





Article

# Estimating Daily Dew Point Temperature Using Machine Learning Algorithms

Sultan Noman Qasem <sup>1,2</sup>, Saeed Samadianfard <sup>3</sup>, Hamed Sadri Nahand <sup>3</sup>, Amir Mosavi <sup>4,5,6</sup>, Shahaboddin Shamshirband <sup>7,8,\*</sup> and Kwok-wing Chau <sup>9</sup>

<sup>1</sup> Computer Science Department, College of Computer and Information Sciences, Al Imam Mohammad Ibn Saud Islamic University (IMSIU), Riyadh 11432, Saudi Arabia; SNMohammed@imamu.edu.sa

<sup>2</sup> Computer Science Department, Faculty of Applied Sciences, Taiz University, Taiz, Yemen

<sup>3</sup> Department of Water Engineering, University of Tabriz, Tabriz 5166616471, Iran; s.samadian@tabrizu.ac.ir (S.S.); hamed.sadri7@yahoo.com (H.S.N.)

<sup>4</sup> School of the Built Environment, Oxford Brookes University, Oxford OX3 0BP, UK; amirhosein.mosavi@qut.edu.au

<sup>5</sup> Institute of Automation, Kando Kalman Faculty of Electrical Engineering, Obuda University, 1034 Budapest, Hungary

<sup>6</sup> The Queensland University of Technology, Institute of Health and Biomedical Innovation, 60 Musk Avenue, Queensland 4059, Australia

<sup>7</sup> Department for Management of Science and Technology Development, Ton Duc Thang University, Ho Chi Minh City, Vietnam

<sup>8</sup> Faculty of Information Technology, Ton Duc Thang University, Ho Chi Minh City, Vietnam

<sup>9</sup> Department of Civil and Environmental Engineering, Hong Kong Polytechnic University, Hong Kong, China; dr.kwok-wing.chau@polyu.edu.hk

\* Correspondence: shahaboddin.shamshirband@tdtu.edu.vn

Received: 4 February 2019; Accepted: 18 March 2019; Published: 20 March 2019



**Abstract:** In the current study, the ability of three data-driven methods of Gene Expression Programming (GEP), M5 model tree (M5), and Support Vector Regression (SVR) were investigated in order to model and estimate the dew point temperature (DPT) at Tabriz station, Iran. For this purpose, meteorological parameters of daily average temperature (T), relative humidity (RH), actual vapor pressure ( $V_p$ ), wind speed (W), and sunshine hours (S) were obtained from the meteorological organization of East Azerbaijan province, Iran for the period 1998 to 2016. Following this, the methods mentioned above were examined by defining 15 different input combinations of meteorological parameters. Additionally, root mean square error (RMSE) and the coefficient of determination ( $R^2$ ) were implemented to analyze the accuracy of the proposed methods. The results showed that the GEP-10 method, using three input parameters of T, RH, and S, with RMSE of  $0.96^\circ$ , the SVR-5, using two input parameters of T and RH, with RMSE of 0.44, and M5-15, using five input parameters of T, RH,  $V_p$ , W, and S with RMSE of 0.37 present better performance in the estimation of the DPT. As a conclusion, the M5-15 is recommended as the most precise model in the estimation of DPT in comparison with other considered models. As a conclusion, the obtained results proved the high capability of proposed M5 models in DPT estimation.

**Keywords:** dew point temperature; prediction; machine learning; meteorological parameters; statistical analysis; big data; gene expression programming (GEP); deep learning; forecasting; M5 model tree; support vector regression (SVR); hydrological model; hydroinformatics; hydrology

## 1. Introduction

Dew point temperature (DPT) is defined as the temperature in which air becomes liquid water due to the high concentration of water molecules. Precise and accurate estimation of DPT has a

