Prediction vs. estimation of dewpoint temperature: assessing GEP, MARS and RF models

Jalal Shiri

ABSTRACT

Dewpoint temperature (T_{dew}) plays a key role in agricultural issues as well as meteorological studies. This paper is aimed at developing and validating prediction and estimation models of T_{dew} values. Gene expression programming (GEP), multivariate adaptive regression spline (MARS), and random forest (RF) models were employed. Data from six weather stations (consisting of a period of ten years) in East Azerbaijan, northwestern Iran were utilized for establishing, testing, and validating the models. In the case of predicting models, chronological records of T_{dew} in previous time steps were introduced as models' inputs to predict T_{dew} values at daily and weekly prediction intervals. In the case of T_{dew} estimating models, daily records of mean air temperature, sunshine hours, relative humidity, and wind speed were utilized as inputs to estimate T_{dew}. Acquired results showed prediction-based GEP surpasses the MARS and RF models in both daily and weekly prediction intervals. Among the estimation models, the MARS models that relied on air temperature, relative humidity, and sunshine hours presented the most accurate results in all studied locations as well as the studied region. The current study proposes the use of MARS models in estimating T_{dew} magnitudes, while it criticizes the use of single data set assignment for both temporal and spatial analysis.

Key words | dewpoint, estimation, heuristic models, prediction

Jalal Shiri
Water Engineering Department, Faculty of
  Agriculture,
University of Tabriz,
Tabriz,
Iran
E-mail: j_shiri2005@yahoo.com

INTRODUCTION

Dewpoint temperature stands for the temperature at which airborne water vapor is condensed (at constant pressure and water vapor contents) and forms liquid dew. Higher dewpoint values correspond to the higher moisture content in the surrounding air (Wallace & Hobbs 2006). In agricultural issues, dew phenomenon might redact the vapor pressure deficit in the immediate neighborhood of the dew drops, resulting in better photosynthesis (Slatyer 1967) and improving water content recuperation after severe water deprivations (Went 1955). Dominant factors influencing the emergence of dew occurrence in unprocessed ecosystems are radiation trade-off between the Earth's surface and atmosphere, turbulent heat and water vapor pressure (Atzema et al. 1990). Accurate estimation of this parameter is of crucial importance since it will specify whether rainfall or snow will occur. Due to its effect on the vapor pressure deficit, T_{dew} is a very important factor for monitoring the seasonal dynamics of important crop factors such as leaf area index (Savoy & Mackay 2015). It also determines the danger level for grass during dry spells. Dewpoint temperature is utilized to identify the available moisture content of the surrounding air (Shank et al. 2008) and to estimate the near surface humidity. Numerous hydro-climatologic models need dewpoint magnitudes as a necessary input parameter to calculate reference evapotranspiration (Mahmood & Hubbarad 2005). In irrigated farmlands, information on this parameter is crucial for irrigation scheduling (Mahmood et al. 2008).

Dewpoint magnitudes are usually measured by hygrometers, although there are some empirical equations to relate dewpoint values to air temperature/relative humidity, as presented by Allen *et al.* (1998).

As a substitute, heuristic data-driven techniques can be utilized to determine the values of this parameter using easily measured meteorological parameters. Among others, Abdel-Aal (2004) and Smith *et al.* (2005) have employed neural network (NN) models for predicting air temperature. Shank *et al.* (2008) applied NN-based models to predict dew-point temperature. Bilgili & Sahin (2010) utilized NN in predicting long-term temperature and rainfall magnitudes in Turkey. Kisi & Shiri (2011) introduced new hybrid wave-let-heuristic models for precipitation forecast. Kisi *et al.* (2013) used NN and neuro-fuzzy models for predicting dew-point temperature. Shiri *et al.* (2014) applied genetic programming and NN models for estimating dewpoint. Kisi & Shiri (2014) developed a neuro-fuzzy-based modeling strategy for predicting long-term monthly air temperature records using geographical inputs. The literature review shows that, in general, the applied models have been applied with single data set assignment (local scale) without trying to generalize the models' capabilities in both prediction or estimation issues. Nevertheless, the prediction of this parameter using the heuristic models has been rarely reported in the literature. The present study aimed at predicting and estimating dewpoint temperature values using the chronological records of this parameter as well as the meteorological inputs at a regional scale through employing heuristic gene expression programming (GEP), multivariate adaptive regression spline (MARS), and random forest (RF)

methodologies. Except using GEP, the current study is the first time MARS and RF have been applied for predicting/estimating dewpoint temperature.

MATERIALS AND METHODS

Study area and data

Data from six sites located in northwestern Iran were used to evaluate the employed methodologies in modeling dew-point magnitudes. The applied data covers a period of 10 years (1st January 2003 to 31st December 2012) of daily values of mean air temperature (T_{mean}), wind speed (W_S), sunshine hours (S), relative humidity (R_H), and dew-point temperature (T_{dew}). All available patterns were carefully analyzed and screened for any inconsistency. A summary of the studied sites and meteorological data can be found in Table 1. As can be seen from the values presented for the coefficient of variation (C_V), dewpoint temperature had the highest magnitudes of this coefficient towards the other applied data in all studied locations, which depicts the higher values of standard deviation for this parameter. Among the studied stations, Marand presented the lowest mean T_{dew} as well as the highest C_V values, while Jolfa had the highest mean T_{dew} and the lowest C_V values. For developing the employed models, data from 1st January 2003 to 31st December 2007 (50% of total available data) were used for training the models, while the remaining data from 1st January 2008 to 31st December 2009 were reserved

