frontiers | Frontiers in Environmental Science

ORIGINAL RESEARCH
published: 04 April 2022
doi: 10.3389/fenvs.2022.826165

Feasibility of Random Forest and Multivariate Adaptive Regression Splines for Predicting Long-Term Mean Monthly Dew Point Temperature

Guodao Zhang[1], *Sayed M. Bateni*[2], *Changhyun Jun*[3]*, *Helaleh Khoshkam*[2], *Shahab S. Band*[4]* and *Amir Mosavi*[5,6,7]

[1]College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou, China, [2]Department of Civil and Environmental Engineering and Water Resources Research Center, University of Hawaii at Manoa, Honolulu, HI, United States, [3]Department of Civil and Environmental Engineering, College of Engineering, Chung-Ang University, Seoul, Korea, [4]Future Technology Research Center, National Yunlin University of Science and Technology, Douliou, Taiwan, [5]John von Neumann Faculty of Informatics, Obuda University, Budapest, Hungary, [6]Institute of Information Society, University of Public Service, Budapest, Hungary, [7]Institute of Information Engineering, Automation and Mathematics, Slovak University of Technology in Bratislava, Bratislava, Slovakia

The accurate estimation of dew point temperature (T_{dew}) is important in climatological, agricultural, and agronomical studies. In this study, the feasibility of two soft computing methods, random forest (RF) and multivariate adaptive regression splines (MARS), is evaluated for predicting the long-term mean monthly T_{dew}. Various weather variables including air temperature, sunshine duration, relative humidity, and incoming solar radiation from 50 weather stations in Iran as well as their geographical information (or a subset of them) are used in RF and MARS as inputs. Three statistical indicators namely, root mean square error (*RMSE*), mean absolute error (*MAE*), and correlation coefficient (*R*) are used to assess the accuracy of T_{dew} estimates from both models for different input configurations. The results demonstrate the capability of the RF and MARS methods for predicting the long-term mean monthly T_{dew}. The combined scenarios in both the RF and MARS methods are found to produce the best T_{dew} estimates. The best T_{dew} estimates were obtained by the MARS model with the *RMSE*, *MAE*, and *R* of respectively 0.17°C, 0.14°C, and 1.000 in the training phase; 0.15°C, 0.12°C, and 1.000 in the validation phase; and 0.18°C, 0.14°C, and 0.999 in the testing phase.

Keywords: dew point temperature, random forest, multivariate adaptive regression splines, machine learning, big data, artificial intelligence

OPEN ACCESS

Edited by:
Hong Liao,
Nanjing University of Information Science and Technology, China

Reviewed by:
Saeid Mehdizadeh,
Urmia University, Iran
Wei Sun,
Sun Yat-Sen University, China

***Correspondence:**
Changhyun Jun
cjun@cau.ac.kr
Shahab S. Band
shamshirbands@yuntech.edu.tw

Specialty section:
This article was submitted to Atmosphere and Climate, a section of the journal Frontiers in Environmental Science

Received: *30 November 2021*
Accepted: *11 March 2022*
Published: *04 April 2022*

Citation:
Zhang G, Bateni SM, Jun C, Khoshkam H, Band SS and Mosavi A (2022) Feasibility of Random Forest and Multivariate Adaptive Regression Splines for Predicting Long-Term Mean Monthly Dew Point Temperature. Front. Environ. Sci. 10:826165. doi: 10.3389/fenvs.2022.826165

INTRODUCTION

Dew point temperature (T_{dew}) is defined as the temperature (at constant pressure) in which water vapor in the air condenses into liquid water. The accurate estimation of T_{dew} is required in many fields such as climatology, hydrology, meteorology, and agronomy (Emmel et al., 2010; Millán et al., 2010; Katul et al., 2012; Feld et al., 2013; Mohammadi et al., 2015; Mohammadi et al., 2016; Alizamir et al., 2020a). T_{dew} along with the wet bulb temperature can be used to compute ambient temperature (Snyder and Melo-Abreu, 2005; Shank, 2006; Mohammadi et al., 2016). The dew point also allows plants to adapt themselves for possible frosts (Mohammadi et al., 2016). T_{dew} is an essential element for plant survival, particularly in regions with low

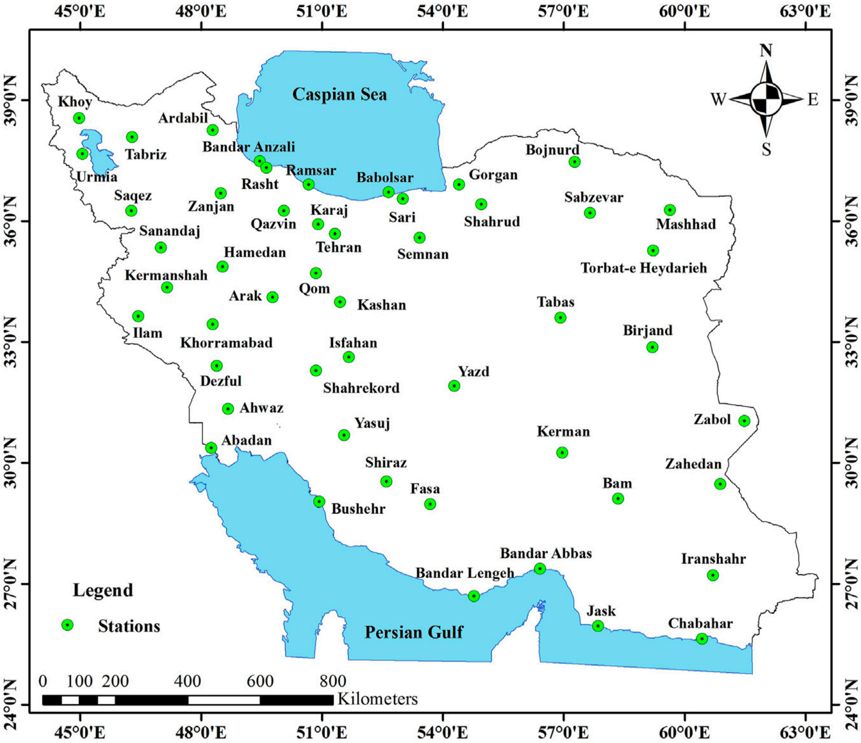

FIGURE 1 | Spatial distribution of the studied stations in Iran.

precipitation (Agam and Berliner, 2006). T_{dew} is necessary for estimating relative humidity and evapotranspiration (Hubbard et al., 2003). Robinson (2000) stated that T_{dew} is important for assessing long-term climate variability.

In recent years, soft computing and data mining approaches have been widely employed as powerful techniques for predicting T_{dew}. A review of the literature shows that random forest (RF) and multivariate adaptive regression splines (MARS) methods have rarely been utilized to estimate T_{dew}; however, they have been extensively used for predicting other hydro-climatological variables (Heddam et al., 2020; Kisi et al., 2021; Tan et al., 2021).