significant role in solving agricultural problems, such as calculating the amount of available moisture in the air and estimating the near surface humidity [1]. DPT and relative humidity are commonly used to measure the air humidity level [2]. The DPT can also be used to estimate the temperature of crops considering glaciation [3]. Many studies have paid attention to the accurate estimation of DPT using regression methods. However, data-driven methods such as Gene Expression Programming (GEP) and Neuro-Fuzzy Inference System (ANFIS) have been developed to identify optimal functions and modeling for complex phenomena. In this regard, several studies have been carried out on the application of the mentioned methods in meteorological studies [4–13]. Shiri [2] compared the capabilities of the artificial neural network (ANN) and GEP to estimate the DPT using meteorological parameters at Seoul and Incheon stations, located in South Korea. They used two management scenarios: In the first scenario, the meteorological information of each station was used to estimate the DPT of the same station; in the second scenario, they used the meteorological information of adjacent stations. Their results showed that in both scenarios GEP was more accurate than ANN. Also, the application of the second scenario showed that GEP had more accurate results in estimating the DPT values of Seoul stations using Incheon station parameters. They also reported that the DPT values at Seoul Station could be estimated using the average temperature and relative humidity of the Incheon station with proper accuracy. Deka et al. [14] examined the ability of a support vector machine (SVM), ANN, and Extreme Learning Machine (ELM) to estimate DPT at two stations in Iran. They showed that the results of the ELM model were more similar to observed DPT at the two mentioned stations. In other research, Zounemat-Kermani [15] implemented two methods of multiple linear regression (MLR) and Levenberg–Marquardt algorithm (LMA) in the artificial neural network (LMA–ANN) in order to estimate DPT values at Ontario Station, Canada. The results of the LMA–ANN model had an appropriate match with observational data. Additionally, Jia et al. [16] investigated dew formation. For this purpose, they used meteorological data of average temperature, sunny hours, wind speed, saturated vapor pressure, relative humidity, and DPT values of three stations of Dagot, Pohang, and Ulsan, South Korea. They reported that the effects of sunny hours, wind speed, and saturated vapor pressure were lower than other parameters. Therefore, it was possible to estimate the DPT using average temperature and relative humidity. Attar et al. [17] used GEP, multivariate adaptive regression splines (MARS), and SVM models to estimate the DPT in arid regions of Iran. Using the meteorological data of 13 synoptic stations during the 55 years (1996 to 2014), and by defining 50 different scenarios. They concluded that the MARS model offers more accurate results than other studied models. In a similar study, Mehdizadeh et al. [18] estimated the DPT values in Tabriz and Urmia cities, in the northwest of Iran, using the GEP method. They defined three scenarios: A parameters-based scenario, a temperature-based scenario, and a periodicity-based scenario considered the meteorological parameters of minimum, maximum, and mean air temperature, actual vapor pressure, and atmospheric pressure. Their results showed that the actual vapor pressure is the most effective meteorological parameter in estimating the DPT in the study area.

Therefore, over the last decade, researchers have tried to estimate DPT values with suitable accuracy. For this reason, the main purpose of the current study was to implement three data-driven methods of GEP, M5, and SVR in order to improve the estimation accuracy and develop some mathematical formulations for obtaining precise estimations of DPT values using explicit formulations. To the best of our knowledge, the application of M5 has not been reported in the literature. In other words, the goals of the study were (i) evaluating the performance of the models above in the estimation of DPT, and (ii) investigating the role of climatic parameters estimation DPT values. The rest of the paper is structured as follows: Section 2 describes implemented methods, evaluation parameters, and characteristics of the study area. Additionally, Section 3 discussed the obtained results and, finally, the conclusion is presented in Section 4.

## 2. Study Area

The Tabriz synoptic station is one of the oldest meteorological stations belonging to the Iranian Meteorological Organization, located in East Azerbaijan province with a latitude of  $38^{\circ} 05' N$  and longitude of  $46^{\circ} 17' E$  and an elevation of 1364 m above sea level. In the current study, the DPT values of Tabriz station in the period of 1998 to 2016 were utilized to evaluate the precision of the considered models. The geographic location of the study area is shown in Figure 1.



Figure 1. Location of the study area [19].

## 3. Materials and Methods

### 3.1. Support Vector Regression (SVR)

SVM, which is established by statistical learning theory, has been broadly applied for identifying complex patterns of different environmental phenomena. Moreover, SVR, as one of the types of SVM, has been used previously for regression problems [20,21]. The basics of the support vector regression are presented briefly below.

If the available data can be linearly split, they can be distinguished using cloud-computing system data. In some cases, data cannot be linearly separated. In these cases, the data is mapped to a larger dimensional space, after which, the data are separated. Figure 2B shows the non-linear mapping of  $\varphi(0): \mathbb{R}^n \rightarrow \mathbb{R}^nh$ , in which, the nonlinear training data are mapped to a higher-dimensional space. Following that, a linear relationship is obtained between the input and output data in the converted space. The general form of the linear relation is as follows:

$$f(x) = W^T \varphi(x) + b \quad (1)$$

where,  $f(x)$  is the estimated variable,  $W^T$  is the transpose of the vector of coefficients and  $b$  is a constant coefficient. The SVM is based on minimizing the value of empirical error.

$$R_{emp}(f) = \frac{1}{N} \sum_{i=1}^N \Theta_{\epsilon} \left( y_i, W^T \varphi(x) + b \right) \quad (2)$$

where,  $\Theta_\epsilon(y, f(x))$  is the  $\epsilon$ -insensitive error function defined as Equation (3):

$$\Theta_\epsilon(y, f(x)) = \begin{cases} |f(x) - y| - \epsilon & \text{if } |f(x) - y| \geq \epsilon \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

