number of decomposed series, which reveals the importance of the preprocessing method of this study.

Despite the acceptability of the FAO Penman-Monteith method, the need for a large amount of reliable weather measurements, such as solar radiation and wind speed, has challenged the method. These data are often not available in developing countries, and the issue is related to the limited number of equipped meteorological stations or inaccuracies of measurement (Berti et al., 2014). Therefore, the need for an alternative ET_0 method seems necessary, and efficient artificial intelligence techniques with a low number of input data can obtain accuracy equal to the FAO method. In this regard, the preprocessing step with a selection of important input data is more important. Many studies were focused on the statistical preprocessing method, and the need for a new method is felt. This study introduces a novel approach by systematically comparing multiple preprocessing methods for ET_0 estimation by integrating decision-making techniques to improve data selection and model accuracy. The preprocessing methods belong to the correlation concept, regression analysis, and decision-making approach, with different normalization methods. To increase the accuracy of decisions, more than one evaluation criteria were considered in the analysis.

2. Materials and methods

2.1. Study area description

The analysis of this study is focused on eleven stations (1992-2021). The station's spatial distribution consists of the North, West, North-West, East, and the center of Iran (Figure 1). The studied stations are designed to examine the validity of different proposed methods in diverse climates. The maximum and minimum annual precipitation are related to the Ramsar and Yazd stations, respectively. Two formulaic (De Martonne (1925) and precipitation effectiveness index, PEI) and graphical methods (climagraph) were used as a way to analyze the climate of stations. The formula section with two different mathematical structures and the graphical section combining meteorological data attempt to extract the prevailing climatic conditions in the region in a different manner. The De Martonne climate classification of stations includes Ramsar (very humid)- Kermanshah, Hamadan, Maku, Zanjan, Shiraz, Ardabil, Urmia, Mashhad (Semi-arid)- Semnan, Yazd (arid). Yazd and Ramsar have the minimum and maximum values of the De Martonne index. The different intensities of the index reflect the climatic diversity of the studied stations. The high values of PEI in Figure 2 are indicative of humidity, which Yazd and Ramsar had the minimum and maximum values of PEI, respectively. As shown in Figure 3, the climograph of Ramsar (marked with a red arrow) displays significant precipitation variability compared to temperature fluctuations. In contrast, other stations show less variation in precipitation. The other climagraphs (except Ramsar) can be divided into two groups, which were shown in Figure 3 with zoom (showing by two blue circles with dash and round dot lines) with the minimum variation in precipitation of Yazd and Semnan stations. Some of the data used are the minimum temperature, maximum temperature, relative humidity, wind speed, and sunshine hours on a monthly time scale.

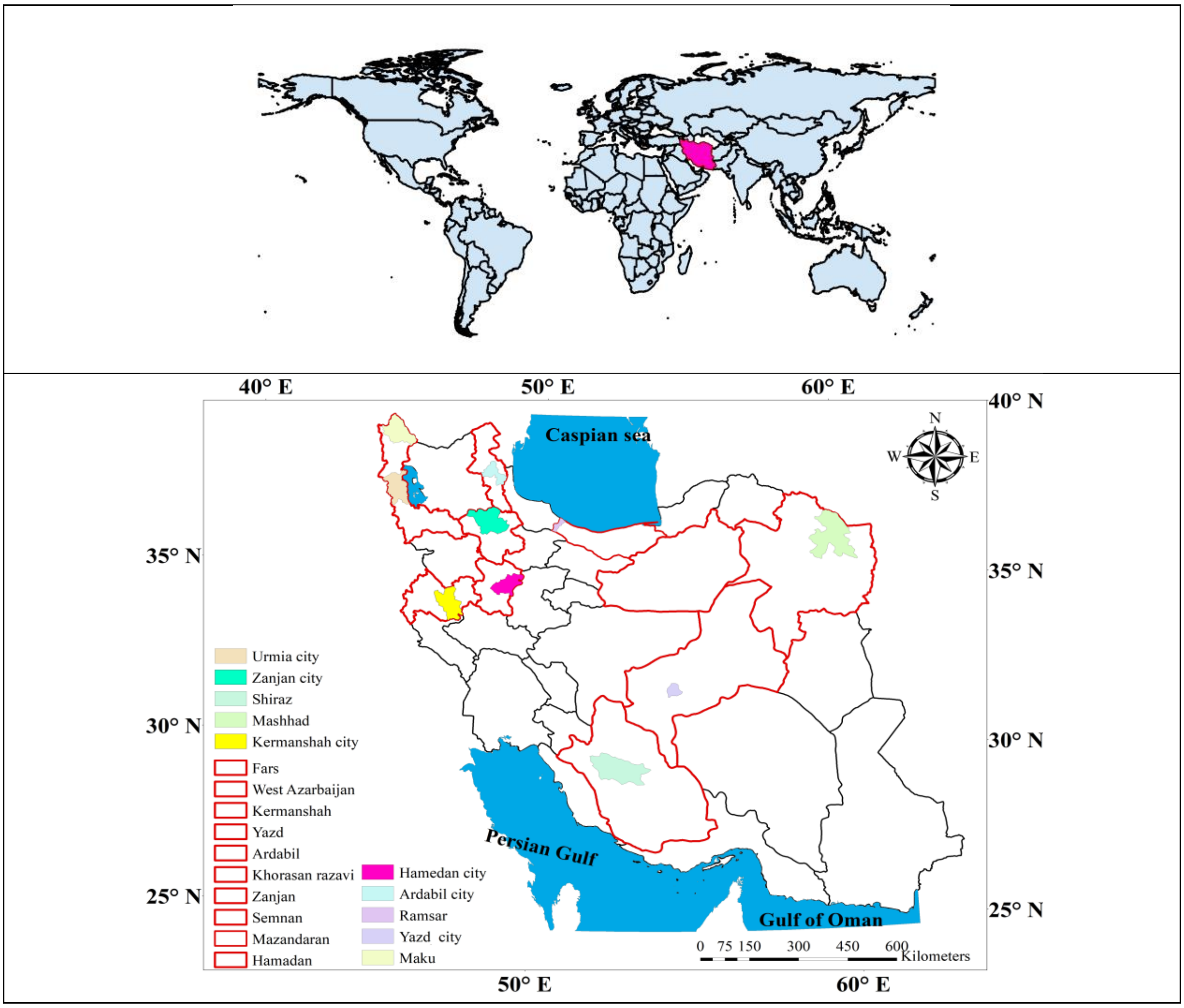


Figure 1. The location of the studied stations in each province in the map of Iran

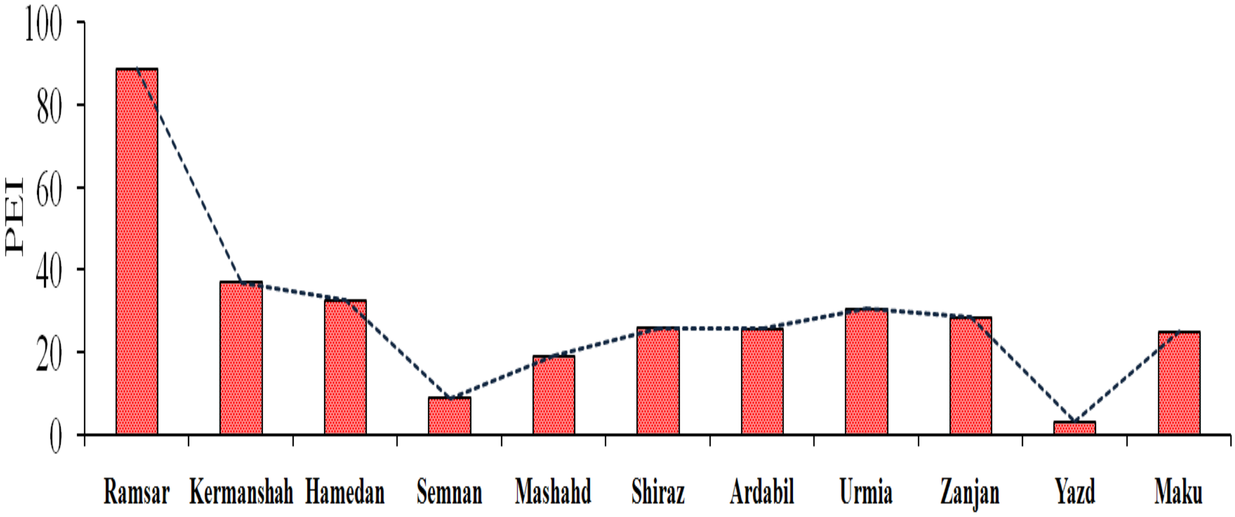


Figure 2. The climate of the studied stations with the PEI climate classification method