Table 1 │ Summary of the studied sites

Station	Geographical coordinates			Mean and coefficient of variation values of data				
	Latitude (°N)	Longitude (°E)	Altitude (m)	T_{mean} (°C)	R_H (%)	W_S (m/s)	S (hrs)	T_{dew} (°C)
Ahar	38.26	47.4	1,390.5	11.365 (0.778)	59.013 (0.248)	2.395 (0.570)	7.250 (0.542)	2.084 (3.571)
Bonab	37.20	46.4	1,290.0	15.100 (0.703)	52.094 (0.314)	1.280 (0.933)	8.149 (0.495)	2.506 (2.428)
Jolfa	38.45	45.4	736.2	14.967 (0.757)	56.398 (0.267)	1.929 (0.882)	7.436 (0.535)	4.650 (1.608)
Marageh	37.24	46.16	1,477.7	13.508 (0.754)	49.410 (0.354)	2.760 (0.441)	8.180 (0.488)	1.347 (4.368)
Marand	38.28	45.46	1,550.0	12.711 (0.840)	50.363 (0.367)	1.654 (0.855)	7.604 (0.527)	0.706 (9.982)
Tabriz	38.50	46.17	1,361.0	13.703 (0.750)	50.671 (0.323)	2.656 (0.443)	7.769 (0.486)	1.489 (4.322)

Note: the values in the brackets show the coefficient of variations (C_V) values of used daily data set. T_{mean}: mean air temperature, R_H: relative humidity, W_S: wind speed, S: sunshine hours, T_{dew}: dewpoint temperature.

for testing and, finally, data between 1st January 2010 and 31st December 2012 were reserved for validation of the models. Dividing the available patterns into three blocks decreases the risk of over fitting (Pour Ali Baba *et al.* 2013).

Methods

Gene expression programming (GEP)

Genetic programming algorithms at first elucidate an impartial function as a qualitative norm. Next, this function is utilized for measuring and evaluating different solutions in a step by step mode of structural rectifying until genetic programming results in a proper solution. This is an evolutionary algorithm and is favored due to its high perfection. GEP can be considered as similar to genetic programming and now even develops computer programs with different sizes and shapes encoded in linear chromosomes of stable lengths. The chromosomes in GEP are composed of multiple genes, each gene encoding a smaller subprogram. Moreover, the structural and functional operation of the linear chromosomes provides the uncompelled functioning of crucial genetic operators, e.g., mutation, transposition, and recombination. In GEP, the formation of genetic variety is highly clarified as genetic operators act at the chromosomes' level. Nonetheless, GEP has a unique, multigenic nature which provides the development of more complex programs consisting of different subprograms (Ferreira 2001a, 2001b). Detailed information about modeling dewpoint temperature using GEP can be seen in, for example, in Shiri *et al.* (2014).

Multivariate adaptive repression spline (MARS)

MARS is a non-parametric regression approach that can be considered as an adjunct of linear models which automatically simulates the nonlinearities and interactions between parameters (Friedman 1991). The MARS algorithm comprises a forward and backward stepwise plan. In the forward stepwise plan, the immoderate forward stepwise chosen plan would produce a complex and over-trained model after a number of splits (Andres *et al.* 2010), which

will have a lower performance accuracy. Hence, the backward stepwise plan ignores the nonobligatory variables among the previously chosen set. This function would project variable X to a new variable Y through utilizing the following two basic functions, utilizing a knot or value of a variable which defines a conjunction point along the range of inputs (Sharda *et al.* 2006; Adamowski *et al.* 2012). MARS utilizes piecewise linear basis functions with the form of $(x - t)_+$ and $(t - x)_+$. The suffix '$+$' stands for the positive part only. Thus:

$$(x - t)_+ = \begin{cases} x - t, & \text{if } x > t \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

and

$$(t - x)_+ = \begin{cases} t - x, & \text{if } x < t \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

The general form of the MARS equation reads:

$$\hat{f}(x) = \sum_i^k c_i B_i(x) \tag{3}$$

which is indeed a weighted sum of function $B_i(x)$. c_i stands for the constant coefficients determined through minimizing the residual sum of squares (standard linear regression). The coefficients might be assumed as weights which depict the importance of each variable.

Random forest (RF)

RF is an ensemble learning algorithm that handles high-dimension regression problems. RF is a tree-based ensemble method, in which all trees are dependent on a group of random variables, and the forest is grown from many regression trees that are put together and form an ensemble (Breiman 2001). The final decision would be found by averaging the output, after fitting individual trees in the ensemble (bagging procedure). The bias of the bagged trees is the same as that of the individual trees, while the variance is decreased by reduction in the correlation between trees (Hastie *et al.* 2009).

RF for regression is formed by growing trees depending on a random vector, and the tree predictor h(x,ϕ) gets numerical values. The output values are numerical and it is assumed that the training set is independently drawn from the random vector distribution. The mean squared generalization error of numerical simulator h(x) can be shown as:

$$E_{X,Y}(Y - h(X))^2 \qquad (4)$$

The RF simulator forms by averaging over k of the trees ($h(x, \phi_k)$). More detailed information on the RF theorem may be found in, for example, Breiman (2001).