Shank et al. (2008) predicted T_{dew} at 20 weather stations in Georgia by using weather data into artificial neural networks (ANN). It was found that ANN could reliably predict T_{dew}. Zounemat-Kermasni (2012) predicted hourly T_{dew} data via the ANN and multiple linear regression (MLR) approaches. Kisi et al. (2013) evaluated the robustness of generalized regression neural networks (GRNN), Kohonen self-organizing feature maps (KSOFM), and adaptive neuro-fuzzy inference system (ANFIS) for estimating T_{dew} at the Daegu, Pohang, and Ulsan stations in South Korea. The accuracy of GRNN and ANFIS were similar and better than that of KSOFM. Shiri et al. (2014) estimated daily T_{dew} data at two weather stations in the Republic of Korea using gene expression programming (GEP) and ANN models. Various combinations of climatic variables were used as inputs, with the accuracy of GEP was found to be higher than that of ANN. Kim et al. (2015) investigated the potential of multi-layer perceptron (MLP), GRNN, and MLR in estimating daily T_{dew} at two weather stations in California. They defined different combinations of weather data as model predictors. The results indicated that the T_{dew} estimates from GRNN were better than those of MLP. Mohammadi et al. (2015) evaluated the accuracy of the extreme learning machine (ELM), ANN, and support vector machine (SVM) approaches in predicting daily T_{dew} at Bandar Abbas and Tabas, Iran. The mean air temperature, relative humidity, atmospheric pressure, solar radiation, and vapor pressure were used as model inputs. The results revealed that ELM and ANN produced the best and worst daily T_{dew} estimates, respectively. Amirmojahedi et al. (2016) utilized a coupled model by combining ELM with wavelet transform (WT) for predicting daily T_{dew} in Bandar Abbas, South Iran. The accuracies of hybrid ELM-WT and single ELM were compared with those of SVM and ANN. Four different input scenarios were used in their models. Mohammadi et al. (2016) estimated daily T_{dew} at two stations in Iran by the ANFIS technique. Different ANFIS models were developed using various input combinations. Their results demonstrated that water vapor pressure was the most influential variable for the accurate prediction of T_{dew}. Mehdizadeh et al. (2017a) employed GEP to estimate daily T_{dew} at the Urmia and Tabriz stations in Northwest Iran. Various input scenarios were developed using meteorological variables and lagged T_{dew} data. Moreover, T_{dew} at each station was predicted using data from a nearby station. Qasem et al. (2019) estimated daily T_{dew} at the Tabriz station in Iran using GEP, SVM, and M5 model tree (M5), with M5 was found to show the highest performance. Naganna et al. (2019) attempted to increase

TABLE 1 | Geographical properties of the stations in Iran and long-term mean annual values of T_{dew}.

Stations	Latitude (°N)	Longitude (°E)	Altitude (m)	Mean Annual T_{dew} (°C)
Abadan	30.37	48.25	6.6	10.03
Ahwaz	31.33	48.67	22.5	9.55
Arak	34.10	49.77	1708	0.02
Ardabil	38.25	48.28	1332	3.57
Babolsar	36.72	52.65	-21	13.45
Bam	29.10	58.35	1066.9	2.60
Bandar Abbas	27.22	56.37	9.8	19.24
Bandar Anzali	37.48	49.45	-23.6	13.26
Bandar Lengeh	26.53	54.83	22.7	19.40
Birjand	32.87	59.20	1491	-0.82
Bojnurd	37.47	57.27	1112	3.72
Bushehr	28.97	50.82	9	16.98
Chabahar	25.28	60.62	8.0	20.70
Dezful	32.40	48.38	143	9.30
Fasa	28.97	53.68	1288.3	2.80
Gorgan	36.90	54.40	0	11.49
Hamedan	34.87	48.53	1741.5	0.64
Ilam	33.63	46.43	1337	0.42
Iranshahr	27.20	60.70	591.1	5.62
Isfahan	32.62	51.67	1550.4	-0.02
Jask	25.63	57.77	5.2	20.67
Karaj	35.92	50.90	1312.5	2.58
Kashan	33.98	51.45	982.3	3.36
Kerman	30.25	56.97	1753.8	-2.58
Kermanshah	34.35	47.15	1318.6	0.64
Khorramabad	33.43	48.28	1147.8	3.06
Khoy	38.55	44.97	1103	3.49
Mashhad	36.27	59.63	999.2	2.98
Qazvin	36.25	50.05	1279.2	2.35
Qom	34.70	50.85	877.4	2.02
Ramsar	36.90	50.67	-20	13.06
Rasht	37.32	49.62	-8.6	12.60
Sabzevar	36.20	57.65	972	1.40
Sanandaj	35.33	47.00	1373.4	0.34
Saqez	36.25	46.27	1522.8	0.81
Sari	36.55	53.00	23	13.13
Semnan	35.58	53.42	1127	2.84
Shahrekord	32.28	50.85	2048.9	–0.82
Shahrud	36.42	54.95	1349.1	2.31
Shiraz	29.53	52.60	1484	1.87
Tabas	33.60	56.92	711	2.34
Tabriz	38.08	46.28	1361	1.37
Tehran	35.68	51.32	1190.8	1.48
Torbat-e Heydarieh	35.27	59.22	1450.8	1.21
Urmia	37.67	45.05	1328	2.72
Yasuj	30.68	51.55	1816.3	–0.06
Yazd	31.90	54.28	1237.2	–1.03
Zabol	31.03	61.48	489.2	4.64
Zahedan	29.47	60.88	1370	–0.74
Zanjan	36.68	48.48	1663	0.90

the accuracy of estimating T_{dew} at two stations in India by coupling the MLP with two bio-inspired optimization algorithms. The hybrid methods outperformed the classic MLP. Alizamir et al. (2020b) recommended a deep echo state network (DESN) to forecast daily T_{dew} at two locations in the Republic of Korea. The proposed model produced the best performance compared to other soft computing methods. Dong et al. (2020) improved the performance of ELM by optimization algorithms to estimate daily T_{dew} in Yangling, China. They indicated the better accuracy of hybrid models compared to the classic ELM.

Given the importance of T_{dew} in various disciplines, particularly agriculture and hydrology, its precise prediction is vital. Therefore, this study investigated the applicability of random forest (RF) and multivariate adaptive regression splines (MARS) for predicting the long-Temperature-, sunshine duration-, radiation-, other climatic variables-, geographical information-, and combined-based input scenarios were considered in this study.

TABLE 2 | Statistical characteristics of long-term mean monthly meteorological data.

Parameter	Minimum	Maximum	Mean	Standard Deviation	Coefficient of Variation	Correlation with T_{dew}
T_{dew}, °C	–7.90	27.62	5.22	7.72	1.48	1.000
T_{min}, °C	–8.69	30.70	11.17	9.22	0.83	0.793
T_{max}, °C	2.30	46.30	24.22	10.28	0.42	0.590
T, °C	–2.79	38.00	17.82	9.72	0.55	0.695
S, hr	2.89	11.87	7.97	2.27	0.28	0.228
S_o, hr	9.35	14.65	12.00	1.55	0.13	0.465
S/S_o, -	0.25	0.88	0.66	0.14	0.22	0.055
R_s, MJ m^{-2}	6.20	27.84	17.83	6.07	0.34	0.400
R_a, MJ m^{-2}	14.65	41.70	30.37	8.55	0.28	0.491
R_s/R_a, -	0.38	0.69	0.58	0.07	0.12	0.055
RH, %	17.00	87.00	51.30	19.34	0.38	0.218
V_p, hpa	3.69	37.21	10.57	6.71	0.63	0.964
P, mm	0.00	308.91	29.63	39.59	1.34	–0.009
α, -	1.00	12.00	6.50	3.45	0.53	0.142
La, °N	25.28	38.55	33.48	3.60	0.11	–0.287
Lo, °E	44.97	61.48	52.56	4.62	0.09	0.113
Alt, m	–23.60	2048.90	933.26	639.82	0.69	–0.737

Only a few studies used RF and MARS to predict T_{dew} (Shiri, 2018). Also, the correct choice of inputs for soft computing models plays an important role in achieving their optimal performance. Hence, this study attempted to find the best input combination.