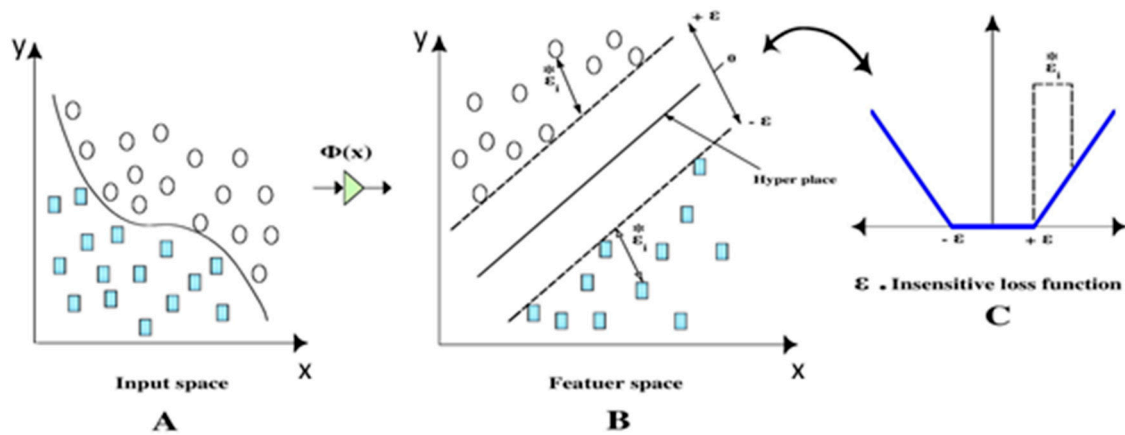


Figure 2. (A) Initial Space. (B) Feature space. (C) Insensitive error function.

In addition, the function  $\Theta_\epsilon(y, f(x))$  is used to find the optimal separator plain in a high dimension space (which may have infinitive dimensions). In a space with a high dimension, an optimal separator plate will maximize the distance between training data. Following this, the SVR model minimizes the general error concerning the constraints. These limitations presented in Equations (4) to (8).

$$\min_{w,b,\zeta_i^*,\zeta_i} R_\epsilon(W, \zeta_i^*, \zeta_i) = \frac{1}{2}W^T W + C \sum_{i=1}^N (\zeta_i^* + \zeta_i) \quad (4)$$

where,  $\zeta_i$  is an error greater than  $-\epsilon$ ,  $\zeta_i^*$  is the error greater than  $+\epsilon$ , and  $c$  is a constant.

$$y_i - W^T \varphi(x_i) - b \leq \epsilon + \zeta_i^*, \quad i = 1, 2, \dots, N \quad (5)$$

$$-y_i + W^T \varphi(x_i) + b \leq \epsilon + \zeta_i, \quad i = 1, 2, \dots, N \quad (6)$$

$$\zeta_i^* \geq 0, \quad i = 1, 2, \dots, \quad (7)$$

$$\zeta_i \geq 0, \quad i = 1, 2, \dots, N \quad (8)$$

After obtaining the optimal separator plate, the vector of the coefficients  $w$  obtained as Equation (9).

$$W = \sum_{i=1}^N (\beta_i^* - \beta_i) \varphi(x_i) \quad (9)$$

in which,  $\beta_i^*$  and  $\beta_i$  are calculated by applying quadratic programming and Lagrange coefficients, respectively. Finally, the SVR function is obtained as Equation (10) in a two-dimensional space, as follows:

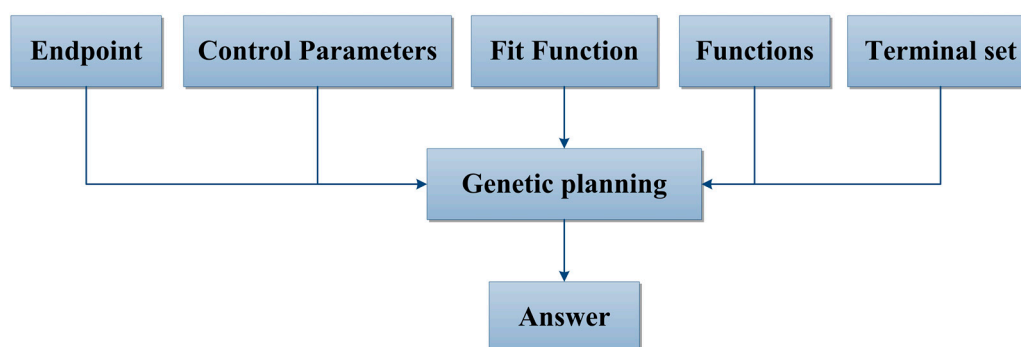
$$f(x) = \sum_{i=1}^N (\beta_i^* - \beta_i) K(X_i, X) + b \quad (10)$$