Methodological structure

Two modeling scenarios were developed and examined in the current research, namely, T_{dew} prediction and T_{dew} estimation. Table 2 presents the different input configurations (prediction as well as estimation models) used in the current paper. Regarding the T_{dew} prediction, the models were fed using the previously recorded dewpoint records to forecast the daily and weekly T_{dew} values. Figure 1 displays the partial autocorrelation functions (PACF) of T_{dew} in the studied locations. The figure shows that the 4–5 time lags seem to be significant in predicting the next step T_{dew} values in all the locations, so, for consistency, a time lag of five steps was selected to feed the prediction models of T_{dew} in different prediction intervals. Therefore, in any prediction model (GEP, MARS, and RF), the T_{dew} values recorded at five previous time steps were utilized

to predict its values in the +1 day and +7 days' prediction horizons. In the case of T_{dew} estimation (estimation models), the simultaneous values of T_{mean}, W_S (indicators of turbulent temperature), S (indicator of radiation exchange), and R_H (indicator of water vapor pressure) were used through different input configurations, to estimate the corresponding T_{dew} values at the same time step. Moreover, since the works deals with a time series analysis, an additional dynamic input configuration was examined to relate the T_{dew} values of the present time step to the recorded values of the mentioned meteorological parameters at one day previous. A study-wide data matrix comprising the pooled training patterns (data between 2003 and 2007 of all stations) was built using four input configurations (Table 2) to train the employed T_{dew} estimation models, and then the developed generalized models were tested and validated using the pooled testing and validation patterns of the stations. In addition to a general assessment of the estimation models, a split up analysis of these models' performance per station was also carried out.

Assessing the models' performances

Three statistical performance analysis indicators, namely, the root mean square error (RMSE), the mean absolute error (MAE), and the variance accounted for (VAF) indices were utilized to assess the models' performances, expressions for which are given as:

$$RMSE = \sqrt{\frac{\sum\limits_{i=1}^{n} (T_{dew}^{io} - T_{dew}^{ie})^2}{n}} \qquad (5)$$

$$MAE = \frac{\sum\limits_{i=1}^{n} |T_{dew}^{io} - T_{dew}^{ie}|}{n} \qquad (6)$$

$$VAF = 1 - \frac{Var(T_{dew}^{io} - T_{dew}^{ie})}{Var(T_{dew}^{o})} \qquad (7)$$

where, T_{dew}^{io} and T_{dew}^{ie} present the recorded and corresponding simulated dewpoint temperature magnitudes at the ith time step, respectively. n shows the number of patterns.

Table 2 | Input configurations used to feed the applied models

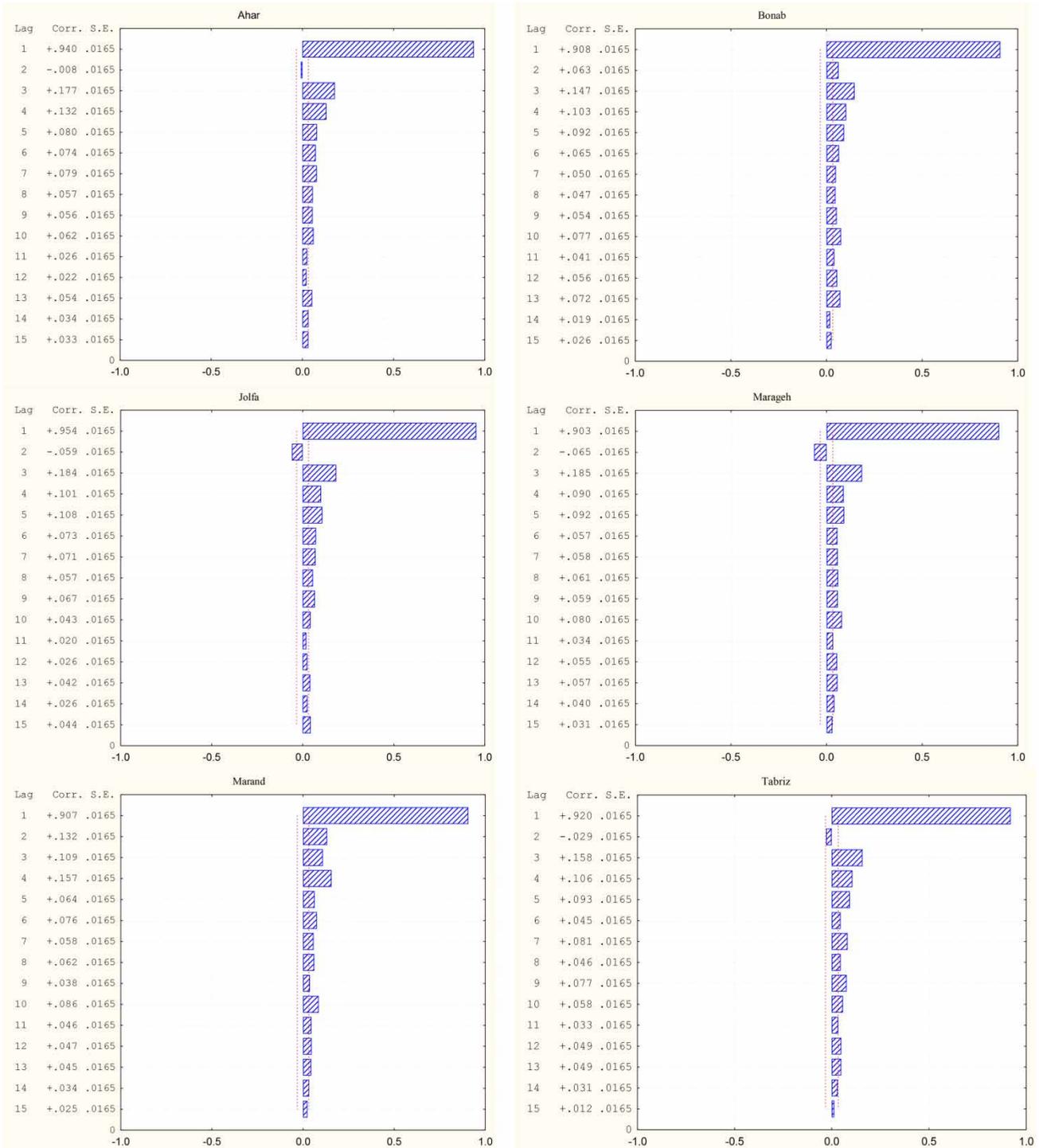|	Inputs	Output	Model
Prediction models	$T_{dew}^{i-5}, T_{dew}^{i-4}, T_{dew}^{i-3}, T_{dew}^{i-2}, T_{dew}^{i-1}$	T_{dew}^{i+1}	–
	$T_{dew}^{i-5}, T_{dew}^{i-4}, T_{dew}^{i-3}, T_{dew}^{i-2}, T_{dew}^{i-1}$	T_{dew}^{i+7}	–
Estimation models	T_{mean}, R_H	T_{dew}	GEP1, MARS1, RF1
	T_{mean}, R_H, S	T_{dew}	GEP2, MARS2, RF2
	T_{mean}, R_H, W_S	T_{dew}	GEP3, MARS3, RF3
	$T_{mean}^{i-1}, R_H^{i-1}, S^{i-1}$	T_{dew}^{i}	GEP4, MARS4, RF4