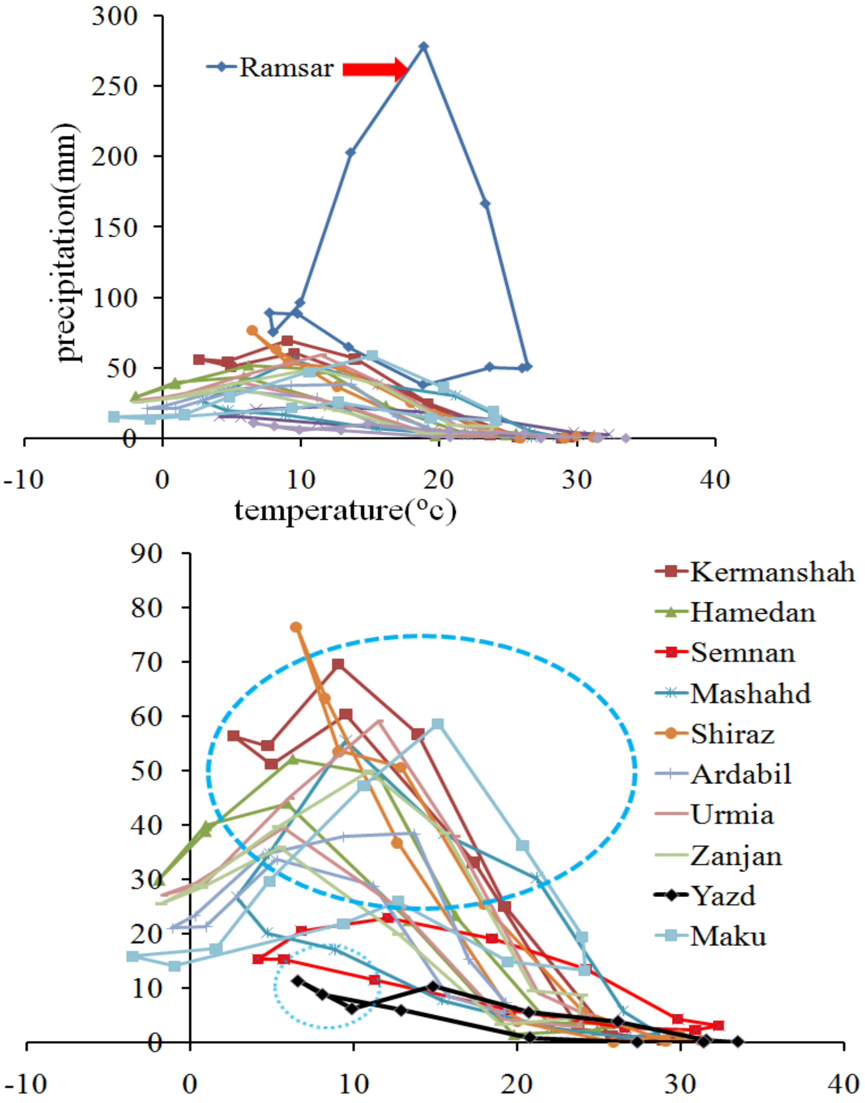


Figure 3. The climograph of stations with zoom on the climographs of all stations except Ramsar to show clearly the temperature and precipitation variations

2.2. Preprocessing methods

The preprocessing step in the modeling process has great importance in deriving the effective and precise factors from the input data.

Several preprocessing methods were investigated in this study to identify the dominant input data for ET_0 estimation. There are different methods for determining ET0, each of which requires different data and methods. In 1948, Penman proposed a formula for ET_0, which was later used and modified by a large number of experts, and the structure of the FAO Penman-Monteith equation is illustrated in equation 1.

$$ET_0 = \frac{0.408\Delta\,(R_n - G) + \gamma\left[\frac{890}{T+273}\right]U_2(e_a - e_d)}{\Delta + \gamma(1 + 0.34U_2)} \quad (1)$$

Where, R_n is the net radiation, G is the soil heat flux, e_a-e_d is the vapour pressure deficit of the air, U2 is the wind speed at 2m, Δ is the slope of saturation vapour pressure temperature relationship, γ and is the psychrometric constant (Shu et al., 2022).

The preprocessing methods include the Pearson correlation coefficient, Kendall's tau-b correlation coefficient, the standardized Beta coefficient, stepwise regression, Shannon's entropy, and simple additive weighting with fuzzy normalization. These methods were selected for their ability to assess important variables with data analysis from different aspects by correlation detection and data normalization, ensuring accurate ET_0 estimation.

2.2.1. Pearson correlation coefficient

The coefficient can distinguish the correlation between independent and dependent variables; higher values indicate higher dependency. The range of coefficients is from -1 to 1, where values close to +1 or -1 indicate strong positive or negative linear relationships, respectively. The correlation coefficient can be defined as equation 2, which the independent variable can explain the changes of the dependent variable (R^2), with the coefficient increasing above 0.6. The Pearson correlation coefficient is brought in equation 3, which is greater than 0.8 is indicating of high correlation.

$$R^2 = \left(\frac{\sum_{i=1}^{N}(O_i-\overline{O})(F_i-\overline{F})}{\sqrt{\sum_{i=1}^{N}(O_i-\overline{O})^2 \times \sum_{i=1}^{N}(F_i-\overline{F})^2}}\right)^2 \qquad (2)$$

$$Corr(F,O) = \frac{Cov(F,O)}{\sigma_F \, \sigma_O} \qquad (3)$$

Where, O is the observed values, F is the forecasted value, $\overline{O}$, $\overline{F}$ are the average of observed and forecasted values, N is the number of data, Cov(O, F) is the covariance between O and F, σ and is the standard deviation (Ahmadpari and Khaustov 2025).

2.2.2. Kendall's tau-b correlation coefficient

The coefficient is another nonparametric statistic that determines the degree of dependency. The efficiency of the coefficient is for data samples with tied ranks. The range of the coefficient is between -1 and 1, and the absolute values of the coefficient can determine the monotonic relation of two variables (Raffinetti and Aimar, 2019).

2.2.3. Standardized Beta coefficient

One of the important parameters in the regression analysis is the standardized Beta coefficient that measures the effect of each independent variable on the dependent variable. The high absolute value of the standardized Beta coefficient is indicative of powerful effectiveness (Nieminen 2022).

2.2.4. Stepwise regression

The emphasis of stepwise regression is on the best and most impressive variables from a large set of variables (Malek et al., 2007). Indeed, the explanatory variables were chosen for a multiple regression model from a group of candidate variables by going through a series of automated steps. At each step, the candidate variables are investigated, one by one, using the t-statistics. The rule of forward selection starts with no explanatory variables, which then adds variables one by one, and the variables with high significance are retained. The stop criterion is the absence of statistically significant variables. The backward elimination rule is based on stopping the discarding when each variable of the equation is statistically significant (Smith, 2018).

The statistical data analysis of the Pearson and Kendall's tau-b correlation coefficient, the standardized Beta coefficient, and stepwise regression are related to the correlation coefficient between ET_0 and meteorological data.

2.2.5. Shannon's entropy

Decision-making is not always between two options, and sometimes we have to make the right selection among several options. In this case, a multi-criteria decision is made, depending on the sensitivity of the problem, for which certain methods can help to reach the best option. Therefore, multi-criteria decision making (MCDM) analysis has many usages in various sciences, such as Dooley et al. (2009) in agriculture, Butchart-Kuhlmann et al. (2018) in hydrology, Jiang et al. (2021) in environmental studies, and management of energy (Musbah et al., 2022). The MCDM by powerful strategies can facilitate decision-making in management problems and operational analysis (Dwivedi and Sharma, 2022a). Some methods are illustrated to solve MCDM problems, such as Shannon's entropy. Shannon (1948) proposed the entropy method. The process of entropy analysis is to assign the weights of the objective criterion (Dwivedi and Sharma, 2022a, b). The assumption of entropy analysis is the importance of data with high-weight indicators relative to the data with low-weight indicators (Dwivedi and Sharma, 2022 b). The Shannon entropy can be defined as the mean value of information that the source transmits by each symbol. The transmission uncertainty can be proportional to the expected value of the probability logarithm of receiving a symbol sequence (Shternshis et al., 2022).

The first step of the entropy method is to construct a decision matrix(DM) which includes the criteria and decision options (Eq.4), where x_{ij}

defines the performance of the *i*th alternative (number of alternatives) and *j*th criteria (number of criteria). Decision matrix normalization is the second step of entropy analysis, and Eq.5 can normalize the DM. The entropy values can be calculated using Eq.6 in the third step of entropy analysis. The final step is the weights, w_j, and determination related to each criterion (Eq.7).

$$DM = \left[X_{ij}\right]_{m\times n} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots\cdot & x_{2n} \\ . & . & \cdots & . \\ . & . & \cdots & . \\ . & . & \cdots\cdot & . \\ x_{m1} & x_{m2} & \cdots\cdot & x_{mn} \end{bmatrix} \tag{4}$$

$$f_{ij} = \frac{x_{ij}}{\sum_{i=1}^{m} x_{ij}} \tag{5}$$

$$E_j = -\frac{\sum_{i=1}^{m} f_{ij} \ \log f_{ij}}{\log m} \tag{6}$$

$$w_j = -\frac{1-E_j}{\sum_{i=1}^{n}(1-E_j)} \tag{7}$$

Where E_j is the value of entropy, and log m is the entropy coefficient (Dwivedi and Sharma, 2022a).