MATERIALS AND METHODS

Study Region and Data

The study area was Iran, which is located in southwest Asia. With an area of about 1,648,000 km^2, Iran spans over the latitude of25°00 N′- 40°00 N′ and longitude of 44°00′ E-63°30′ E. The locations of the study stations are shown in **Figure 1**. **Table 1** presents the geographical properties of the selected stations. As can be seen in **Table 1**, the long-term mean annual T_{dew} ranges from -2.58 °C at Kerman to 20.70 °C at Chabahar.

Meteorological data from 50 stations (compiled by the Iran Meteorological Organization, IMO) were utilized in this study. The data include long-term mean monthly dew point temperature (T_{dew}), minimum, maximum, and mean air temperatures (T_{min}, T_{max}, T), solar radiation (R_s), sunshine duration (S), relative humidity (RH), vapor pressure (V_p), and precipitation (P) between 1951 and 2015. Statistical characteristics of these variables are presented in **Table 2**. In this table, S_o and R_a denote the maximum possible sunshine duration and extraterrestrial radiation, respectively, which were calculated based on the relationships presented by Allen et al. (1998). La, Lo and Alt are the latitude, longitude, and altitude of study stations, respectively. We can observe that T_{min}, S_o, R_a and V_p respectively in the temperature-sunshine duration- radiation- and other meteorological variables-based input scenarios have the highest correlations with T_{dew} (**Table 2**). **Figure 2** illustrates the long-term mean monthly of meteorological variables in the study stations.

The data were split into three parts. 70% (420 months), 15% (90 months), and 15% (months) of the data were used for training, testing, and validating the models, respectively.

Random Forest

Random forest (RF), first developed by Breiman (2001), is a powerful ensemble learning algorithm. This model can be employed for regression, classification, and unsupervised learning problems (Liaw and Wiener, 2002). Many decision trees are created using the RF technique via permutation and continual variation of the elements influencing the intended parameter, before all created trees are incorporated for the prediction. Over-fitting, which may occur in the decision tree approach, is eliminated when the number of trees increases. Hence, at every phase of tree growth, the developed model becomes more accurate, and the error rate is reduced. In the RF, the bagging process is utilized to choose random samples of variables as the training dataset. Next, for each variable, if the values of that variable are permuted across the out-of-bag observations, the function specifies the model prediction error (Trigila et al., 2015). Various bootstrap samples of the data, a sampling approach with permutations, were involved in the construction of the RF. Therefore, some out-of-bag datasets were generated from the training dataset via the repetition of the sampling operation.

The number of trees is the most important feature affecting the accuracy of RF (Breiman, 2001). The optimal number of trees is determined by trial and error. 500 trees were used in the RF as increasing the number of trees did not improve its performance.

Multivariate Adaptive Regression Splines

Multivariate adaptive regression splines (MARS) were initially presented by Friedman (1991). This is a non-parametric regression technique, in which the response/target variable can be estimated by using a series of coefficients and functions called basis functions. Cheng and Cao (2014) stated that one of the advantages of MARS is its ability to estimate the contributions of these basis functions. Therefore, the additive and interactive influences of input predictors are allowed to specify the target variable.

The typical form of a MARS model can be defined as follows:

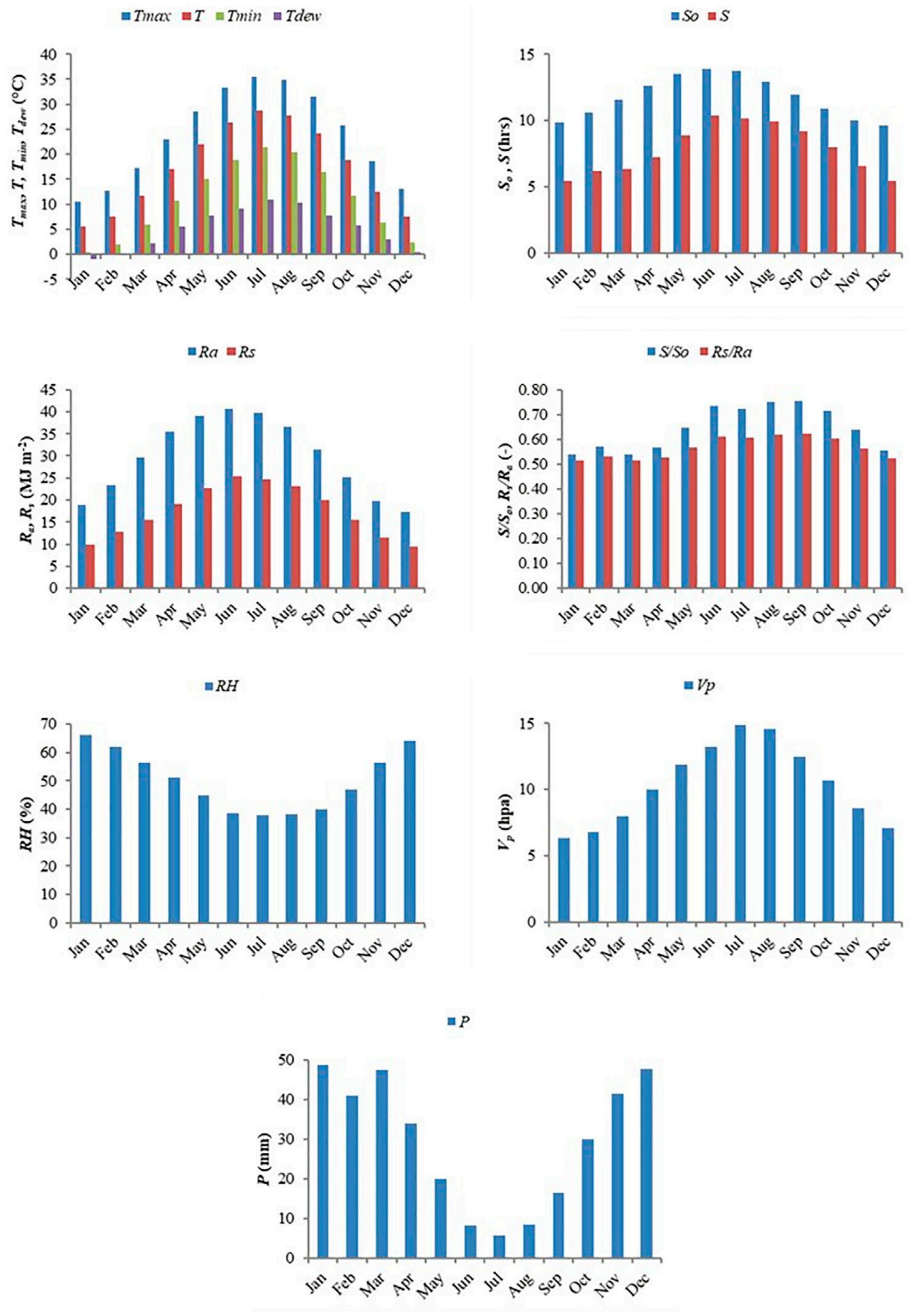

FIGURE 2 | Long-term mean monthly meteorological variables in the study stations.

$$y = f(x) = c_o + \sum\nolimits_{i=1}^{m} c_i b_i(x) \tag{1}$$

where y is the dependent variable predicted by MARS, x is the independent variable(s), c_o is a primary constant or bias, c_i is the coefficient for the ith basis function, and $b_i(x)$ indicates the ith basis function.