Figure 1 │ Partial autocorrelation functions (PACFs) of T_{dew} in the studied locations.

The optimum value of the *RMSE* and *MAE* is zero describing the minimum (zero) error in the models, while the optimum value of *VAF* is unity which depicts the minimum assumable variance of the residuals.

RESULTS AND DISCUSSION

Prediction models

Table 3 sums up the statistical indices of the prediction models in all studied locations for both the daily and weekly prediction intervals in the validation period. As could be predicted, daily predictions gave more accurate results than the weekly predictions for all locations. Figure 2 displays the VAF reduction (due to increasing prediction interval) values of all the applied models in the studied stations. The maximum reduction (0.235) was observed for the GEP model of Marageh, while the minimum reduction (0.039) belonged to the RF model of the same station. Analyzing the values presented in Table 2 showed that the VAF values of GEP models for daily and weekly predictions in this station were 0.832 and 0.636, respectively, which exhibited much better performance accuracy for the daily prediction interval. In the case of the RF model, however, neither daily nor weekly intervals presented higher performance accuracy (VAF = 0.633 and 0.608, respectively). This clearly shows that the GEP overall performance is much

better than RF in this station for both the intervals. Nonetheless, in this specific case, it was seen that the models' principal differences play key roles in generating such discrepancies, rather than the nature of the T_{dew} time series. Noticeable differences for the models are also observed for Bonab, Jolfa, and Marand. GEP accuracy reduction was higher than those of MARS and RF (which showed similar reductions) in Bonab and Jolfa, while RF presented the highest VAF reduction in Marand. In contrast, the VAF reductions of all three applied models were almost the same in Tabriz and Ahar (with the exception of small differences for RF). Such discrepancies support the preliminary assumption of the significant role of the models' nature and structure on producing such differences. GEP evolves the structure and constants of a specified solution, simultaneously. Thus, the degrees of freedom over other function fitting methods are increased by using GEP. Combining physically based models with subject matter expertise and site-specific data has been demonstrated to produce higher quality representations over methods that rely solely on a single approach (Deschaine 2014; Shiri 2017). Additionally, in the case of the GEP-based predictions,

Table 3 | Validation statistics of the prediction models (period: 2010–2012)

Prediction interval		+1 day prediction			+7 days' prediction		
Station	Model	RMSE	MAE	VAF	RMSE	MAE	VAF
Ahar	GEP	2.452	1.890	0.898	3.991	3.140	0.728
	MARS	2.387	1.824	0.902	3.983	3.139	0.730
	RF	2.405	1.855	0.901	3.834	3.074	0.749
Bonab	GEP	2.494	1.893	0.842	3.670	2.861	0.652
	MARS	2.983	2.345	0.773	3.669	2.860	0.656
	RF	3.063	2.415	0.761	3.666	2.875	0.656
Jolfa	GEP	2.134	1.611	0.913	3.482	2.658	0.770
	MARS	2.934	2.281	0.831	3.334	2.544	0.788
	RF	2.950	2.317	0.836	3.371	2.580	0.784
Marageh	GEP	2.556	1.945	0.832	3.858	2.995	0.636
	MARS	3.733	2.942	0.642	3.990	3.041	0.611
	RF	3.775	2.987	0.633	3.998	3.110	0.608
Marand	GEP	3.038	2.319	0.836	3.258	2.111	0.800
	MARS	3.003	2.285	0.840	3.358	2.678	0.798
	RF	2.921	2.228	0.848	3.625	2.725	0.758
Tabriz	GEP	2.507	1.900	0.847	3.566	3.118	0.68
	MARS	2.421	1.845	0.857	3.756	3.211	0.688
	RF	2.417	1.859	0.858	3.856	3.220	0.679