2.2.6. Fuzzy simple additive weighting

The first step of this method is to number a decision matrix. The second step tries to normalize the values of the decision matrix, for which, in this study, fuzzy normalization was proposed. In this method, the elements of the decision matrix are different for positive and negative criteria, and Eqs. 8 and 9 are for positive and negative criteria, respectively (Dwivedi and Sharma, 2002a).

$$n_{ij} = \frac{a_{ij} - Min\, a_j}{Max\, a_j - Min\, a_j} \tag{8}$$

$$n_{ij} = \frac{Max\, a_j - a_j}{Max\, a_j - Min\, a_j} \tag{9}$$

The next step after DM construction and normalization is multiplying the normalized matrix by the matrix of weights. At the end step, the best result can be obtained according to relation 10.

$$A^* = \left\{ A_i \middle| Max \sum_{j=1}^{n} n_{ij}\, w_j \right\} \tag{10}$$

Where n is the normalized matrix, and w is the weight matrix.

Different combinations of data were defined to implement the preprocessing methods.

The evaluation criteria of decision making analysis methods, Shannon׳s entropy and fuzzy simple additive weighting, in this study are related to the root mean square error (RMSE), mean absolute error (MAE), correlation coefficient (R), Nash-Sutcliffe efficiency coefficient (NSE), and geometric mean error ratio (GMER), which their structures are in Eqs.11-14.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i-1}^{N}(F_i - O_i)^2} \tag{11}$$

$$MAE = \frac{1}{N}\sum_{i-1}^{N}\left|O_i - F_i\right| \tag{12}$$

$$GMER = \exp\left(\frac{1}{n}\sum_{i=1}^{n}\ln\left(\frac{F_i}{O_i}\right)\right) \tag{13}$$

$$NSE = 1 - \frac{\sum_{i=1}^{n}(F_i - O_i)^2}{\sum_{i=1}^{n}(O_i - \overline{O})^2} \tag{14}$$

Where F*ᵢ* is the forecasted value, O_i is the observed value, $\overline{O}$ is the mean of the observed value (Abdelbaki, 2016; Jamil and Bellos, 2019). The effective data determination on the ET_0 process using the mentioned methods should be done in the context of artificial intelligence models; therefore, support vector regression (SVR) was used in this research, due to its high capability in forecasting issues.

2.3. Support vector regression (SVR)

The regression analysis aims to minimize the error between observed and forecasted values; this matter can be possible by SVR with the estimation foremost value in a certain threshold, such as the distance of the boundary line and hyperplane. The structure of SVR and the definition of the kernel function in the SVR process allow it to upgrade the values of error tolerance (Rezaei et al., 2023). SVR tries to decrease the operational risk with reorganization and increase the margins of data (Cai et al., 2023). Some of the SVR strengths are: the independence of the computational complexity of SVR on the scale of the input space, SVR preferential performance for low sample size with a high feature space dimension, the high generalizability with high predictive accuracy, and the foundation matter is the flexibility of SVR in determination the values of acceptable error of model to know a proper line (Rezaei et al., 2023). In the SVR structure, ***w*** and *b* are the coefficients of the modeling between the mapping function and the dependent variable, and in one kind of SVR, *ε*-SVR, can be obtained from the optimization problem (Eq. 15).

$$\min_{w,b} C\sum_{i=1}^{N} l(y_i - w^T\phi(x_i) - b) + \frac{1}{2}\|w\|^2 \qquad (15)$$

Where $\phi(\boldsymbol{x})$ is the mapping function, $l(.)$ is a loss function.

The Lagrange function $L(w,b,\xi_i,\xi_i^*,\alpha_i,\alpha_i^*,\mu_i,\mu_i^*)$ with a soft margin loss and five slack variables definition can be made from an optimization problem (objective function and constraint or constraints), which finally ends in a dual optimization problem (Eq.16).

$$\min_{w,b} C\sum_{i=1}^{N} l(y_i - w^T\phi(x_i) - b) + \frac{1}{2}\|w\|^2 \qquad (16)$$

$$\max_{\alpha_i\alpha_i^*} -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}(\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j)k(x_i,x_j) + \sum_{i=1}^{N}\left[y_i(\alpha_i^* - \alpha_i) - \varepsilon(\alpha_i^* + \alpha_i)\right]$$

$$s.t. \quad \sum_{j=1}^{N}(\alpha_i^* - \alpha_i) = 0$$

$$0 \le \alpha_i^*, \alpha_i \le C$$

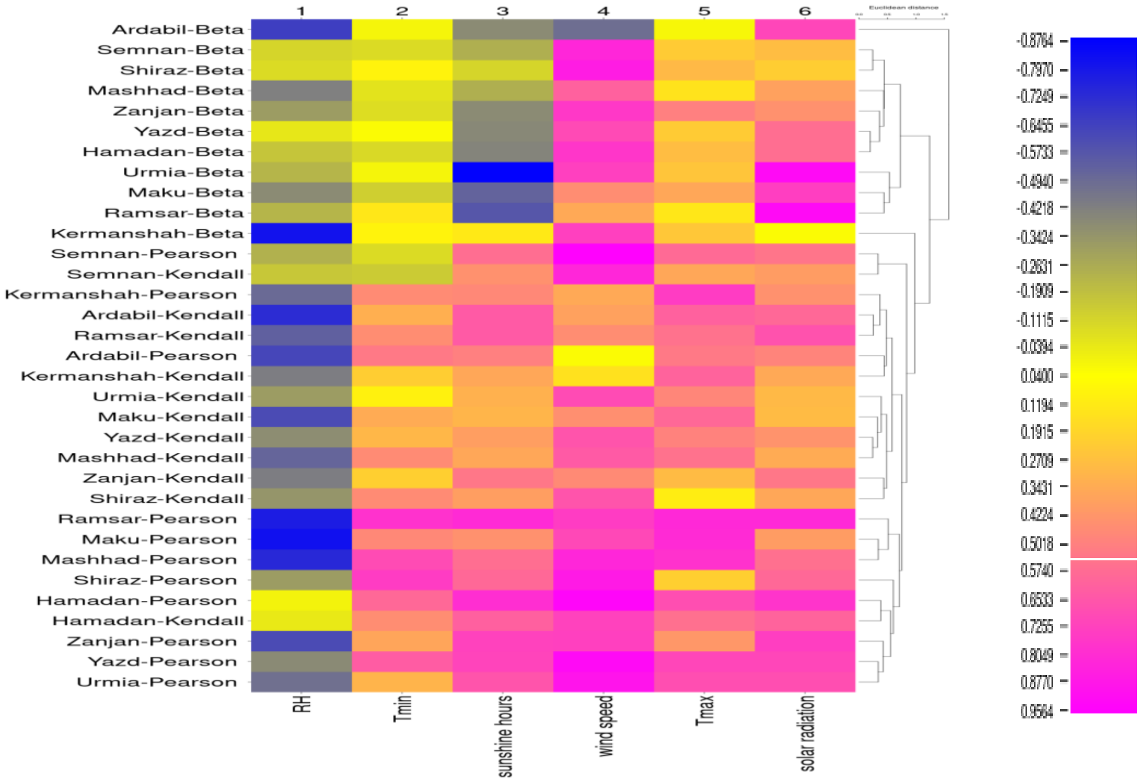
Where k is the kernel function (Haoyuan et al., 2023).

3. Results

The selection of effective meteorological data on the ET$_0$ simulation was focused on the minimum temperature (T$_{min}$), maximum temperature (T$_{max}$), relative humidity (RH), wind speed (U), sunshine hours(S), and solar radiation (R). The considered time scale is annual and monthly. The period division is based on 80% and 20%, wherein 1992-2015 and 2016-2021 are calibration and verification periods, respectively. FAO Penman-Monteith was employed as the reference method for ET$_0$ comparison. Some preprocessing methods based on the statistical approach were applied to measure the dependency of data and ET$_0$, and the results are displayed in Figure 4.