The MARS model consists of two phases: forward and backward. The prediction process begins using an intercept, which is the

TABLE 3 | Statistical indices of T_{dew} estimates from the RF model for the training, validation, and testing phases.

Type of Scenarios	Inputs	Training			Validation			Testing		
		***RMSE* (°C)**	***MAE* (°C)**	***R***	***RMSE* (°C)**	***MAE* (°C)**	***R***	***RMSE* (°C)**	***MAE* (°C)**	***R***
Temperature-based	T_{min}	4.92	4.06	0.782	4.28	3.24	0.886	2.62	1.79	0.922
	T_{max}	6.17	4.78	0.628	6.56	4.71	0.657	3.60	2.89	0.899
	T	5.63	4.47	0.700	5.31	3.89	0.796	3.00	2.30	0.921
	T_{min}, T_{max}	3.44	2.80	0.906	2.38	1.89	0.965	1.61	1.23	0.941
	T_{min}, T	3.78	3.07	0.881	2.84	2.22	0.954	1.90	1.35	0.930
	T_{min}, T_{max}, T	3.74	3.11	0.887	2.85	2.22	0.953	1.77	1.29	0.938
Sunshine duration-based	S	7.35	5.75	0.366	7.67	5.43	0.448	5.06	4.25	0.673
	S_o	6.44	5.12	0.593	6.53	4.86	0.672	4.38	3.86	0.838
	S/S_o	7.60	5.81	0.280	7.71	5.63	0.424	5.12	4.01	0.543
	S_o, S	5.98	4.66	0.664	5.80	4.23	0.756	4.04	3.32	0.783
	S_o, S/S_o	5.91	4.61	0.672	5.67	4.13	0.764	3.94	3.14	0.788
	S_o, S, S/S_o	5.85	4.55	0.686	5.80	4.32	0.762	3.82	3.14	0.807
Radiation-based	R_s	7.00	5.61	0.466	7.07	5.35	0.584	4.63	3.89	0.789
	R_a	6.55	5.17	0.573	6.70	4.97	0.655	4.40	3.76	0.824
	R_s/R_a	7.60	5.81	0.280	7.71	5.63	0.424	5.12	4.01	0.543
	R_a, R_s	6.20	4.87	0.627	6.02	4.57	0.736	4.18	3.40	0.785
	R_a, R_s/R_a	5.80	4.50	0.690	5.46	4.02	0.791	3.84	3.04	0.809
	R_a, R_s, R_s/R_a	5.90	4.48	0.674	5.66	4.19	0.772	3.71	3.03	0.822
Other meteorological variables-based	RH	6.90	5.40	0.484	6.79	5.28	0.628	4.84	3.78	0.231
	V_p	0.53	0.31	0.998	0.67	0.34	0.997	0.39	0.21	0.996
	P	7.37	5.73	0.370	7.54	5.77	0.554	5.35	4.28	0.495
	V_p, RH	0.57	0.32	0.997	0.70	0.37	0.997	0.35	0.21	0.997
	V_p, P	0.58	0.32	0.997	0.71	0.37	0.997	0.36	0.21	0.997
	V_p, P, RH	0.85	0.50	0.996	0.87	0.46	0.997	0.54	0.34	0.994
Combined	V_p, T_{min}	0.56	0.32	0.998	0.68	0.36	0.997	0.35	0.21	0.997
	V_p, S_o	**0.53**	**0.31**	**0.998**	**0.65**	**0.35**	**0.997**	**0.35**	**0.21**	**0.997**
	V_p, R_a	0.54	0.31	0.998	0.67	0.36	0.997	0.36	0.22	0.997
	V_p, T_{min}, R_a	0.77	0.55	0.996	0.70	0.45	0.998	0.45	0.29	0.995
	V_p, T_{min}, S_o	0.76	0.54	0.996	0.71	0.45	0.998	0.43	0.27	0.995
	V_p, T_{min}, R_a, S_o	0.63	0.43	0.997	0.64	0.38	0.998	0.38	0.23	0.996
Geographical information-based	La, Lo, Alt, α	2.31	1.78	0.959	2.29	1.83	0.968	1.69	1.36	0.943

Note: Bold values indicate the statistical metrics of the best input.

average of the dependent parameter values. The basis functions are subsequently added continuously to the developed model. It should be noted that when the basis functions are added, the model considers the functions that cause a significant reduction in the sum of square errors. In the forward stage, an over-fitted MARS that include a large number of knots is realized. Then, the backwards stage prunes the model until a suitable MARS is presented based on the lowest value for the generalized cross-validation criterion.

Performance Investigation Metrics

The accuracies of the models were evaluated using three statistical metric: root mean square error ($RMSE$), mean absolute error (MAE), and correlation coefficient (R). These metrics can be expressed as follows:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}\left(T_{o,i} - T_{p,i}\right)^2}{N}} \tag{2}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|T_{o,i} - T_{p,i}\right| \tag{3}$$

$$R = \frac{\sum_{i=1}^{N}\left(T_{o,i} - \overline{T_o}\right)\left(T_{p,i} - \overline{T_p}\right)}{\sqrt{\left[\sum_{i=1}^{N}\left(T_{o,i} - \overline{T_o}\right)^2\right]\left[\sum_{i=1}^{N}\left(T_{p,i} - \overline{T_p}\right)^2\right]}} \tag{4}$$

where $T_{o,i}$ and $T_{p,i}$ are the ith measured and predicted long-term mean monthly T_{dew}, respectively; $\overline{T_o}$ and $\overline{T_p}$ denote the mean of the measured and predicted values of the long-term mean monthly T_{dew}, respectively, and N is the number of data points.

Low values for the $RMSE$ and MAE indices, and a high value of the R index indicate higher performance of the model for predicting the long-term mean monthly T_{dew}.

RESULTS AND DISCUSSION

This study evaluated the performance of two soft computing approaches, RF and MARS, for predicting the long-term mean monthly T_{dew} at 50 stations in Iran. Thirty-one scenarios in six categories were considered to identify the most important

TABLE 4 | Statistical indices of T_{dew} estimates from the MARS model for training, validation, and testing phases.