Note: RMSE: root mean square error (°C), MAE: mean absolute error (°C), VAF: variance accounted for (-).
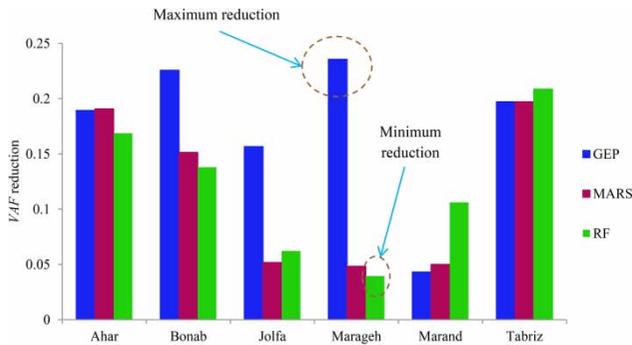
Figure 2 │ *VAF* reduction of the models by increasing prediction intervals (calculated as $\frac{VAF_{1day} - VAF_{7days}}{VAF_{1day}}$).

the model is able to select the most relevant variables among the pre-defined input-target matrix, so some of the introduced input variables may not be picked up by GEP for producing the target values. Although the input variables have been identified using the PACFs of T_{dew} time series at each station, these functions can only depict the linear correlations between successive time lags, so any nonlinear relations would be eliminated. On the other hand, the employed models can evaluate the nonlinearity among the input-target set. Table 4 presents the importance (weights) of input variables in prediction models. As seen from the weights in the table, the applied models gave different importance (weights) to the introduced input variables although some similarities are also observed. For instance, both daily GEP and MARS models gave the same importance (=zero) to the dewpoint temperature recorded at 4-days ago (T_{dew}^{i-4}) in Bonab, Jolfa, Marageh, and Tabriz. Also, weekly GEP and MARS models presented zero weight for T_{dew}^{i-3}, in Marand and for T_{dew}^{i-2} in Tabriz. The

scale for the input variables' weights in RF is different from GEP and MARS. Again, it may be stated that the way the input variables participate in the modeling (prediction) phase is different for each location, and even in models that show a complex relationship between the inputs and the predicted variables (nonlinear relations) even the variables are both the same (dewpoint). Criticizing the correlogram-based making of the input–output matrixes for prediction issues, the fact shows the dominant effect of non-linearities between the T_{dew} variations among successive time intervals (here, successive days), although the general trend might not show such important variations in terms of central-tendency statistical measures (e.g., mean, C_V, etc.). Comparing the models' performances among the stations, the highest *VAF* belonged to Ahar (with the lowest skewness value for the T_{dew} records) and the minimum values were observed for Marageh. Lower performance accuracy of the prediction models in Marageh might be linked to higher differences between the maximum and minimum air temperature values (not presented here), which made it difficult to extrapolate the dewpoint values using the previously recorded time series magnitudes. Overall, predictions relying on daily intervals performed much better than those that relied on weekly intervals, which clearly depicted the effect of the time series memory in producing the future magnitudes.

Estimation models

Table 5 sums up the global error statistics of the applied T_{dew} estimation models in the studied region. It is clear from the table that MARS models have produced the most accurate

Table 4 │ Variables importance (weights) in prediction models

	Daily prediction			Weekly prediction		
Station	GEP	MARS	RF	GEP	MARS	RF
Ahar	3,2,0,1,0	2,2,4,1,1	0.95,0.41,0.71,0.76,1	2,0,3,1,2	3,1,1,2,2	0.62,0.48,0.77,0.85,1
Bonab	2,2,0,0,1	3,1,1,0,3	1,0.43,0.48,0.56,0.64	1,1,0,1,2	3,2,1,1,1	0.6,0.67,0.93,0.93,1
Jolfa	2,1,2,0,1	3,3,2,0,3	1,0.47,0.57,0.72,0.54	2,1,1,3,0	5,0,1,2,2	0.93,0.55,0.85,0.89,1
Marageh	3,1,0,0,1	3,2,3,0,2	1,0.49,0.59,0.62,0.75	2,1,1,1,3	3,0,2,2,2	0.63,0.59,0.74,1,0.81
Marand	3,1,0,1,2	2,3,0,1,2	1,0.52,0.56,0.61,0.55	1,1,0,2,3	3,1,0,2,1	1,0.40,0.67,0.93,0.66
Tabriz	2,2,1,0,1	3,2,1,0,4	1,0.65,0.64,0.63,0.49	3,0,1,1,2	3,0,2,1,1	1,0.73,0.80,0.85,0.94

Note: The weights presented for each model belong to T_{dew}^{i-1}, T_{dew}^{i-2}, T_{dew}^{i-3}, T_{dew}^{i-4}, T_{dew}^{i-5}, respectively.