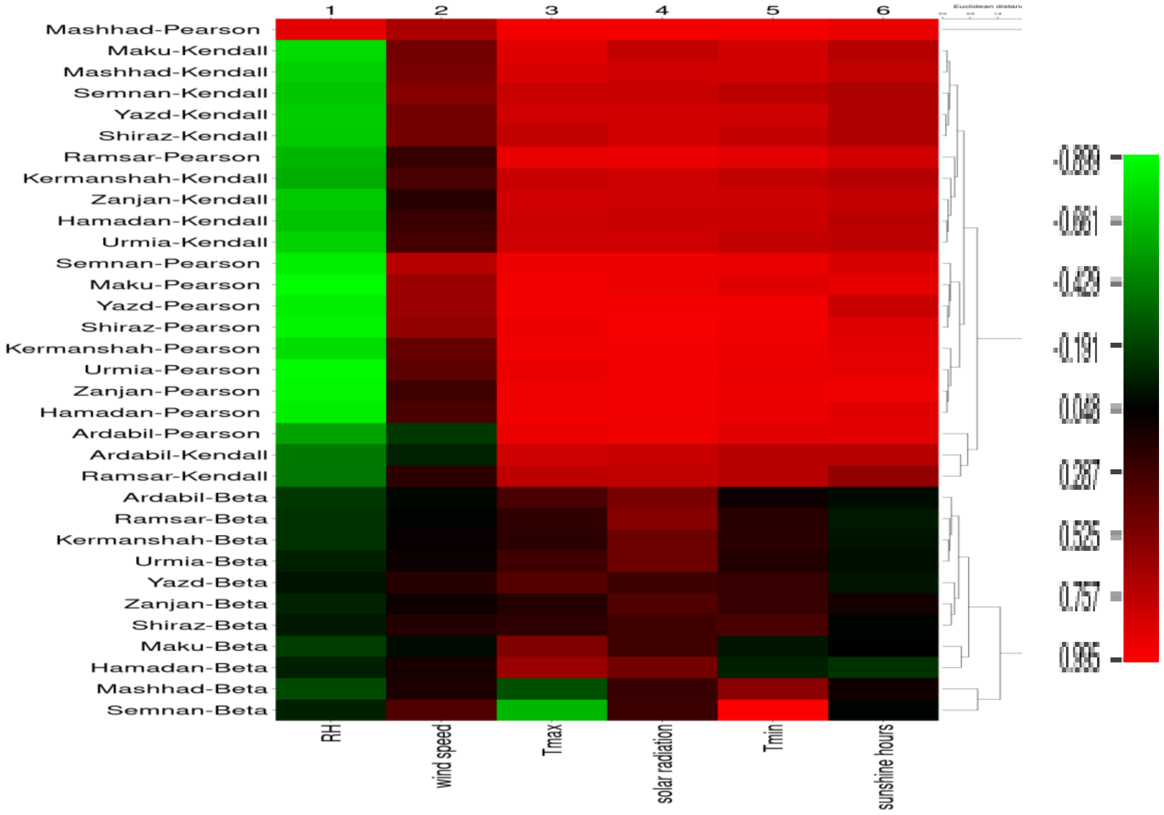
Fig. 4 applies a heat map, which is a two-dimensional representation of data with a system of color-coding. In this map, the Pearson, τ Kendall correlation coefficients and standardized Beta coefficient values for different stations are represented by colors. The maximum Pearson and τ Kendall correlation coefficients in the annual scale is related to the maximum temperature, wind speed, solar radiation, maximum temperature, wind speed, wind speed, maximum temperature, wind speed, wind speed, wind speed and maximum temperature in the Maku, Yazd, Zanjan, Ardabil, Mashhad, Semnan, Kermanshah, Shiraz, Hamadan, Urmia and Ramsar, respectively. Generally, the pink cell, which shows high correlation in Fig. 4.a, is low for relative humidity.

The maximum Pearson correlation coefficient in the monthly scale is related to the solar radiation, maximum, and minimum temperature in all stations. This matter was preserved by τ Kendall correlation coefficient. The green cell, which shows the low correlation in Fig. 4b, is high for relative humidity. The maximum Beta coefficient in the annual scale is solar radiation, wind speed, wind speed, solar radiation, wind speed, wind speed, relative humidity, wind speed, wind speed, solar radiation, and solar radiation in the Maku, Yazd, Zanjan, Ardabil, Mashhad, Semnan, Kermanshah, Shiraz, Hamadan, Urmia, and Ramsar, respectively. The maximum Beta coefficient in the monthly scale in the Maku, Yazd, Zanjan, Ardabil, Mashhad, Semnan, Kermanshah, Shiraz, Hamadan, Urmia, and Ramsar is the maximum temperature, maximum temperature, solar radiation, solar radiation, minimum temperature, minimum temperature, solar radiation, solar radiation, maximum temperature, solar radiation, and solar radiation, respectively.



(a)

Figure 4. Heat map of Pearson, τ, Kendall correlation coefficients, and standardized Beta coefficient in annual, a, and monthly, b, time scales for all stations.



(b)

Figure 4. cont. Heat map of Pearson, τ, Kendall correlation coefficients, and standardized Beta coefficient in annual, a, and monthly, b, time scales for all stations.

The derived meteorological data in the stepwise regression at the annual scale can be described as the relative humidity, wind speed, solar radiation, maximum temperature in Maku, wind speed, maximum temperature, solar radiation, sunshine hours in Yazd, solar radiation, wind speed, maximum temperature, relative humidity, sunshine hours in Zanjan, relative humidity, wind speed in Ardabil, wind speed, relative humidity, solar radiation in Mashhad, wind speed, maximum temperature, relative humidity in Semnan, maximum temperature, relative humidity, wind speed, solar radiation in Kermanshah, wind speed, maximum temperature, solar radiation in Shiraz, wind speed, sunshine hours, relative humidity, maximum temperature in Hamadan, wind speed, maximum temperature, relative humidity in Urmia, relative humidity, solar radiation, wind speed, minimum temperature, sunshine hours in Ramsar. The derived data in the monthly scale consist of the maximum temperature, solar radiation, relative humidity in Maku, minimum temperature, wind speed, solar radiation, maximum temperature in Yazd, solar radiation, maximum temperature, relative humidity, wind speed, minimum temperature, sunshine hours in Zanjan, solar radiation, maximum temperature, relative humidity, minimum temperature, wind speed in Ardabil, solar radiation, maximum temperature, wind speed, relative humidity, minimum temperature, sunshine hours in Mashhad, solar radiation, wind speed, minimum temperature, maximum temperature, relative humidity in Semnan, solar radiation, maximum temperature, wind speed, relative humidity, minimum temperature in Kermanshah, solar radiation, minimum temperature, relative humidity, maximum temperature, relative humidity in Shiraz, solar radiation, maximum temperature, wind speed, sunshine hours, relative humidity in Hamadan, solar radiation, maximum temperature, wind speed, minimum temperature, relative humidity in Urmia, solar radiation, maximum temperature, relative humidity, minimum temperature, wind speed, sunshine hours in Ramsar. Different combinations of data were defined to implement the MCDM methods (Table 1).

Table 1. Different combinations of meteorological data to model ET_0

Number of input data 5; **first group**	case 1	T_{max}	T_{min}	RH	U	S
	case 2	T_{max}	T_{min}	RH	U	R
	case 3	T_{max}	RH	U	R	S
	case 4	T_{min}	RH	U	R	S
Number of input data 4; **second group**	case 1	T_{max}	T_{min}	RH	U	
	case 2	T_{max}	T_{min}	RH	S	
	case 3	T_{max}	T_{min}	RH	R	
	case 4	T_{max}	T_{min}	U	R	
	case 5	T_{max}	T_{min}	R	S	
	case 6	T_{max}	T_{min}	U	R	
	case 7	RH	U	S	R	
Number of input data 3; **third group**	case 1	T_{max}	T_{min}	RH		
	case 2	T_{max}	T_{min}	U		
	case 3	T_{max}	T_{min}	S		
	case 4	T_{max}	T_{min}	R		
	case 5	RH	U	S		
	case 6	RH	U	R		
	case 7	U	R	S		

Decision-making analysis needs some criteria, and five criteria, RMSE, R, MAE, NSE, and GMER, were applied in Shannon’s entropy method. The selected criteria of the method include negative criteria, such as RMSE, MAE, GMER, and positive criteria, such as R and NSE, which are used to find the best solution from all data (T_{max}, T_{min}, RH, U, S, and R), and different combinations of data (Table 1). For example, the information about the value of weights in Shannon’s entropy is shown in Figure 5.