Type of Scenarios	Inputs	Training			Validation			Testing		
		RMSE (°C)	*MAE* (°C)	*R*	*RMSE* (°C)	*MAE* (°C)	*R*	*RMSE* (°C)	*MAE* (°C)	*R*
Temperature-based	T_{min}	5.01	4.10	0.771	4.15	3.16	0.888	2.63	1.76	0.923
	T_{max}	6.37	4.99	0.591	6.44	4.60	0.671	3.84	3.07	0.901
	T	5.80	4.67	0.678	5.63	4.09	0.763	3.25	2.43	0.928
	T_{min}, T_{max}	2.93	2.30	0.929	2.36	1.77	0.965	1.61	1.22	0.927
	T_{min}, T	3.15	2.54	0.917	2.40	1.80	0.966	1.54	1.14	0.935
	T_{min}, T_{max}, T	2.88	2.27	0.931	2.21	1.61	0.970	1.52	1.19	0.936
Sunshine duration-based	S	7.52	5.96	0.307	7.74	5.60	0.437	5.07	4.33	0.660
	S_o	6.79	5.39	0.513	6.96	5.26	0.581	4.65	3.95	0.782
	S/S_o	7.70	5.94	0.231	7.89	5.73	0.389	5.25	4.21	0.499
	S_o, S	6.04	4.70	0.646	5.57	4.14	0.755	3.99	3.22	0.733
	S_o, S/S_o	6.04	4.64	0.645	5.57	4.00	0.761	3.93	3.11	0.743
	S_o, S, S/S_o	5.68	4.29	0.695	4.82	3.37	0.832	3.79	2.98	0.747
Radiation-based	R_s	7.08	5.73	0.445	7.23	5.58	0.543	4.83	4.10	0.743
	R_a	6.88	5.48	0.493	7.14	5.51	0.548	4.60	3.91	0.807
	R_s/R_a	7.70	5.94	0.231	7.89	5.73	0.389	5.25	4.21	0.499
	R_a, R_s	5.79	4.43	0.682	5.46	3.91	0.764	4.28	3.26	0.666
	R_a, R_s/R_a	5.81	4.45	0.678	5.61	4.33	0.750	3.98	3.07	0.745
	R_a, R_s, R_s/R_a	6.14	4.75	0.630	6.04	4.58	0.701	4.13	3.37	0.722
Other meteorological parameters-based	RH	6.99	5.43	0.464	6.92	5.39	0.584	5.40	4.07	0.083
	V_p	0.48	0.38	0.998	0.48	0.36	0.998	0.58	0.44	0.991
	P	7.55	5.96	0.296	7.79	5.92	0.439	5.43	4.48	0.508
	V_p, RH	0.35	0.27	0.999	0.28	0.22	1.000	0.37	0.27	0.997
	V_p, P	0.44	0.34	0.998	0.39	0.30	0.999	0.45	0.33	0.995
	V_p, P, RH	0.24	0.18	1.000	0.19	0.14	1.000	0.23	0.17	0.999
Combined	V_p, T_{min}	0.28	0.22	0.999	0.20	0.15	1.000	0.23	0.19	0.999
	V_p, S_o	0.35	0.26	0.999	0.27	0.20	1.000	0.34	0.26	0.997
	V_p, R_a	0.32	0.24	0.999	0.25	0.18	1.000	0.31	0.22	0.998
	V_p, T_{min}, R_a	0.24	0.18	1.000	0.16	0.12	1.000	0.21	0.17	0.999
	V_p, T_{min}, S_o	0.19	0.15	1.000	0.16	0.12	1.000	0.19	0.15	0.999
	V_p, T_{min}, R_a, S_o	**0.17**	**0.14**	**1.000**	**0.15**	**0.12**	**1.000**	**0.18**	**0.14**	**0.999**
Geographical information-based	La, Lo, Alt, α	2.60	2.04	0.944	2.16	1.67	0.971	2.51	1.83	0.866

Note: Bold values indicate the statistical metrics of the best input.

variables affecting T_{dew}, and to determine the best input combinations. The *RMSE*, *MAE*, and *R* values were employed to assess the accuracy of the methods.

Performance of RF and MARS Approaches

The statistical indices of dew point estimates from the RF and MARS approaches for various input scenarios are presented in **Tables 3**, **4**, respectively.

In the temperature-based input scenarios, T_{min} and T both produced better results than T_{max}., T_{dew} was found to have a higher correlation with T_{min} than T and T_{max}. Therefore, better results were obtained by employing T_{min} as the input. The superiority of T_{min} compared to T and T_{max} was also found by Mohammadi et al. (2016) and Mehdizadeh et al. (2017a). T_{dew} is more correlated with T_{min} as cool air cannot retain water vapor much longer, meaning the effect of T_{min} on T_{dew} is greater than those of T_{max} and T (Mehdizadeh et al., 2017a). To develop scenarios with more inputs, T and T_{max} were added to T_{min}. A similar strategy was followed to develop scenarios with multiple inputs for other categories. The input combination of T_{min} and T_{max} exhibited a better accuracy than T_{min} and T. Also, the scenarios with all inputs generally yielded better results in comparison with the scenarios with fewer inputs, particularly single-input scenarios. Air temperature is typically measured at all weather stations. Therefore, it can be easily used as a possible input predictor to predict T_{dew}.

Among the sunshine duration-based scenarios, S_o and S/S_o were the best and the worst predictors, respectively. Input combinations S_o and S, and S_o and S/S_o generally produced a similar accuracy, particularly for the MARS model. Interestingly, the S_o and S/S_o scenario was slightly better than the S_o and S scenario in the RF approach. The full-input scenario performed best in both the RF and MARS approaches. However, the performance of this scenario is still not accurate enough for predicting T_{dew}. Additionally, a sunshine duration sensor is needed to measure the sunny hours, which may not be available at some locations. Therefore, the application of sunshine duration variables as the only input of the models is not recommended.