$K(X_i, X_j)$  is called the Kernel function and is equal to the inner product of the two vectors of  $X_i$  and  $X_j$  in a space with a high dimension. In the SVR method, several Kernel functions are used, including the polynomial Kernel function, the normalized polynomial Kernel function, the radial-base Kernel function, and the Pearson Kernel function [22].

### 3.2. Gene Expression Programming

GEP, which is a variant of genetic programming (GP), is a generalized genetic algorithm. GEP is considered to be a circular method based on Darwin’s theory of evolution. GEP at the beginning of the process does not take into account the functional relationship and can optimize the structure of the model and its components [23]. Unlike the genetic algorithm, GEP acts on the tree structure of formulas rather than a series of binary numbers. The tree structures created from the set of functions (mathematical operators used in formulas) and terminals (problem variables and constant numbers). Before the implementation of GEP, the following factors are determined in the following:

1. Terminal set (problem variables, randomized constant numbers),
2. The mathematical operators used in formulas,
3. Select the fitness function (RMSE, MSE, MAE, . . . ) to measure the fitness of the formulas,
4. Select the parameters controlling the implementation of the program (population size, the probability associated with the use of genetic operators and other details related to the implementation of the program),
5. The completion benchmark and the presentation of the results of the program implementation (the number of new population production, the determination of the specified amount for the fitness of the formulas if the fitness level is equal to or greater than that value stopped) [24]. The outlines of the mentioned steps are shown in Figure 3. Moreover, the parameters used in the implementation of the GEP presented in Table 1.



**Figure 3.** The general form of the initial steps of gene expression programming (GEP) (Alvisi, Mascellani, Franchini, and Bardossy, 2005).

**Table 1.** Parameters used in the GEP method.

Parameter	Quantity
Functions used	$+, -, \times, \div, \sqrt{\quad}, \ln(x), exp, r, Sin, Cos, Arctan$
Number of chromosomes	30
Number of genes	3
Linking function	Sum
Jump speed	0
Mutation rate	0.044
Inversion rate version	0.1
One-point recombination rate T	0.3
Two-point recombination rate two points	0.3
Gene recombination rate the gene	0.1
Gene transposition rate	0.1

### 3.3. M5 Model Tree

The M5 model tree is a subset of the machine learning and data mining methods developed by Quinlan [25]. Data mining refers to the process of searching and discovering various models,



summarizing and obtaining quantities from the collections. Learning machine and data mining methods have the ability to discover data patterns, semi-automatically. The main reasons for using a model tree are as follows:

1. The model tree is directly related to estimative variables; therefore, the results of the model are easy to understand.
2. Model trees are non-parametric, and there is no user intervention on them.
3. The output of the model has a high degree of accuracy that can be compared to other models.

The structure of the model trees includes roots, branches, nodes, and leaves. The nodes are represented by a circle and the branches represent the connection between the nodes [26]. The generation of the model tree structure consists of different steps of creating a tree and pruning it. In the first step, an inferential algorithm or division criterion is used for the production of the tree. The decision criterion for the M5 is the standard deviation of the class values, which is calculated as a quantity of error to a node and calculates the expected reduction in this error as the result of the test of each attribute in that node. The standard deviation ratio (SDR) is calculated as follows (Equation (11)):

$$SDR = sd(T) - \sum \frac{|T_i|}{|T|} sd(T_i) \quad (11)$$

In which, T is a collection of input samples to each node,  $T_i$  is a subset of the samples that have the  $i$  output of the potential series and  $sd$  denotes the standard deviation [27]. As a result of the branching process, the data in the child nodes has a lower standard deviation than the parent node and is purer. After maximizing all possible branches, the M5 selects an attribute that maximizes the expected reduction. This division forms a large quasi-tree structure, which causes over-fitting. To overcome the mentioned problem, the tree should be pruned. This is done by replacing a sub-tree with a leaf; therefore, the second step in the design of a model tree is to prune the grown tree and to replace the sub-trees with linear regression functions. This method for creating the model tree divides the space of the input parameters into smaller areas or sub-areas. In each area, a linear regression model is fitted. After obtaining a linear model, the simplification of the model can minimize the estimation error by deleting the model parameters [28].