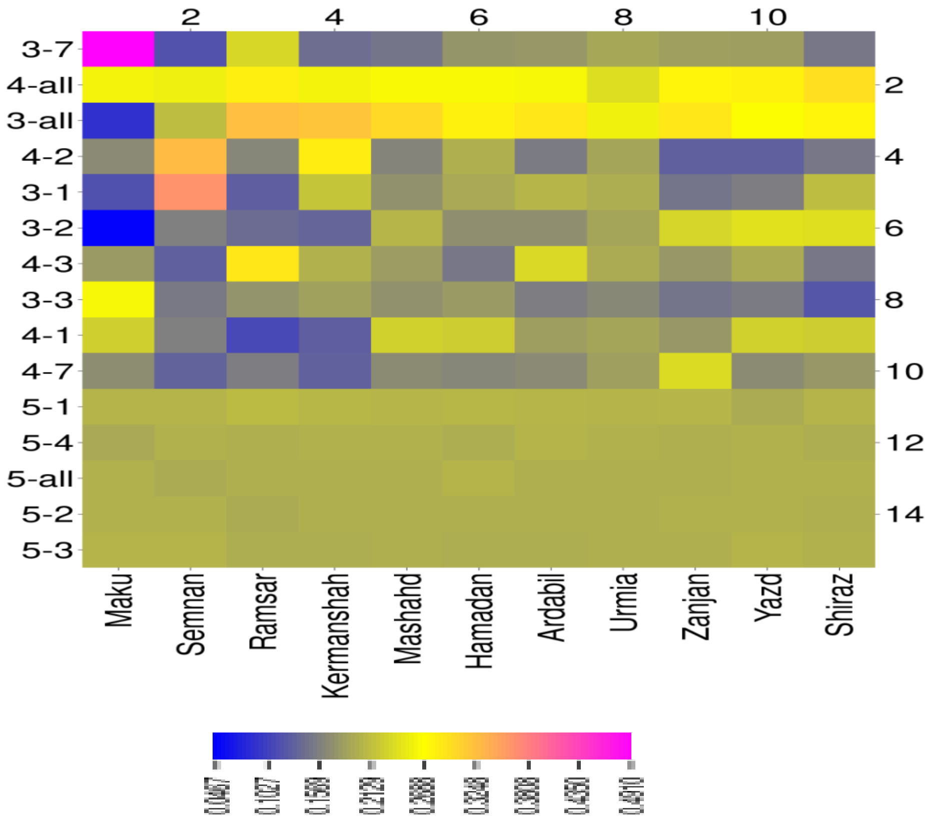


Figure 5. Heat map of weights obtained from Shannon᾽s entropy in the monthly scale for all stations

In Fig. 5, the color distribution is the same as the data combination by five numbers, but the data combinations with four and three data have color diversity. The combination 3-7; the number of input data is equal to 3, and the data are wind speed, solar radiation, and sunshine hours, has the highest weight, pink in Maku. The variety of colors is seen is all stations, but it is high in Maku, Semnan, Ramsar, Kermanshah, and Mashhad. The proposed method in the study to compare the performance of Shannon᾽s entropy and fuzzy normalization is the error criteria computation of groups with high weight in each station, and it is shown in Figure 6.

In the monthly scale and the combination with five input data, the RMSE of all stations related to Shannon᾽s entropy is higher than fuzzy normalization, except Mashhad, with the same RMSE in the two methods, and Zanjan and Yazd with a low error of Shannon᾽s entropy. In the combination with four input data, the RMSE of Shiraz and Urmia is the same in the two methods, and at the other stations, the RMSE of Shannon's entropy is greater than the RMSE of fuzzy normalization. In the combination with three input data, the RMSE of the two methods in Maku, Mashahd, and Kermanshah is the same, and in the other stations except Urima, the RMSE of Shannon᾽s entropy is high. In the annual scale and the combination with five input data, the MAE of Mashhad is the same in two methods, the MAE of Shannon᾽s entropy in Maku, Ardabil, Semnan, Kermanshah, Shiraz, Hamadan, Urmia and Ramsar is greater than fuzzy normalization, and in the Yazd and Zanjan, the MAE of Shannon᾽s entropy is lower than fuzzy normalization. In combination with four input data, except Shiraz and Urmia, with the same MAE of the two methods, the MAE of Shannon᾽s entropy is higher than fuzzy normalization. In the combination with three input data, the MAE of the two methods in Maku, Mashhad, and Kermanshah are the same, the MAE of Shannon᾽s entropy in Urmia is lower than fuzzy normalization, and in the other stations, fuzzy normalization has better performance. In the other case, the average of evaluation criteria was calculated for all stations to select the precise combination (Figures 7 and 8).

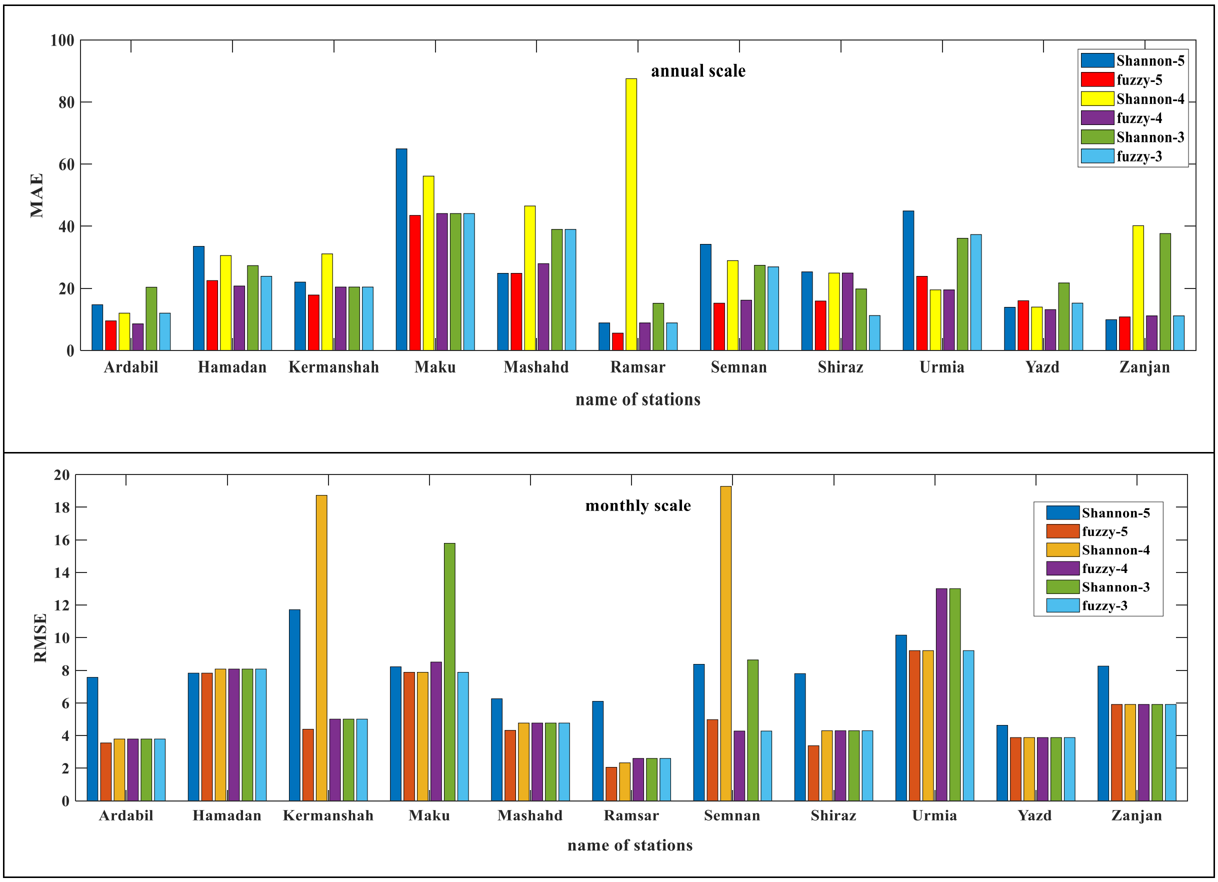


Figure 6. Compassion, the efficiency of Shannon᾽s entropy and fuzzy normalization methods, with a high weight of each group in all stations

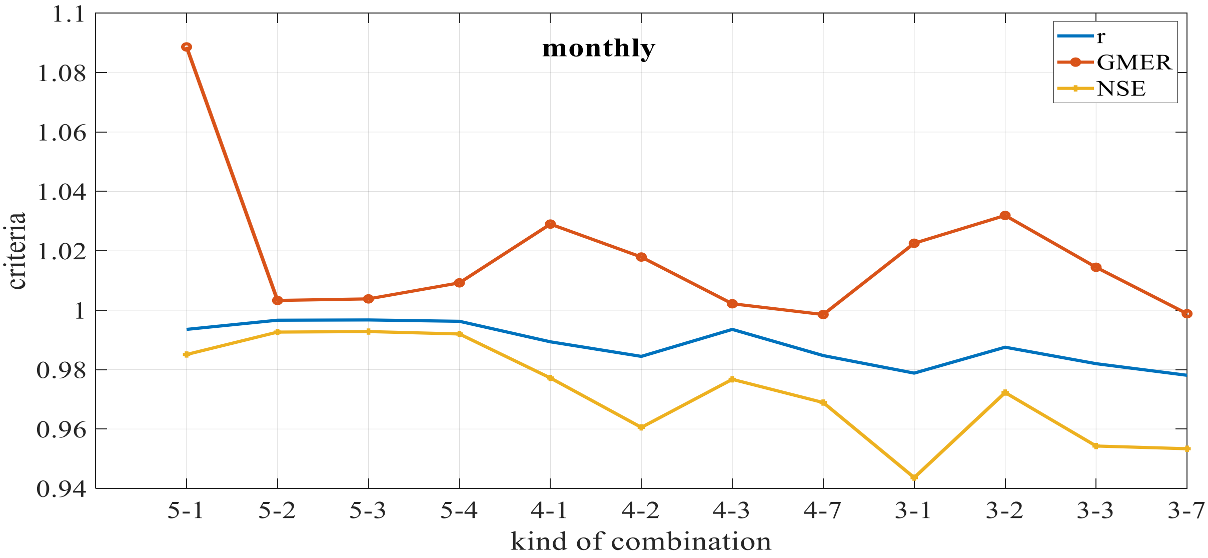


Figure 7. GMER, NSE, and R as the evaluation criteria average for all combinations and all stations in the annual and monthly scale.