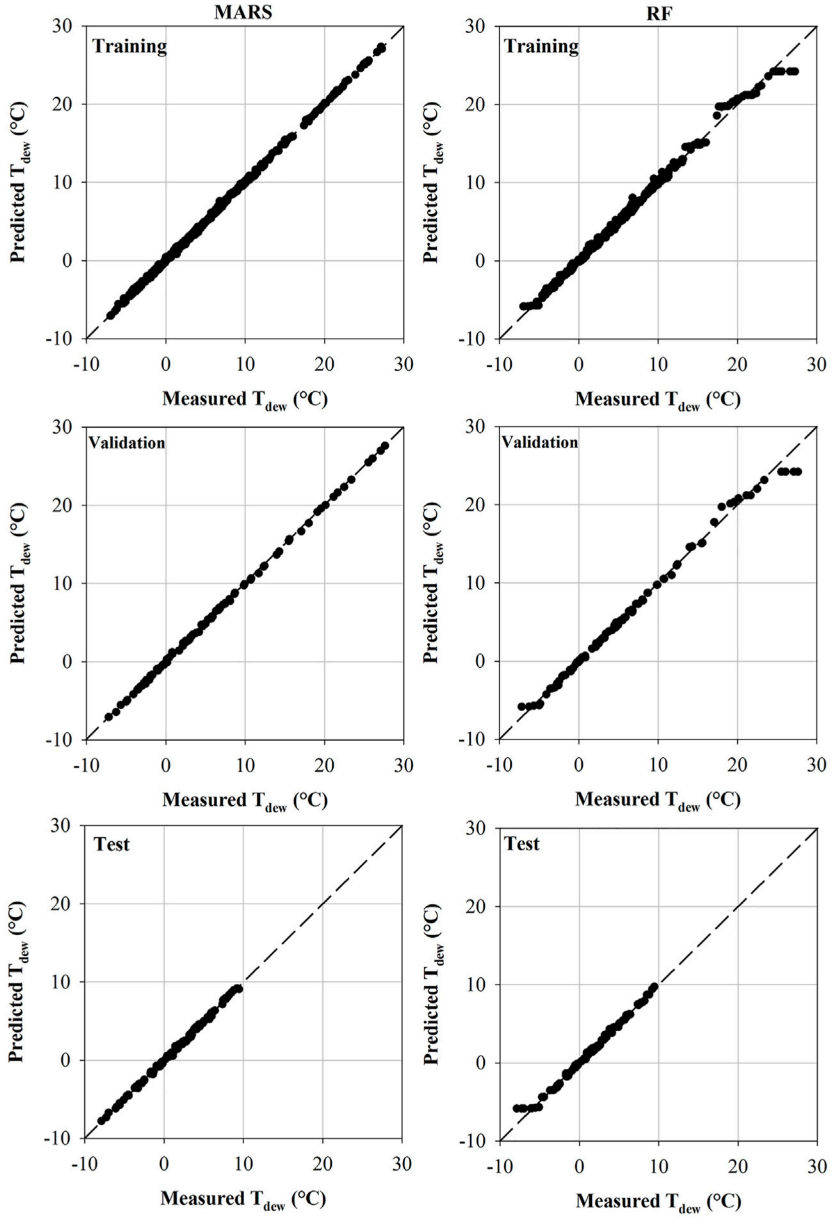

FIGURE 3 | Dew point temperature (T_{dew}) predicted by the superior scenarios of RF and MARS approaches versus the measured values for the training, validation, and test phases.

In the radiation-based scenarios, the input R_a showed the best accuracy, while the performance of the clearness index (R_s/R_a) was not as good as R_s. In general, the performance of the R_a and R_s/R_a input combinations was slightly better than that of R_a and R_s single-input predictors. The RF approach generally produced the highest accuracy with the full-input scenario in the radiation-based classes. However, for the MARS models, two-input scenarios exhibited better performance than the full-input scenario. Similar to the sunshine duration scenarios, radiation-based input combinations did not perform satisfactorily, resulting in higher values of *RMSE* and *MAE* and lower values of *R*. Solar radiation is measured by pyranometer, a relatively expensive device that may not be available at weather stations in developing countries. Therefore, the use of radiation-based scenarios may be limited.

In the other meteorological scenarios, various combinations of *RH*, V_p, and *P* were examined. The results for the single-input scenarios show that V_p is the most influential input variable for the accurate prediction of T_{dew}. Also, the performance of this predictor is better than the most effective variables in temperature- (i.e., T_{min}), sunshine duration- (i.e., S_o), and radiation-based (i.e., R_a) scenarios. For the V_p predictor, the *RMSE, MAE,* and *R* of T_{dew} estimates from the RF method in the testing phase were 0.39°C, 0.21°C, and 0.996, respectively.

Corresponding values from the MARS method were0.58°C, 0.44°C and 0.991. Furthermore, the model with *RH* as input performed better than *P*. Comparing the statistical indices of single *RH* and *P* scenarios with the two- and full-input scenarios shows that the accuracy of T_{dew} predictions significantly increased by adding V_p to *RH* and *P*. For the two-input and three-input scenarios, the *Vp* and *RH* combination in the RF method, and the *Vp*, *P*, and *RH* combination in the MARS method were the best performers.

The most important variables of the four classes (i.e., T_{min}, S_o, R_a, and V_p) were employed to develop the combined scenarios. The performance of T_{min}, S_o, and R_a was not as good as that of V_p. However, the feasibility of T_{min}, S_o, and R_a was considerably improved by adding V_p into them. In the combined-based classes with two inputs, V_p and T_{min} in the MARS model, and V_p and S_o in the RF model yielded slightly better T_{dew} estimates. Interestingly, utilizing three-input and four-input scenarios did not necessarily increase the accuracy of the RF method. But, the accuracy of the MARS method was enhanced by increasing the number of predictors. All combined scenarios produced reliable results due to the higher *R* values and lower *RMSE* and *MAE* values. Unfortunately, these scenarios require many weather variables, which is typically unavailable in developing countries. These scenarios can only be used to predict T_{dew} at weather stations, which are able to measure all required meteorological parameters.

The long-term mean monthly T_{dew} can also be predicted from the geographical characteristics (i.e., latitude, longitude, and altitude) and periodicity (α), which denotes the number of months (i.e., one for January and 12 for December). These predictors can be applied to predict the long-term mean monthly T_{dew} without using meteorological data. These results support the outcomes of previous studies (Kisi et al., 2015; Kisi and Sanikhani, 2015; Mehdizadeh et al., 2017b; Sanikhani et al., 2018) in which the geographical information and number of month were successfully utilized in soft computing models to predict mean monthly time series of hydrological variables such as air and soil temperatures, precipitation, and reference evapotranspiration.

As can be seen in **Tables 3**, **4**, T_{min}, S_o, R_a, and V_p variables showed more accurate results than the other sole-input scenarios. The better performance of these predictors in their respective scenario classes can be attributed to their high correlations with T_{dew} (see **Table 2**).

Comparison of MARS and RF Approaches for Different Input Scenarios

It can be concluded that the RF method is generally superior to the MARS method for the single-input temperature-, sunshine duration-, and radiation-based scenarios. However, the MARS approach generally showed a better performance for the multi-input scenarios. The geographical information-based scenario was superior in the RF method compared to the MARS method. In contrast, the other weather variable-based classes (except the single *RH* and single *P* inputs, and the combined scenarios) performed better in MARS than RF.

Comparison of predicted and measured long-term mean monthly T_{dew} values by the best inputs for the training, validation, and testing phases are depicted in **Figure 3**. As can be seen in **Figure 3**, these inputs can accurately predict long-term mean monthly T_{dew}. As shown in **Tables 3**, **4**, the input combination of V_p and S_o in the RF approach, and V_p, T_{min}, R_a, and S_o in the MARS model were the superior combinations in all of the three study periods (bold text in **Tables 3**, **4**). The estimates of long-term mean monthly T_{dew} using these inputs are very close to the measured data, particularly for the MARS method.

The results revealed that the other weather variable-based (except the single *RH* and single *P* variables) and combined scenarios outperformed the other scenarios (Table … …). However, for both methods, combined scenarios indicated a slightly better performance over other weather variables-based scenarios. Temperature-based combinations had better performance compared to sunshine duration- and radiation-based scenarios, which both had the lowest prediction accuracies. Furthermore, the accuracy of the geographical information-based combinations was better than the temperature-, sunshine duration-, and radiation-based scenarios. This confirms the feasibility of RF and MARS for predicting the long-term mean monthly T_{dew} from the geographical information and the periodicity term.