#### 3.4. Evaluation Criteria

Error values between observed and estimated data were determined by RMSE and coefficient of determination ( $R^2$ ). The RMSE was used to evaluate the accuracy of the estimations. The consistent estimated values of the model lead to the minimization of the statistical index. Furthermore,  $R^2$  is a statistical tool for determining the type and degree of the relationship of a variable with other variables. This coefficient varies from 0 to 1; when there is no relationship between two variables, its value is equal to zero [29,30]. Furthermore, Taylor diagrams [31] were used to check the accuracy of the mentioned models. It is noteworthy that Taylor suggested a diagram, in which measured parameters and some characteristics of the model are summed up, coincidentally. Surprisingly, Taylor diagrams utilize several points on a polar plot for comparing the accuracy of measured and estimated values. In these diagrams, the coefficient of determination and normalized standard deviation are represented by an azimuth angle and radial distances from the base point, respectively [31,32].

## 4. Results and Discussion

In order to reach the research objectives, daily average temperature, relative humidity, actual vapor pressure, wind speed, and sunny hours at the Tabriz synoptic station were collected from the Meteorological Organization of East Azerbaijan province, Iran during the period 1998 to 2016. The statistical characteristics of the implemented data are presented in Table 2. There is no basic way of separating training and testing data. For example, the study of Kurup and Dudani [33] used a total of 63% of their data for model development, whereas Samadianfard et al. [4] and Samadianfard et al. [7]

used 67% of total data, and Deo et al. [34] used 70% of total data to develop their models. Thus, to develop the studied GEP, M5, and SVR models for estimation DPT, we divided the data into training (67%) and testing (33%). Therefore, the accuracy of the models in estimating DPT evaluated through Taylor diagrams. Additionally, the effects of considered meteorological parameters were inspected by defining 15 different input combinations (Table 3).

**Table 2.** Statistical characteristics of the meteorological data.

CC	Skewness	Standard Deviation	Max	Min	Mean	Parameter
0.59	−0.13	10.26	34.0	−15.0	13.3	T <sub>avg</sub> (°C)
0.12	0.24	17.45	96.0	10.0	50.0	RH (%)
0.01	0.13	4.33	880.0	848.3	864.3	V <sub>p</sub> (kpa)
0.21	0.86	1.57	13.0	0.00	3.40	W (m/s)
0.23	−0.71	3.78	14.0	0.00	7.90	S (h)

**Table 3.** Different combinations of input parameters in the estimation of dew point temperature.

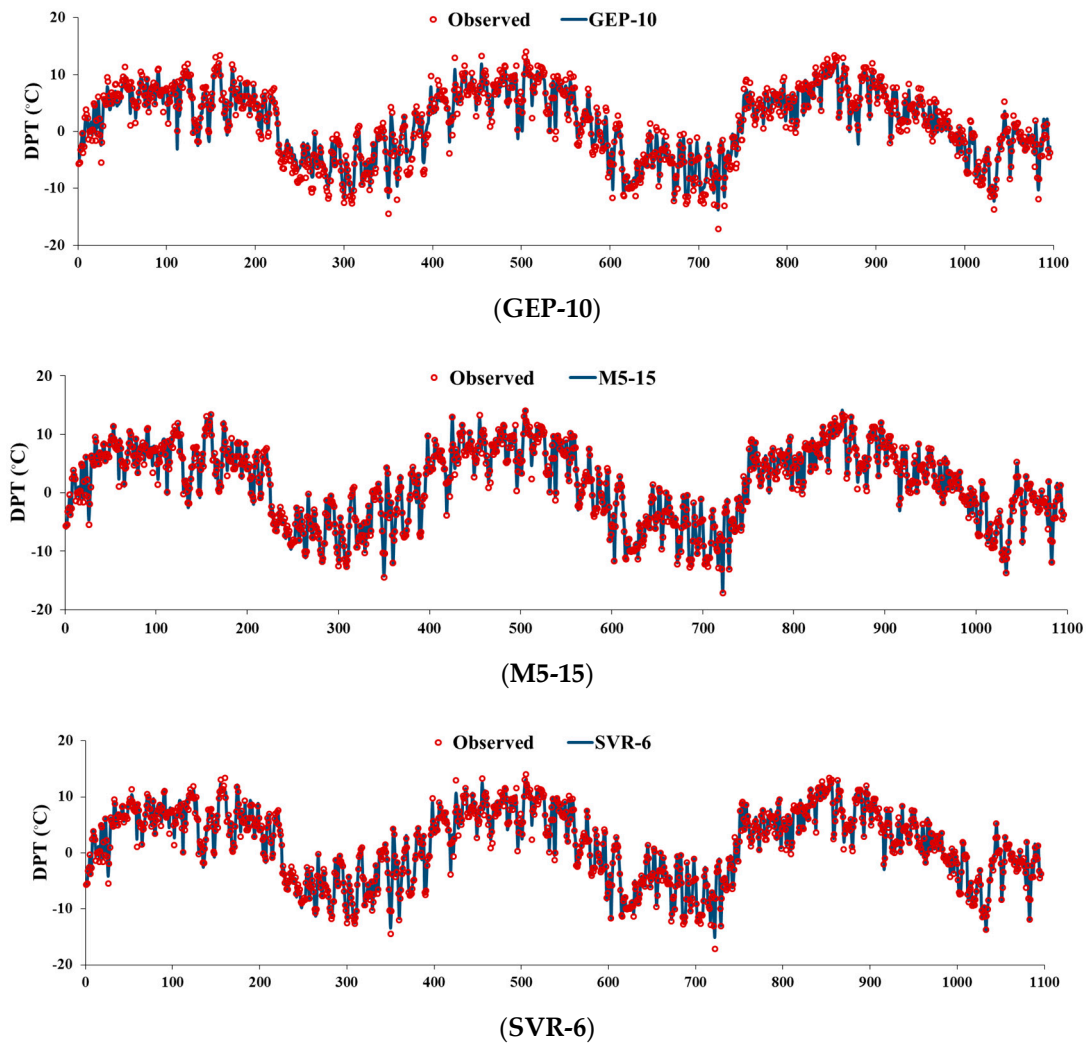
Number	Input Parameters	Number	Input Parameters
1	T	9	T, S
2	RH	10	T, S, RH
3	V <sub>p</sub>	11	T, S, V <sub>p</sub>
4	W	12	T, S, W
5	S	13	T, S, RH, V <sub>p</sub>
6	T, RH	14	T, S, RH, W
7	T, V <sub>p</sub>	15	T, S, RH, W, V <sub>p</sub>
8	T, W		