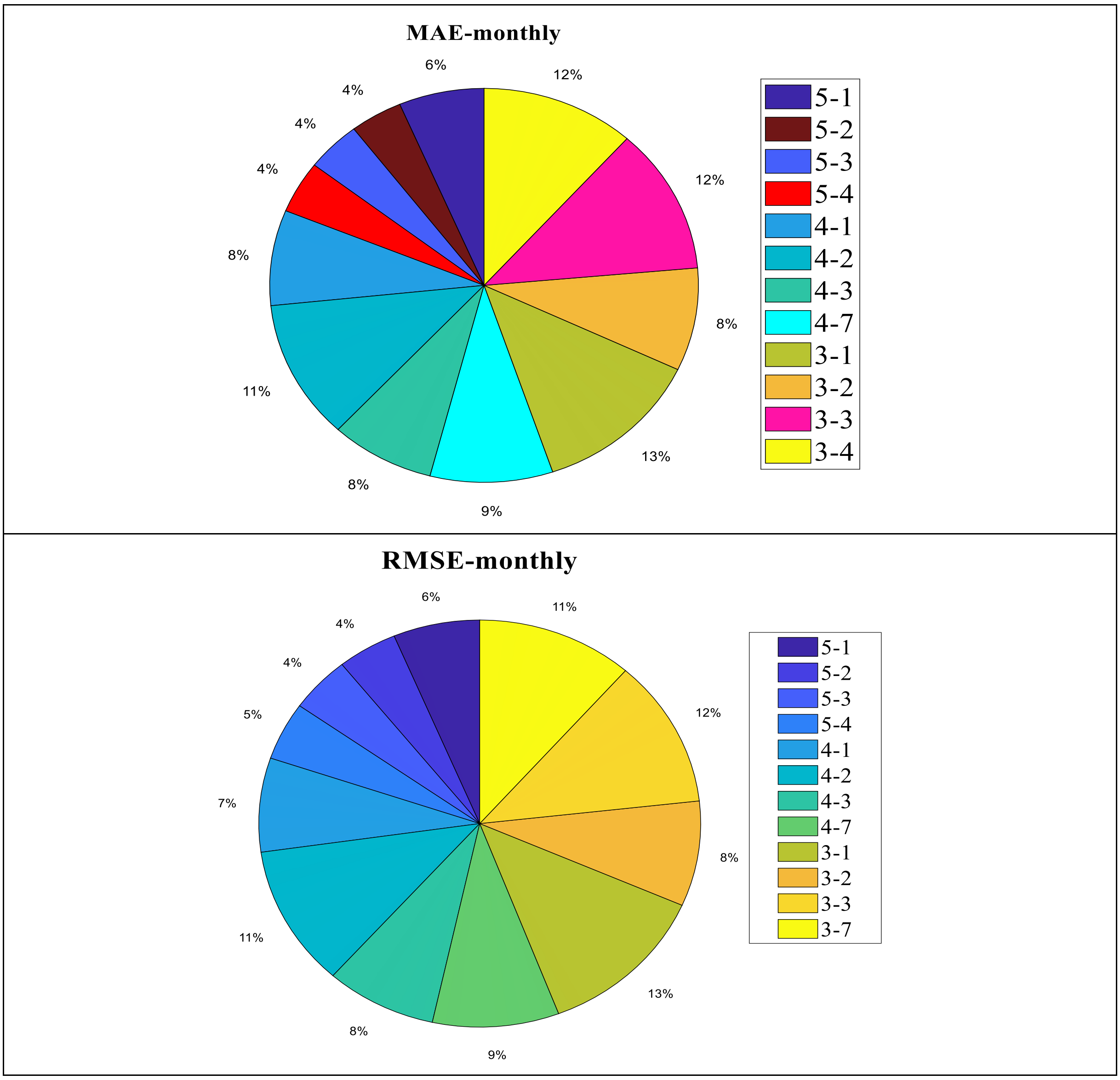


Figure 8. Evaluation criteria average, RMSE, and MAE, for all combinations and all stations in two scales: monthly and annual

In the annual scale, the minimum values of RMSE, MAE are related to the 5-3 combination; the maximum values of R and NSE are related to the 5-1 and 5-3 combinations, respectively. GMER has the minimum values in the 3-2 combination. In the monthly scale, the 5-2 combination has the minimum values of RMSE, MAE, and maximum values of R. The minimum values of GMER are related to the 3-7 combination. If the average of the criteria is regarded in each group of Table 1, the RMSE decreases in the annual scale from three to four and five input data, is 10.01% and 48.73%, respectively. The NSE increases on a monthly scale from three to four and five input data are 2.06% and 3.66%, respectively. The combination of five input data has the best performance. The next part tries to compare the performance of preprocessing methods (Figure 9). The differences between Pearson and τ Kendall correlation coefficients are low; therefore, only the Pearson coefficient was used in this part. The comparison was conducted on the average of the evaluation criteria in all situations. In the fuzzy normalization, the selected combination was based on the maximum weight of the combination.

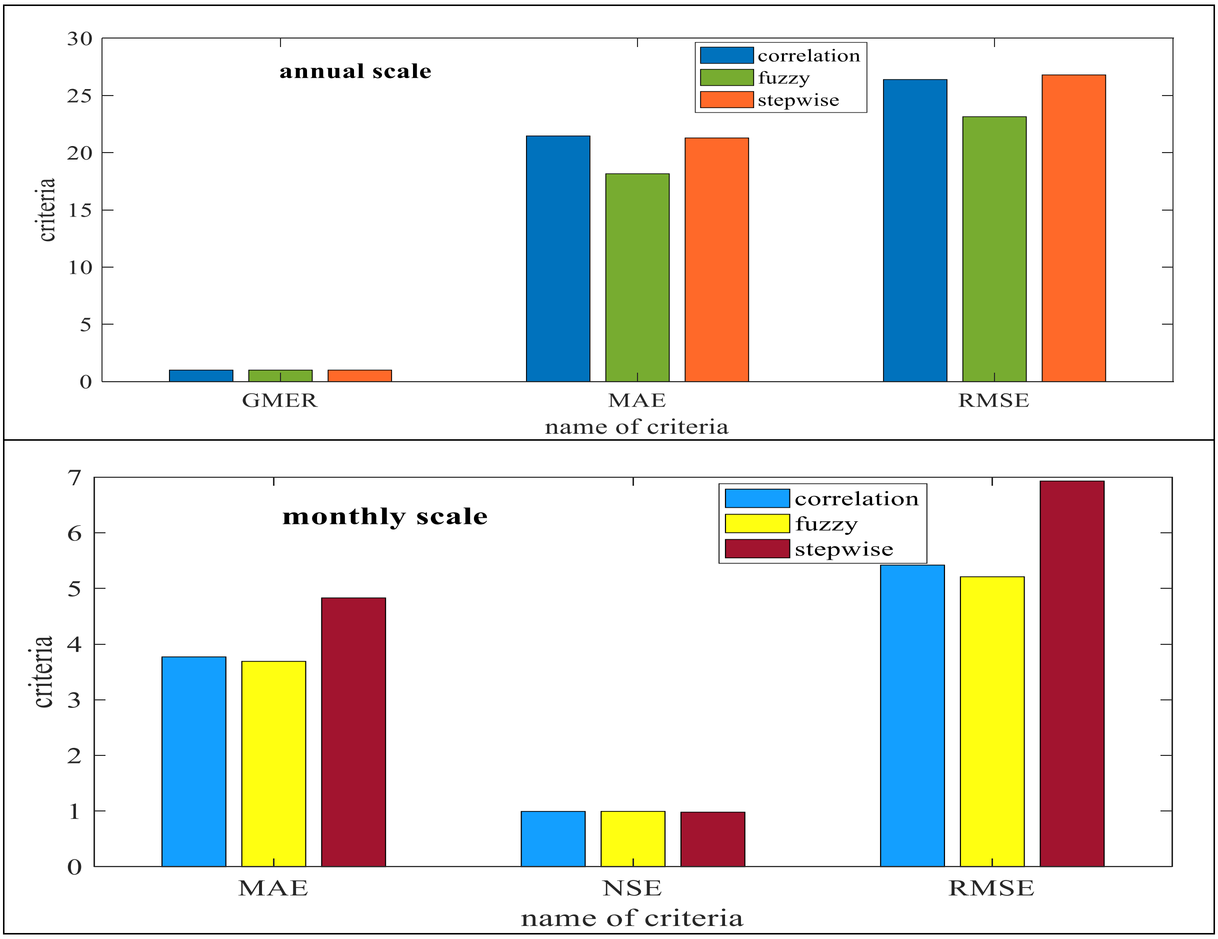


Figure 9. Comparison of the efficiency of preprocessing methods in two scales: annual and monthly

In two scales, the performance of fuzzy normalization is in a good state. In the annual scale, the Pearson correlation and stepwise regression have the same function. In the monthly scale, stepwise regression has poor performance. The selection of input data based on fuzzy normalization could decrease the error of the simulation.