Table 5 │ Global error statistics of the estimation models' validation period

	RMSE (°C)	MAE (°C)	VAF (−)
GEP1	1.893	1.694	0.895
MARS1	1.188	0.890	0.970
RF1	1.790	1.335	0.932
GEP2	1.813	1.628	0.921
MARS2	1.152	0.861	0.971
RF2	1.637	1.224	0.945
GEP3	2.446	1.861	0.876
MARS3	1.170	0.879	0.971
RF3	1.897	1.406	0.925
GEP4	3.023	2.441	0.801
MARS4	2.667	2.089	0.845
RF4	2.943	2.346	0.813

Table 6 │ Correlation values between the applied variables and the error magnitudes of the heuristic models for the validation period

	T_{mean}	R_H	W_S	S	T_{dew}
T_{mean}	1.000				
R_H	−0.660	1.000			
W_S	0.279	−0.286	1.000		
S	0.609	−0.687	0.122	1.000	
T_{dew}	0.815	−0.135	0.179	0.298	1.000
GEP1	−0.057	−0.159	−0.015	0.043	−0.227
MARS1	0.007	0.006	0.111	−0.090	0.186
RF1	−0.154	−0.274	−0.042	0.148	−0.490
GEP2	−0.289	0.033	−0.081	−0.127	−0.393
MARS2	0.006	0.007	0.094	−0.019	0.184
RF2	−0.230	−0.321	−0.025	0.087	−0.608
GEP3	0.182	−0.179	0.041	0.043	−0.017
MARS3	0.013	−0.002	0.014	−0.074	0.188
RF3	−0.199	−0.322	−0.026	0.164	−0.576

results with the lowest *RMSE* and *MAE* as well as the highest *VAF* magnitudes, followed by RF and GEP. Among the applied input configurations, the models which rely on T_{mean}, R_H, and S (GEP2, MARS2, and RF2) presented the most accurate results. The dynamic input configuration could not improve the performance accuracy of the models. This might be due to the weak influence of the previously recorded meteorological parameters on estimating T_{dew} at the current time step. Nevertheless, further time steps might be considered to derive the chronological dependence of this parameter to the other recorded variables, which would involve a high-dimensional matrix analysis and lots of computational costs. Moreover, since the used meteorological parameters are easily measured in lots of weather stations, using preceding measurements for providing simulations of T_{dew} would not be of practical interest.

Analyzing the error statistics in Table 5 shows that introducing sunshine hours (S) as an input variable improved the performance accuracy of the employed models, while wind speed's (W_S) inclusion into the input matrix reduced the models' accuracy, which might be linked to the correlations between the input variables and T_{dew}, as can be seen in Table 6 (T_{dew} presents the highest positive correlations with T_{mean} and S). From Table 6 it can be seen that the correlations between the meteorological parameters with the errors of the heuristic models are different for the applied models. MARS models presented the highest correlations with the wind speed values for all three static input

configurations (1–3), while GEP and RF models showed the highest correlations with air temperature and relative humidity. As stated before, this can be explained by the differences between the ways the models handle the data matrix to produce the outcomes. On the other hand, the relations between the inputs and T_{dew} are different for varying range of the magnitudes of these variables. Lawrence (2005) argued that the relationships between the relative humidity and dewpoint temperature are linear for $R_H >$ 50% while they are nonlinear for $R_H <$ 50%. For a constant R_H value, T_{dew} increases with increasing the air temperature (and vice versa), while for a fixed T_{dew} value, the relative humidity increases with decreasing the air temperature. This is clearly seen from the linear dependency (correlation) values presented in Table 6, which can affect the T_{dew} modeling using different models due to the different effect of such variables on the target parameter.

Figure 3 shows the validation *VAF* values of the models split per station. The figure clearly shows the superiority of the MARS model over the GEP and RF in all studied stations and the input configurations. Further, the MARS simulations showed a similar trend in all stations, as *VAF* fluctuations are monotonously observed in the charts. In the case of the GEP models, however, the performance accuracy presented considerable fluctuations among the
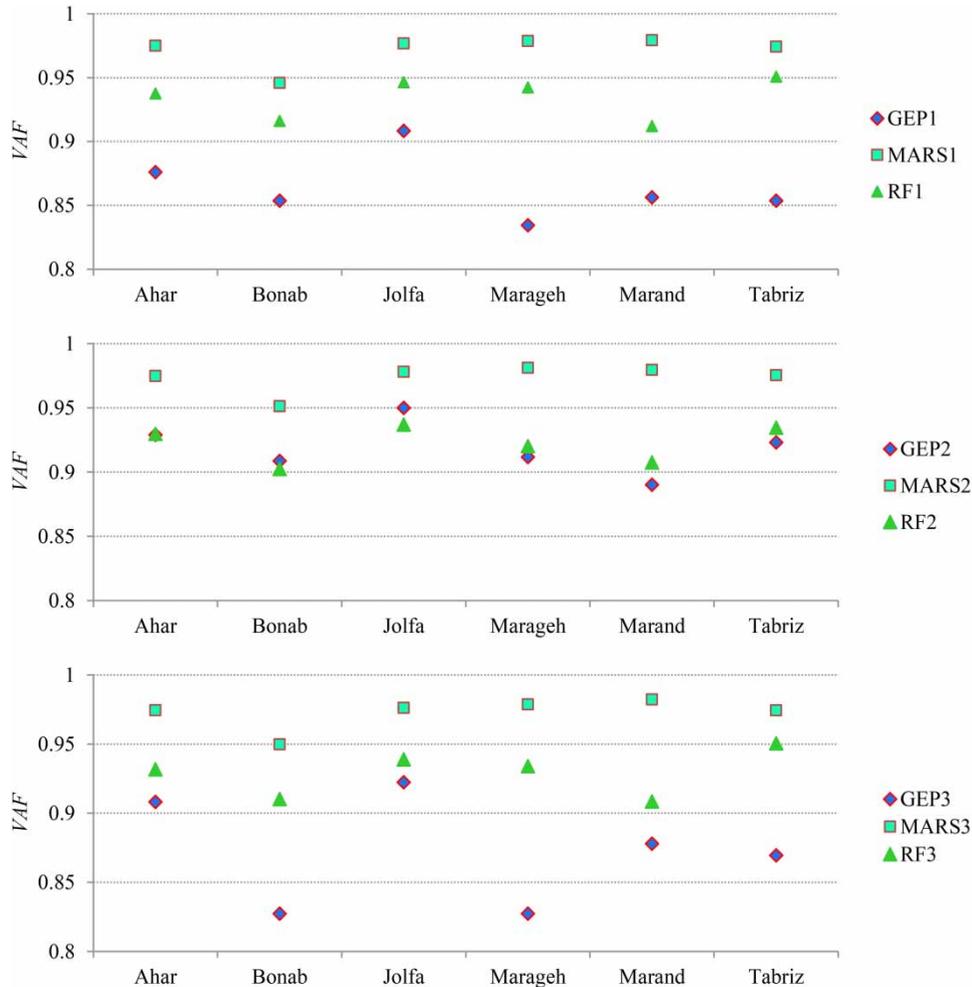
Figure 3 │ Statistical indices of the models during the validation period-split per station.