After performing the computations for different input combinations, the accuracy of the considered models was determined in the testing phase based on the statistical criteria (Equations (9) and (10)) and Taylor diagrams. The obtained results are presented in Table 4.

**Table 4.** Evaluation of the performance of GEP, M5 model tree (M5), and support vector regression (SVR) models in the testing phase.

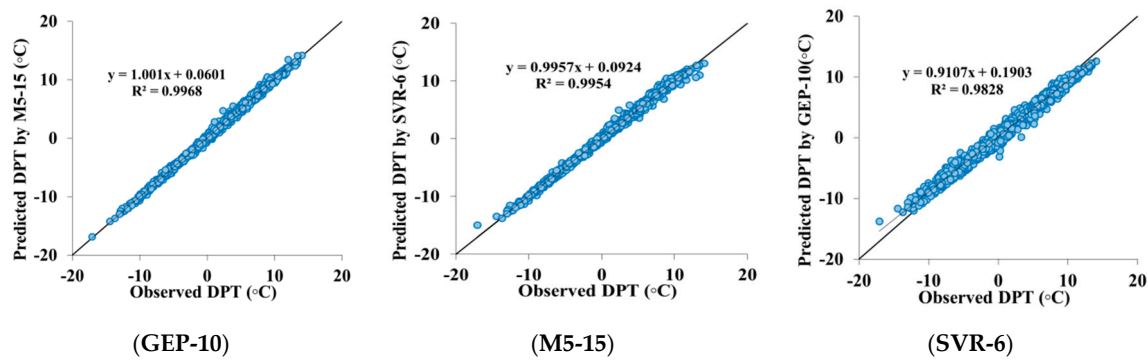
Scenarios	GEP		M5		SVR	
	RMSE (Degree)	R <sup>2</sup>	RMSE (Degree)	R <sup>2</sup>	RMSE (Degree)	R <sup>2</sup>
1	3.40	0.719	3.36	0.727	3.37	0.724
2	6.20	0.087	6.11	0.092	6.15	0.102
3	5.74	0.241	5.58	0.243	5.59	0.247
4	5.90	0.168	5.85	0.173	5.98	0.158
5	5.20	0.403	5.77	0.188	5.76	0.187
6	1.56	0.935	0.40	0.996	0.44	0.996
7	3.44	0.714	3.34	0.731	3.33	0.731
8	3.50	0.701	3.30	0.734	3.30	0.736
9	3.18	0.751	2.98	0.787	3.00	0.783
10	0.96	0.902	0.40	0.996	0.46	0.994
11	3.10	0.760	2.96	0.788	2.99	0.784
12	3.21	0.748	2.90	0.795	2.91	0.796
13	2.57	0.840	0.38	0.996	0.54	0.994
14	1.05	0.974	0.38	0.996	0.47	0.994
15	2.60	0.835	0.37	0.996	0.55	0.989