4. Discussion

Some elements can affect the ET_0 modeling, and in the meantime, the type of data considered has high importance. To prevent an increase in the operational volume, the account of sufficient and effective data can increase the efficiency of the ET_0 simulation. Therefore, some concepts as the preprocessing methods, Pearson and Kendall's tau-b correlation coefficients, the standardized Beta coefficient, stepwise regression, Shannon's entropy, and simple additive weighting with fuzzy normalization, were tested in this study to select the effective data. The basis of the correlation coefficient method is measuring the correlation between ET_0 and meteorological data. In the monthly scale, there were no differences between the numbers of data with significant coefficients between the Kendall and Pearson correlation coefficient methods. In the stepwise regression, the number of selected data points in each station is different; for example, the selected data points in Ardabil are two, and it is five in Zanjan for the annual scale. In some stations, all of the considered data are selected, such as in Mashhad and Shiraz, for the monthly scale.

The comparison of preprocessing methods, Pearson-fuzzy normalization-stepwise regression, indicates that from stepwise regression and Pearson to fuzzy methods in the annual scale, RMSE decreased by 13.62% and 12.13% respectively, and MAE decreased by 14.7% and 15.42%, respectively. The evaluation of Shannon's and fuzzy normalization showed that in most stations fuzzy normalization method has better performance. All of the methods use a mathematical structure to find the dependency of ET_0 on the input data, but the kind of structure and how they work make a difference in the obtained results and the performance of the methods. In the decision-making analysis, the procedure of assigning weight to the criteria has an important role. Ahmadi et al. (2021) proved the significance of the preprocessing method in the ET_0 mechanism investigation.

Fuzzy performance investigation among all stations stated that Zanjan, Ardabil, and Urmia in annual scale and Semnan, Shiraz, and Ramsar in monthly scale have better performance relative to the other stations in terms of evaluation criteria.

In order to investigate the type of climate effect, humid or arid, the average of NSE and GMER was calculated in each climate: Ramsar in the wet group, Maku, Zanjan, Ardabil, Mashhad, Kermanshah, Shiraz, Hamadan, Urmia in semi-arid, Yazd and Semnan in arid climate. The maximum NSE in the annual scale belongs to a semiarid climate. The NSE among the three kinds of climate on a monthly scale is the same and rounded to 0.99. The GMER among different climates is almost the same and rounded to 1. Therefore, there is no significant difference in the criteria of various climates in all stations. This issue can be of great help in meteorological studies.

The different data combinations have several reactions on ET_0 estimation, for example in Maku, Ramsar and in annual scale, all of combination with four input data had in a worse situation relative to using all data (the RMSE decreasing from 4-1,4-2,4-3,4-4,4-5,4-6,4-7 combinations to using all data is 22.12%, 41.15%, 39.79%, 7.4%, 35.32%, 14.91% and 29.46% in Maku. The MAE decreasing from 4-5,4-6,4-7 combinations to using all data is 61.8%, 51.23%, and 32.64% in Ramsar.) Some combinations can strengthen the ET_0 estimation on the annual scale. For example, NSE increasing from all input data to 5-3 combination in Urmia, Ramsar, and Semnan is 12.34%, 8.1%, and 40.65%, respectively. The variation of evaluation criteria in the monthly scale also verified the previous item, which RMSE decreased 75%, 52.59%, 65.18%, 67.7% from 3-1, 3-2, 3-3, 3-7 to 5-2 combinations in Ardabil. NSE increased from 4-2 to 5-2 and from 4-2 to 5-4 combination in Maku and Zanjan, respectively. The combination with the minimum values of MAE and RMSE in Maku, Yazd, Zanjan, Ardabil, Mashhahd, Semnan, Kermanshah, Shiraz, Hamadan1, Urmia, Ramsar in annual scale is 5-2, 4-7, 5-2, 4-6, 5-3, 5-4, 5-2, 4-4, 4-1, 5-3, 3-1, respectively. The optimum combination in a monthly scale is 5-2, 4-1, 5-4, 5-3, 5-2, 5-2, 5-3, 5-4, 5-1, 5-3, 5-2. In the monthly and annual scale, the 5-2 combination has high frequency with maximum and minimum temperature, relative humidity, wind speed, and solar radiation input data. In the study of Tabari and Talaee (2013), mean temperature, relative humidity, and wind speed were regarded as the data with high performance. Yadeta et al. (2020) expressed that solar radiation and maximum temperature are the remarkable data of the study. The maximum values of RMSE and MAE was also investigated, which the combination in annual sale, in Maku, Yazd, Zanjan, Ardabil, Mashhahd, Semnan, Kermanshah, Shiraz, Hamadan1, Urmia, Ramsar is 3-2, 4-3, 3-4, 3-7, 4-3, 4-5, 3-7, 4-5, 3-1, 3-3, 3-3, respectively. The minimum and maximum values of the error criteria in the monthly scale are focused on different combinations of five and three input data. The diversity of evaluation criteria among all stations showed that in the annual and monthly scales, the minimum RMSE, MAE, and maximum NSE, and R are related to the 5-3 and 5-2 combinations, respectively.

5. Conclusion

One of the requirements in ET_0 modeling is the application of an accurate and effective amount of meteorological data. Therefore, implementation of that before the modeling process is necessary, which in this study, two approaches based on statistical and multi-criteria decision making were applied as the preprocessing methods.

Some findings of the study are given below:

-In the monthly scale, the combination of five and three input data had the best and worst performance, respectively. The different combinations with various numbers of data had different impacts on the ET_0 modeling. In addition to the kind of input data, the number of data points has high importance in stating clearly the mechanism of ET_0. However, the diversity of climate led to finding different combinations in each station with high performance, and it is indicative of each combination's potential in capturing the governing mechanism of the ET_0 in nature.

-The results indicated that the normalization process had better performance in the preprocessing method based on the MCDM approach relative to the other methods. The average of the criteria showed that the best method has no limitations regarding the three types of different climates, wet, semiarid, and arid, and the fuzzy normalization had good performance. This method has no geographical limitation. Determining an efficient method for the preprocessing step has an acceptable response in all climates, which is one of the strengths and innovations of the research.

- One of the things that can strongly affect the preprocessing method-based MCDM approach is the type of decision-making method. In the decision-making problem, the method used for normalization of the decision matrix has high importance in information extraction.

-In general, maximum temperature, relative humidity, wind speed, solar radiation, sunshine hours (annual), and minimum temperature (monthly) were introduced as the effective data. The reason for the better performance of certain data combinations is related to the high dependency of these combinations on ET_0 variation.

In general, the normalization method used in the structure of the MCDM method, the kind of machine learning, and the MCDM method can be regarded as the limitations of the study. One of the recommendations for the mentioned result improvement is to use an efficient MCDM method in the preprocessing. Generally, using the exact method as the preprocessing step in each climate, based on the data capabilities of the area and selection of the effective data, can upgrade the efficiency of ET_0 estimation. It can lead to the precise determination of water availability and strong policymaking in irrigation planning, agricultural studies.

Acknowledgments: We would like to express our sincere gratitude to the Azarbaijan Shahid Madani University for the financial and logistical support that significantly contributed to the research project.

Author contribution:

Laleh Parviz: Analysis, findings, methodology development, results, and discussion, introduction, literature review, paper writing, including result analysis, and grammar check. **Fardin Ghanbari-Maleki:** Data collection, methodology. All authors significantly contributed to the conception and design of the study. The manuscript has been thoroughly reviewed and approved by all named authors.

Conflict of interest: The authors of this article declared no conflict of interest regarding the authorship or publication of this article.

Availability of data and materials: The datasets are available upon a reasonable request to the corresponding author.