CONCLUSION

This study evaluated the performance of two soft computing approaches, random forest (RF) and multivariate adaptive regression splines (MARS), for predicting the long-term mean monthly T_{dew}. To specify the influential variables, different input combinations consisting of meteorological variables, geographical characteristics, and the periodicity component were employed as inputs in the RF and MARS models. The meteorological variables included minimum, maximum, and mean air temperatures (T_{min}, T_{max}, and *T*); actual sunshine duration, maximum possible sunshine duration, and sunshine duration ratio (*S*, S_o, and S/S_o); actual solar radiation, extraterrestrial radiation, and clearness index (R_s, R_a, and R_s/R_a); and relative humidity (*RH*), vapor pressure (V_p), and precipitation (*P*). Thirty-one input scenarios were considered in six different categories: temperature-, sunshine duration-, radiation-, other weather variable-, geographical information-based, and combined scenarios. The results obtained are summarized as follows:

- For the single-input scenarios, T_{min}, S_o, R_a, and V_p were the optimum inputs for the temperature-, sunshine duration-, radiation-, and other weather variables r-based scenarios, respectively. Among these variables, V_p had the best performance.
- sunshine duration- and radiation-based scenarios showed the lowest accuracy, while the combined scenarios performed the best.
- The geographical information-based scenarios were superior to the temperature-, sunshine duration-, and radiation-based scenarios. Therefore, the geographical properties and periodicity term can be used to predict

the long-term mean monthly T_{dew} without using any meteorological data.

- In general, the single-input scenarios had a higher accuracy for the RF model compared to the MARS model. While, the multi-input scenarios in the MARS model outperformed the RF method.
- The best multi-input combinations were V_p and S_o for RF, and V_p, T_{min}, R_a and S_o for MARS.
- V_p can be used as the sole input in both the RF and MARS approaches to predict the long-term mean monthly T_{dew} with acceptable accuracy.

Often only a few input configurations were used to estimate different hydrologic variables such as evaporation, solar radiation, soil temperature. The various inputs scenarios used in this study can be tested in future works to find the best input combinations for estimating different variables of interest. Other standalone and coupled models can be used in future studies to estimate T_{dew} and compare it with the outcomes of this work.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

All the authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

REFERENCES

Agam, N., and Berliner, P. R. (2006). Dew Formation and Water Vapor Adsorption in Semi-arid Environments-A Review. *J. Arid Environments* 65 (4), 572–590. doi:10.1016/j.jaridenv.2005.09.004

Alizamir, M., Kim, S., Zounemat-Kermani, M., Heddam, S., Kim, N. W., and Singh, V. P. (2020a). Kernel Extreme Learning Machine: an Efficient Model for Estimating Daily Dew point Temperature Using Weather Data. *Water* 12 (9), 2600. doi:10.3390/w12092600

Alizamir, M., Kim, S., Kisi, O., and Zounemat-Kermani, M. (2020b). Deep echo State Network: a Novel Machine Learning Approach to Model Dew point Temperature Using Meteorological Variables. *Hydrological Sci. J.* 65 (7), 1173–1190. doi:10.1080/02626667.2020.1735639

Allen, R. G., Pereira, L. S., Raes, D., and Smith, M. (1998). *Crop Evapotranspiration. GuideLines for Computing Crop Evapotranspiration*. Rome, Italy: FAO Irrigation and Drainage Paper No. 56.

Amirmojahedi, M., Mohammadi, K., Shamshirband, S., Seyed Danesh, A., Mostafaeipour, A., and Kamsin, A. (2016). A Hybrid Computational Intelligence Method for Predicting Dew point Temperature. *Environ. Earth Sci.* 75, 1–12. doi:10.1007/s12665-015-5135-7

Breiman, L. (2001). Random Forests. *Mach. Learn.* 45, 5–32. doi:10.1023/a:1010933404324

Cheng, M.-Y., and Cao, M.-T. (2014). Accurately Predicting Building Energy Performance Using Evolutionary Multivariate Adaptive Regression Splines. *Appl. Soft Comput.* 22, 178–188. doi:10.1016/j.asoc.2014.05.015

Dong, J., Wu, L., Liu, X., Li, Z., Gao, Y., Zhang, Y., et al. (2020). Estimation of Daily Dew point Temperature by Using Bat Algorithm Optimization Based Extreme Learning Machine. *Appl. Therm. Eng.* 165, 114569. doi:10.1016/j.applthermaleng.2019.114569

Emmel, C., Knippertz, P., and Schulz, O. (2010). Climatology of Convective Density Currents in the Southern Foothills of the Atlas Mountains. *J. Geophys. Res.* 115 (D11). doi:10.1029/2009jd012863

Feld, S. I., Cristea, N. C., and Lundquist, J. D. (2013). Representing Atmospheric Moisture Content along Mountain Slopes: Examination Using Distributed Sensors in the Sierra Nevada, California. *Water Resour. Res.* 49 (7), 4424–4441. doi:10.1002/wrcr.20318

Friedman, J. H. (1991). Multivariate Adaptive Regression Splines. *Ann. Statist.* 19, 1–67. doi:10.1214/aos/1176347963

Heddam, S., Ptak, M., and Zhu, S. (2020). Modelling of Daily lake Surface Water Temperature from Air Temperature: Extremely Randomized Trees (ERT) versus Air2Water, MARS, M5Tree, RF and MLPNN. *J. Hydrol.* 588, 125130. doi:10.1016/j.jhydrol.2020.125130

Hubbard, K. G., Mahmood, R., and Carlson, C. (2003). Estimating Daily Dew point Temperature for the Northern Great Plains Using Maximum and Minimum Temperature. *Agron. J.* 95 (2), 323–328. doi:10.2134/agronj2003.0323

Katul, G. G., Oren, R., Manzoni, S., Higgins, C., and Parlange, M. B. (2012). Evapotranspiration: a Process Driving Mass Transport and Energy Exchange in the Soil-Plant-Atmosphere-Climate System. *Rev. Geophys.* 50 (3). doi:10.1029/2011RG000366

Kim, S., Singh, V. P., Lee, C.-J., and Seo, Y. (2015). Modeling the Physical Dynamics of Daily Dew point Temperature Using Soft Computing Techniques. *KSCE J. Civ Eng.* 19 (6), 1930–1940. doi:10.1007/s12205-014-1197-4

Kisi, O., Kim, S., and Shiri, J. (2013). Estimation of Dew point Temperature Using Neuro-Fuzzy and Neural Network Techniques. *Theor. Appl. Climatol.* 114 (3-4), 365–373. doi:10.1007/s00704-013-0845-9

Kisi, O., Sanikhani, H., Zounemat-Kermani, M., and Niazi, F. (2015). Long-term Monthly Evapotranspiration Modeling by Several Data-Driven Methods without Climatic Data. *Comput. Elect. Agric.* 115, 66–77. doi:10.1016/j.compag.2015.04.015

Kisi, O., Khosravinia, P., Heddam, S., Karimi, B., and Karimi, N. (2021). Modeling Wetting Front Redistribution of Drip Irrigation Systems Using a New Machine Learning Method: Adaptive Neuro- Fuzzy System Improved by Hybrid Particle Swarm Optimization - Gravity Search Algorithm. *Agric. Water Manag.* 256, 107067. doi:10.1016/j.agwat.2021.107067

Kisi, O., and Sanikhani, H. (2015). Prediction of Long-Term Monthly Precipitation Using Several Soft Computing Methods without Climatic Data. *Int. J. Climatol.* 35 (14), 4139–4150. doi:10.1002/joc.4273

Liaw, A., and Wiener, M. (2002). Classification and Regression by Random forest. *R. News* 2 (3), 18–22.