stations. Although relative humidity and sunshine hours' trends and mean values in the studied stations were very close, noticeable differences between the mean and C_V values of T_{dew} in these stations would create such discrepancy in the GEP performance accuracy. However, as it can be seen from the figure, MARS is not sensitive to such discrepancies. Table 7 presents the mathematical expressions of the GEP2 and MARS2 models. Different responses of the models to the data in different locations can be explained taking into consideration the principal differences as well as the basic functions considered for the GEP and MARS models, as stated before. Finally, although the data have been pooled together to make a general input-target matrix, the obtained results suffer from information being obtained from the same stations for the

testing, data from which have been incorporated in the models' training. This might be remedied by introducing exogenous inputs from the nearby stations as inputs for producing the T_{dew} values of the target stations.

Nevertheless, the current study aimed at providing general heuristic-based functions that capture the process physics based on the training, testing, and validation patterns. In the current paper, a single chronological data set assignment was employed which might be established in future research via data set scrutinizing approaches to involve all the possible patterns in train–test phases. Nonetheless, a complete spatial scanning of the available patterns using more weather stations might be considered to examine the generalizability of the heuristic models in estimating dewpoint temperature using ancillary data.

Table 7 | Mathematical expression of GEP2 and MARS2 models

GEP2	$T_{dew} = Arctg[\cos^3(S) - S + T_{mean} + 2.919] + 0.0001 T_{mean}[T_{mean} + R_H]^2 - 6.987$
MARS2	$T_{dew} = 5.98706909541185e + 000 + 9.41021229696271e - 001*\max(0; T_{mean} - 1.58400000000000e + 001) - 9.36347451296909e - 001*\max(0; 1.58400000000000e + 001 - T_{mean}) + 4.44731609739038e - 001*\max(0; R_H - 4.70000000000000e + 001) - 4.86429176276841e - 001*\max(0; 4.70000000000000e + 001 - R_H) - 1.73049333411989e - 001*\max(0; R_H - 3.65000000000000e + 001) - 8.73281480638373e - 002*\max(0; R_H - 6.50000000000000e + 001) - 1.03949596102562e - 001*\max(0; S - 6.00000000000000e + 000) - 1.05726621642050e - 001*\max(0; T_{mean} - 2.46800000000000e + 001) - 4.98258058329862e - 002*\max(0; T_{mean} - 2.90000000000000e + 000) - 4.39711149166199e - 001*\max(0; R_H - 8.95000000000000e + 001) - 6.62147495338194e - 001*\max(0; T_{mean} - 3.39600000000000e + 001) + 4.03627060993806e - 001*\max(0; R_H - 9.05000000000000e + 001)$

CONCLUSIONS

The ability of GEP, MARS, and RF models in predicting and estimating dewpoint temperature (T_{dew}) was evaluated using daily records of meteorological parameters of six weather stations in Iran. First, the T_{dew} values were predicted using its previously recorded values in daily and weekly prediction intervals. The results showed that, as could be anticipated, +1 day predictions provided more accurate results than +7 day predictions in all studied locations and employed models. Moreover, GEP surpassed both the MARS and RF models in all locations and both prediction intervals. Regarding the estimation models (where air temperature, humidity, sunshine hours, and wind speed were used to feed the employed models), the models relied on air temperature, relative humidity, and sunshine hours to provide the most accurate outputs. In all cases of the estimation models, MARS outperformed GEP and RF models. No strong relation was observed between T_{dew} and previously recorded meteorological parameters. The present paper showed that the MARS model gave promising results with regard to the most commonly used GEP or RF models in estimation of T_{dew}. However, only six stations were considered in the current research and a double data set assignment was used for assessing the performance accuracy. Further studies will be needed using stations of different climatological context and through utilizing more robust data management scenarios, e.g., k-fold testing, for assessing the temporal and spatial generalizability of the models. These might be interesting subjects for future studies.

REFERENCES

Abdel-Aal, R. E. 2004 Hourly temperature forecasting using abductive networks. *Engineering Applications of Artificial Intelligences* **17**, 543–556.