As can be seen in Table 4, GEP-10 with RMSE of 0.96 degrees and  $R^2$  equal to 0.902 with the parameters of T, RH, and S shows better performance compared to GEP models. However, SVR-6 with RMSE of 0.44 degree and  $R^2$  of 0.996 presents more accurate estimation compared to the SVR models. Furthermore, the best estimation of the DPT, based on M5 models, was related to M5-15 with RMSE of 0.37 degree and  $R^2$  of 0.996 and using all considered meteorological parameters as the input. In other words, a comprehensive comparison between the mentioned models exhibited that M5-15 had the best performance in estimation DPT values by using input combinations of T, S, RH, W,  $V_p$ . After selecting the most accurate models for estimation DPT values, the time series plots and scatterplots are finalized and illustrated in the Figures 4 and 5.



**Figure 4.** Observed and estimated values of dew point temperature (DPT) with the best models including (GEP-10), (M5-15), and (SVR-6).

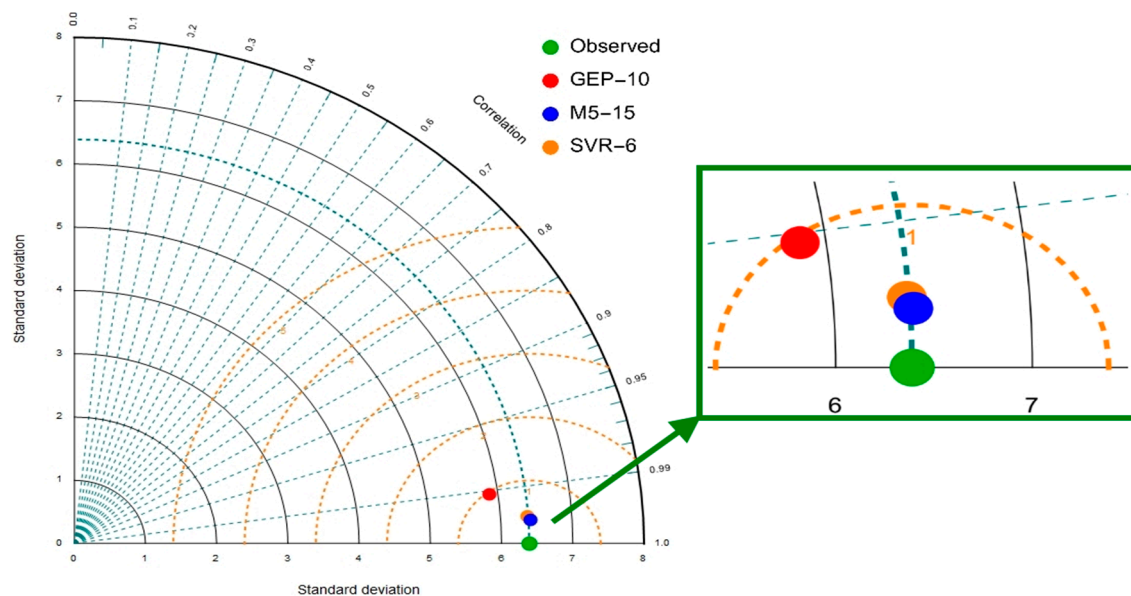




**Figure 5.** The scatter plots of observed and estimated DPT values with most precise models including (GEP-10), (M5-15), and (SVR-6).

It can be comprehended from Table 4 and Figure 4 that the estimation accuracy of the M5-15 was higher than the GEP-10 and SVR-6. The above-mentioned conclusion, regarding the high accuracy of the M5-15 model in estimation the DPT for Tabriz station, can be deduced from Figure 5. In this figure, it can be seen that the distribution of the points around the bisector line in the M5-15 model was less than the corresponding points of GEP-10 and SVR-6.

Furthermore, Taylor charts were used to examine the standard deviation and correlation values among estimated and measured DPT values for the GEP, M5, and SVR models with different input parameters. Taylor diagrams for models mentioned above are shown in Figure 6. The length of the space from the reference point (a green color point) to each point is defined as centered RMSE [31]. Therefore, the most accurate model has a minimum distance between the green point and its corresponding point. According to Figure 6, M5-15 (a blue color point) offered the most accurate estimations of DPT values at Tabriz station.