Mehdizadeh, S., Behmanesh, J., and Khalili, K. (2017a). Application of Gene Expression Programming to Predict Daily Dew point Temperature. *Appl. Therm. Eng.* 112, 1097–1107. doi:10.1016/j.applthermaleng.2016.10.181

Mehdizadeh, S., Behmanesh, J., and Khalili, K. (2017b). Evaluating the Performance of Artificial Intelligence Methods for Estimation of Monthly Mean Soil Temperature without Using Meteorological Data. *Environ. Earth Sci.* 76, 1–16. doi:10.1007/s12665-017-6607-8

Millán, H., Ghanbarian-Alavijeh, B., and García-Fornaris, I. (2010). Nonlinear Dynamics of Mean Daily Temperature and Dewpoint Time Series at Babolsar, Iran, 1961-2005. *Atmos. Res.* 98, 89–101. doi:10.1016/j.atmosres.2010.06.001

Mohammadi, K., Shamshirband, S., Motamedi, S., Petković, D., Hashim, R., and Gocic, M. (2015). Extreme Learning Machine Based Prediction of Daily Dew point Temperature. *Comput. Elect. Agric.* 117, 214–225. doi:10.1016/j.compag.2015.08.008

Mohammadi, K., Shamshirband, S., Petković, D., Yee, P. L., and Mansor, Z. (2016). Using ANFIS for Selection of More Relevant Parameters to Predict Dew point Temperature. *Appl. Therm. Eng.* 96, 311–319. doi:10.1016/j.applthermaleng.2015.11.081

Naganna, S. R., Deka, P. C., Ghorbani, M. A., Biazar, S. M., Al-Ansari, N., and Yaseen, Z. M. (2019). Dew point Temperature Estimation: Application of Artificial Intelligence Model Integrated with Nature-Inspired Optimization Algorithms. *Water* 11 (4), 742. doi:10.3390/w11040742

Qasem, S. N., Samadianfard, S., Nahand, H. S., Mosavi, A., shamshirband, S., and Chau, K.-w. (2019). Estimating Daily Dew point Temperature Using Machine Learning Algorithms. *Water* 11 (3), 582. doi:10.3390/w11030582

Robinson, P. J. (2000). Temporal Trends in United States Dew point Temperatures. *Int. J. Climatol.* 20 (9), 985–1002. doi:10.1002/1097-0088(200007)20:9<985::aid-joc513>3.0.co;2-w

Sanikhani, H., Deo, R. C., Samui, P., Kisi, O., Mert, C., Mirabbasi, R., et al. (2018). Survey of Different Data-Intelligent Modeling Strategies for Forecasting Air Temperature Using Geographic Information as Model Predictors. *Comput. Elect. Agric.* 152, 242–260. doi:10.1016/j.compag.2018.07.008

Shank, D. B. (2006). *Dew point Temperature Prediction Using Artificial Neural Networks*. MS thesis. United Kingdom: Harding University.

Shank, D. B., Hoogenboom, G., and McClendon, R. W. (2008). Dewpoint Temperature Prediction Using Artificial Neural Networks. *J. Appl. Meteorol. Climatol.* 47 (6), 1757–1769. doi:10.1175/2007jamc1693.1

Shiri, J., Kim, S., and Kisi, O. (2014). Estimation of Daily Dew point Temperature Using Genetic Programming and Neural Networks Approaches. *Hydrol. Res.* 45 (2), 165–181. doi:10.2166/nh.2013.229

Shiri, J. (2018). Prediction vs. Estimation of Dewpoint Temperature: Assessing GEP, MARS and RF Models. *Hydrol. Res.* 50 (2), 633–643. doi:10.2166/nh.2018.104

Snyder, R. L., and Melo-Abreu, J. P. D. (2005). *Frost Protection: Fundamentals, Practice and Economics*, 1. Rome: Food and Agricultural Organization of the United Nations.

Tan, J., Xie, X., Zuo, J., Xing, X., Liu, B., Xia, Q., et al. (2021). Coupling Random forest and Inverse Distance Weighting to Generate Climate Surfaces of Precipitation and Temperature with Multiple-Covariates. *J. Hydrol.* 598, 126270. doi:10.1016/j.jhydrol.2021.126270

Trigila, A., Iadanza, C., Esposito, C., and Scarascia-Mugnozza, G. (2015). Comparison of Logistic Regression and Random Forests Techniques for Shallow Landslide Susceptibility Assessment in Giampilieri (NE Sicily, Italy). *Geomorphology* 249, 119–136. doi:10.1016/j.geomorph.2015.06.001

Zounemat-Kermasni, M. (2012). Hourly Predictive Levenberg–Marquardt ANN and Multi Linear Regression Models for Predicting of Dew point Temperature. *Meteorol. Atmos. Phys.* 117, 181–192. doi:10.1007/s00703-012-0192-x

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

NOMENCLATURE

$\mathbf{T_{dew}}$ Dew point temperature

MARS Multivariate adaptive regression splines

RF Random forest

ANN Artificial neural networks

MLR Multiple linear regression

GRNN Generalized regression neural networks

KSOFM Kohonen self-organizing feature maps

ANFIS Adaptive neuro-fuzzy inference system

GEP Gene expression programming

MLP Multi-layer perceptron

ELM Extreme learning machine

SVM Support vector machine

WT Wavelet transform

M5 M5 model tree

DESN Deep echo state network

$\mathbf{T_{min}}$ Minimum air temperature

$\mathbf{T_{max}}$ Maximum air temperature

T Mean air temperature

S sunshine duration

$\mathbf{S_o}$ Maximum possible sunshine duration

$\mathbf{R_s}$ Solar radiation

$\mathbf{R_a}$ Extraterrestrial radiation

RH Relative humidity

$\mathbf{V_p}$ Vapor pressure

P Precipitation

La Latitude

Lo Longitude

Alt Altitude

y Dependent variable predicted using the MARS

x Independent variable in MARS

$\mathbf{c_o}$ Bias

$\mathbf{c_i}$ Coefficient for the *i*th basis function of the MARS

$\mathbf{b_i(x)}$ *i*th basis function

RMSE Root mean square error

MAE Mean absolute error

R Correlation coefficient

$\mathbf{T_{o,i}}$ *i*th measured long-term mean monthly T_{dew}

$\mathbf{T_{p,i}}$ *i*th predicted long-term mean monthly T_{dew}

$\overline{\mathrm{T_o}}$ Mean of the measured values of the long-term mean monthly T_{dew}

$\overline{\mathrm{T_p}}$ Mean of the predicted values of the long-term mean monthly T_{dew}