Adamowski, J., Chan, H. F., Prasher, S. O. & Sharda, V. N. 2012 Comparison of multivariate adaptive regression splines with coupled wavelet transform artificial neural networks for runoff forecasting in Himalayan micro-watersheds with limited data. *Journal of Hydroinformatics* **14** (3), 731–744.

Allen, R. G., Pereira, L. S., Raes, D. & Smith, M. 1998 *Crop Evapotranspiration-Guidelines for Computing Crop Water Requirements*. FAO Irrigation and Irrigation Drainage Paper 56, FAO, Rome, Italy.

Andres, J. D., Lorca, P., de Cos Juez, F. J. & Sánchez-Lasheras, F. 2010 Bankruptcy forecasting: a hybrid approach using Fuzzy c-means clustering and Multivariate Adaptive Regression Splines (MARS). *Expert Systems with Applications* **38**, 1866–1875.

Atzema, A. J., Jacobs, A. F. G. & Wartena, L. 1990 Moisture distribution within a maize crop due to dew. *Netherlands Journal of Agricultural Science* **38** (2), 117–129.

Bilgili, M. & Sahin, B. 2010 Prediction of long-term monthly temperature and rainfall in Turkey. *Energy Sources, Part A* **32**, 60–71.

Breiman, L. 2001 Random forests. *Machine Learning* **45**, 5–32.

Deschaine, L. M. 2014 *Decision Support for Complex Planning Challenges: Combining Expert Systems, Engineering-Oriented Modeling, Machine Learning, Information Theory, and Optimization Technology*. Chalmers University of Technology, Sweden, p. 233.

Ferreira, C. 2001a Gene expression programming in problem solving. In: *6th Online World Conference on Soft Computing in Industrial Applications (Invited Tutorial)*.

Ferreira, C. 2001b Gene expression programming: a new adaptive algorithm for solving problems. *Complex Systems* **13** (2), 87–129.

Friedman, J. H. 1991 Multivariate adaptive regression splines. *Annals of Statistics* **19**, 1.

Hastie, T., Tibshirani, R. & Friedman, J. 2009 *The Elements of Statistical Learning*. Springer, New York.

Kisi, O. & Shiri, J. 2011 Precipitation forecasting using wavelet-genetic programming and wavelet-neuro-fuzzy conjunction models. *Water Resources Management* **25** (13), 3135–3152.

Kisi, O. & Shiri, J. 2014 Prediction of long-term monthly air temperature using geographical inputs. *International Journal of Climatology* **34**, 179–186.

Kisi, O., Kim, S. & Shiri, J. 2013 Estimation of dew point temperature using neuro-fuzzy and neural network techniques. *Theoretical and Applied Climatology* **114**, 365–373.

Lawrence, M. G. 2005 The relationship between relative humidity and the dewpoint temperature in moist air. *Bulletin of the American Meteorological Society* **86** (2), 225–230.

Mahmood, R. & Hubbarad, K. G. 2005 Assessing bias in evapotranspiration and soil moisture estimates due to the use of modeled solar radiation and dew point temperature data. *Agricultural and Forest Meteorology* **130**, 71–84.

Mahmood, R., Hubbard, K. G., Leeper, R. D. & Foster, S. A. 2008 Increase in near-surface atmospheric moisture content due to land use changes: evidence from the observed dewpoint temperature data. *Monthly Weather Reviews* **136**, 1554–1561.

Pour Ali Baba, A., Shiri, J., Kisi, O., FakheriFard, A., Kim, S. & Amini, A. 2013 Estimating daily reference evapotranspiration using available and estimated climatic data by adaptive neuro-fuzzy inference system (ANFIS) and artificial neural network (ANN). *Hydrology Research* **44** (1), 131–146.

Savoy, P. & Mackay, D. S. 2015 Modeling the seasonal dynamics of leaf area index based on environmental constrains to canopy development. *Agricultural and Forest Meteorology* **200**, 46–56.

Shank, D. B., Hoogenboom, G. & McClendon, R. W. 2008 Dewpoint temperature prediction using artificial neural networks. *Journal of Applied Meteorology and Climatology* **47**, 1757–1769.

Sharda, V., Prasher, S. O., Patel, R. M., Ojavasi, P. R. & Prakash, C. 2006 Modeling runoff from middle Himalayan watersheds employing artificial intelligence techniques. *Agriculture Water Management* **83**, 233–242.

Shiri, J. 2017 Evaluation of FAO56-PM, empirical, semi-empirical and gene expression programming approaches for estimating daily reference evapotranspiration in hyper-arid regions of Iran. *Agricultural Water Management* **188**, 101–114.

Shiri, J., Kim, S. & Kisi, O. 2014 Estimation of daily dew point temperature using genetic programming and neural networks approaches. *Hydrology Research* **45** (2), 165–181.

Slatyer, R. O. 1967 *Plant–Water Relationships*. Academic Press, London.

Smith, B. A., McClendon, R. W. & Hoogenboom, G. 2005 An enhanced artificial neural network for air temperature prediction. *Proceedings of World Academy of Science, Engineering and Technology (PWASET)* **7**, 7–12.

Wallace, J. M. & Hobbs, P. V. 2006 *Atmospheric Science: An Introductory Survey*, 2nd edition. Academic Press, Canada, p. 504.

Went, F. W. 1955 Fog, mist dew and other sources of water. In: *Yearbook Agriculture*. US Department of Agriculture, pp. 103–109.