References

Gong, A. M. (2016). Using automatic calibration method for optimizing the performance of Pedotransfer functions of saturated hydraulic conductivity. Ain Shams Engineering Journal, *7*(2), 653-662. doi: 10.1016/j.asej.2015.05.012

Ahmadi, F., Mehdizadeh, S., Mohammadi, B., Pham, Q. B., Doan, T. N. C., Vo, N. D. (2021). Application of an artificial intelligence technique enhanced with intelligent water drops for monthly reference evapotranspiration estimation. Agricultural Water Management, 244, 106622. doi: 10.1016/j.agwat.2020.106622

Ahmadpari, H., Khaustov, V. (2025). Analyzing meteorological and hydrological droughts in the Darreh Dozdan River basin through drought indices. Environment and Water Engineering, 11(2), 174-184. doi: 10.22034/ewe.2025.506959.2004

Berti, A., Tardivo, G., Chiaudani, A., Rech, F., Borin, M. (2014). Assessing reference evapotranspiration by the Hargreaves method in north-eastern Italy. Agricultural Water Management, 140, 20-5. doi: 10.1016/j.agwat.2014.03.015

Butchart-Kuhlmann, D., Kralisch, S., Fleischer, M., Meinhardt, M., Brenning, A. (2018). Multicriteria decision analysis framework for hydrological decision support using environmental flow components. Ecological Indicators, 93, 470-480. doi:10.1016/j.ecolind.2018.04.057

Cai, W., Wen, X., Li, C., Shao, J., Xu, J. (2023). Predicting the energy consumption in buildings using the optimized support vector regression model. Energy, 273, 127188. doi: 10.1016/j.energy.2023.127188

Chauhan, S., Shrivastava, R. K. (2009). Performance evaluation of reference evapotranspiration estimation using climate based methods and artificial neural networks. Water Resources Management, 23(5), 825-837. doi: 10.1007/s11269-008-9301-5

De Martonne, E. (1925). TraitéGéographie. Physique: 3 tomes. Max leclcrc and H. Bourrclier, proprietors of LibrairicArmard Colin: Paris.

Dooley, A.E., Smeaton, D. C., Sheath, G. W., Ledgard, S. F. (2009). Application of multiple criteria decision analysis in the New Zealand agricultural industry. The Journal of Multi-Criteria Decision Analysis, 16(1-2), 39-53. doi: 10.1002/mcda.437

Dwivedi, P. P., Sharma, D. K. (2022a). Application of Shannon Entropy and COCOSO techniques to analyze performance of sustainable development goals: The case of the Indian Union Territories. Results in Engineering, 14, 100416. doi: 10.1016/j.rineng.2022.100416

Dwivedi, P. P., Sharma, D. K. (2022b). Application of Shannon entropy and CoCoSo methods in selection of the most appropriate engineering sustainability components. Cleaner Materials, 5, 100118. doi: 10.1016/j.clema.2022.100118

Ellenburg, W. L., Cruise, J., Singh, V. P. (2017). The Role of Evapotranspiration in Streamflow Modeling-an Analysis Using Entropy Theory. In AGU Fall Meeting Abstracts 2017 Dec (Vol. 2017, pp. H23C-1677).

Fu, T., Li, X., Jia, R., Feng, L. (2021). A novel integrated method based on a machine learning model for estimating evapotranspiration in dryland. Journal of Hydrology, 603, 126881. doi: 10.1016/j.jhydrol.2021.126881

Ghabaei Sough, M., Mosaedi, A., Hesam, M., Hezarjaribi, A. (2010). Evaluation Effect of Input Parameters Preprocessing in Artificial Neural Networks (Anns) by Using Stepwise Regression and Gamma Test Techniques for Fast Estimation of Daily Evapotranspiration. Water and Soil, 24(3), 610-624. doi: 10.22067/jsw.v0i0.3631

Gong, D., Hao, W., Gao, L., Feng, Y., Cui, N. (2021). Extreme learning machine for reference crop evapotranspiration estimation: Model optimization and spatiotemporal assessment across different climates in China. Computers and Electronics in Agriculture, 187, 106294. doi: 10.1016/j.compag.2021.106294

Haoyuan, S., Yizhong, M., Chenglong, L., Jian, Z., Lijun, L. (2023). Hierarchical Bayesian support vector regression with model parameter calibration for reliability modeling and prediction. Reliability Engineering and System Safety, 229, 108842. doi: 10.1016/j.ress.2022.108842

Hu, X., Shi, L., Lian, X., Bian, J. (2023). Parameter variability across different timescales in the energy balance-based model and its effect on evapotranspiration estimation. Science of the Total Environment, 871, 161919. doi: 10.1016/j.scitotenv.2023.161919.

Jiang, G. J., Chen, H. X., Sun, H. H., Yazdi, M., Nedjati, A., Adesina, K. A. (2021). An improved multi-criteria emergency decision-making method in environmental disasters. Soft Computing, 25(15), 10351-10379. doi: 10.1007/s00500-021-05826-x

Kim, H. J., Chandrasekara, S., Kwon, H. H., Lima, C., Kim, T. W. (2023). A novel multi-scale parameter estimation approach to the Hargreaves-SamaniEq. for estimation of Penman-Monteith reference evapotranspiration. Agricultural Water Management, 275, 108038. doi: 10.1016/j.agwat.2022.108038

Malek, M. H., Berger, D. E., Coburn, J. W. (2007). On the inappropriateness of stepwise regression analysis for model building and testing. European Journal of Applied Physiology, 101, 263-264. doi: 10.1007/s00421-007-0485-9

Maroufpoor, S., Bozorg-Haddad, O., Maroufpoor, E. (2020). Reference evapotranspiration estimating based on optimal input combination and hybrid artificial intelligent model: Hybridization of artificial neural network with grey wolf optimizer algorithm. Journal of Hydrology, 588, 125060. doi: 10.1016/j.jhydrol.2020.125060

Musbah, H., Ali, G., Aly, H.H., Little, T. A. (2022). Energy management using multi-criteria decision making and machine learning classification algorithms for intelligent system. Electric Power Systems Research, 203, 107645. doi: 10.1016/j.epsr.2021.107645

Nieminen, P. (2022). Application of standardized regression coefficient in meta-analysis. BioMedInformatics, 2(3), 434-458. doi:10.3390/biomedinformatics2030028

Raffinetti, E., Aimar, F. (2019). MDCgo takes up the association/correlation challenge for grouped ordinal data. AStA Advances in Statistical Analysis, 103(4), 527-561. doi: 10.1007/s10182-018-00341-1

Rezaei, I., Amirshahi, S. H., Mahbadi, A. A. (2023). Utilizing support vector and kernel ridge regression methods in spectral reconstruction. Results in Optics, 11, 100405. doi: 10.1016/j.rio.2023.100405

Saroughi, M., Mirzania, E., Achite, M., Katipoğlu, O. M., Al-Ansari, N., Vishwakarma, D. K., Chung, I. M., Alreshidi, M. A., Yadav, K. K. (2024). Evaluate effect of 126 pre-processing methods on various artificial intelligence models accuracy versus normal mode to predict groundwater level (case study: Hamedan-Bahar Plain, Iran). Heliyon, 10(7). doi: 10.1016/j.heliyon.2024.e29006

Shternshis, A., Mazzarisi, P., Marmi, S. (2022). Measuring market efficiency: The Shannon entropy of high-frequency financial time series. Chaos, Solitons & Fractals, 162, 112403. doi: 10.1016/j.chaos.2022.112403

Shu, Z., Zhou, Y., Zhang, J., Jin, J., Wang, L., Cui, N., Wang, G., Zhang, J., Wu, H., Wu, Z., Chen, X. (2022). Parameter regionalization based on machine learning optimizes the estimation of reference evapotranspiration in data deficient area. Science of the Total Environment, 844, 157034. doi: 10.1016/j.scitotenv.2022.157034

Smith, G. (2018). Step away from stepwise. Journal of Big Data, 5(32), 1-12. doi: 10.1186/s40537-018-0143-6

Su, Q., Singh, V. P., Karthikeyan, R. (2022). Improved reference evapotranspiration methods for regional irrigation water demand estimation. Agricultural Water Management, 274, 107979. doi: 10.1016/j.agwat.2022.107979

Tabar, H., Hosseinzadeh Talaee, P. (2013). Multilayer perceptron for reference evapotranspiration estimation in a semiarid region. Neural Computing & Applications, 23, 341-348. doi: 10.1007/s00521-012-0904-7

Yadeta, D., Kebede, A., Tessema, N. (2020). Potential evapotranspiration models evaluation, modelling, and projection under climate scenarios, Kesem sub-basin, Awash River basin, Ethiopia. Modeling Earth System and Environment, 6, 2165-2176. doi: 10.1007/s40808-020-00831-9

Yao, Y., Mallik, A. U. (2022). Estimation of actual evapotranspiration and water stress in the Lijiang River Basin, China using a modified Operational Simplified Surface Energy Balance (SSEBop) model. Journal of Hydro-environment Research, 41, 1-11. doi: 10.1016/j.jher.2022.01.003

Zhu, N., Wang, J., Luo, D. (2024). Unveiling evapotranspiration patterns and energy balance in a subalpine forest of the Qinghai–Tibet Plateau: observations and analysis from an eddy covariance system. Journal of Forestry Research, 35, 53. doi: 10.1007/s11676-024-01708-8