**Figure 6.** Taylor diagrams of estimated DPT values in the test period.

One of the advantages of GEP and the M5 models, in comparison with other data-driven methods, is their ability to provide explicit relationships to calculate the output parameter. Therefore, for the

current study, Equation (12) was obtained for estimation DPT values using GEP-10 as the most accurate GEP model.

$$T_{dew} = \sqrt{RH - 9.7} - 9.7 - e^{\left(\frac{S}{RH}\right)} + \frac{T}{\sqrt{\left(e^{\left(\frac{3.4}{RH}\right)}\right)^3}} + \sqrt{RH - 9.7} - 9.7 - \sqrt{\frac{S}{RH}} \quad (12)$$

Additionally, the list of linear equations (presented in Table 5) was the outcome of M5-15 estimation of the DPT values using meteorological parameters of T, S, RH, W, V<sub>p</sub>.

**Table 5.** Obtained Equation from the M5 model tree for scenario No. 15.

Obtained Equation from the M5 Model Tree	Conditions of Input	
	RH	T
$T_{dew} = 0.9196*T + 0.2162*RH - 0.0032*EA + 0.0798*W + 0.0124*S - 17.7731$	$RH \leq 65.5$	$T \leq -7.95$
$T_{dew} = 0.9311*T + 0.189*RH - 0.0032*EA + 0.0399*W + 0.0127*S - 15.7956$	$RH > 65.5$	$T \leq -7.95$
$T_{dew} = 1.0015*T + 0.248*RH - 0.0055*EA + 0.0959*W + 0.0316*S - 17.3586$	$RH \leq 61.5$	$-7.95 < T \leq -5.05$
$T_{dew} = 0.915*T + 0.1943*RH - 0.0043*EA + 0.0266*W + 0.0049*S - 15.237$	$61.5 < RH \leq 73.5$	$-7.95 < T \leq -5.05$
$T_{dew} = 0.9199*T + 0.1804*RH + 0.0016*EA + 0.0293*W - 0.0011*S - 19.2341$	$RH > 73.5$	$-7.95 < T \leq -5.05$
$T_{dew} = 0.8274*T + 0.2887*RH - 0.0005*EA + 0.076*W - 0.0087*S - 23.8787$	$RH \leq 50.5$	$-5.05 < T \leq -0.15$
$T_{dew} = 0.9152*T + 0.2308*RH - 0.0005*EA + 0.0213*W - 0.0276*S - 20.454$	$50.5 < RH \leq 61.5$	$-5.05 < T \leq -0.15$
$T_{dew} = 0.9186*T + 0.1929*RH + 0.0033*EA + 0.0302*W - 0.0131*S - 21.4432$	$61.5 < RH \leq 74.5$	$-5.05 < T \leq -0.15$
$T_{dew} = 0.9565*T + 0.1694*RH - 0.001*EA + 0.0095*W - 0.0093*S - 15.8692$	$RH > 74.5$	$-5.05 < T \leq -0.15$
$T_{dew} = 0.8804*T + 0.3409*RH + 0.0005*EA + 0.0036*W - 0.0043*S - 26.7889$	$RH \leq 48.5$	$-0.15 < T \leq 5.75$
$T_{dew} = 0.9219*T + 0.2566*RH + 0.0111*EA + 0.0036*W - 0.030*S - 31.9987$	$48.5 < RH \leq 61.5$	$-0.15 < T \leq 5.75$
$T_{dew} = 0.9029*T + 0.3429*RH + 0.0108*EA + 0.0036*W - 0.0556*S - 35.5285$	$RH \leq 49.5$	$5.75 < T \leq 9.95$
$T_{dew} = 0.9108*T + 0.2683*RH + 0.0189*EA + 0.0036*W - 0.0193*S - 39.3861$	$49.5 < RH \leq 61.5$	$5.75 < T \leq 9.95$
$T_{dew} = 0.8932*T + 0.2097*RH - 0.0008*EA + 0.0033*W - 0.0173*S - 18.9323$	$61.5 < RH \leq 73.5$	$-0.15 < T \leq 1.55$
$T_{dew} = 0.917*T + 0.2395*RH + 0.0204*EA - 0.0033*W + 0.0002 - 39.2768$	$61.5 < RH \leq 65.5$	$1.55 < T \leq 4.05$
$T_{dew} = 0.9644*T + 0.206*RH - 0.0003*EA + 0.0033*W + 0.0002*S - 19.2912$	$65.5 < RH \leq 73.5$	$1.55 < T \leq 4.05$
$T_{dew} = 0.9634*T + 0.1762*RH - 0.0017*EA + 0.0033*W - 0.0078*S - 15.844$	$RH > 73.5$	$1.55 < T \leq 4.05$
$T_{dew} = 0.9456*T + 0.2034*RH + 0.0097*EA + 0.0033*W - 0.0182*S - 27.6514$	$RH > 73.5$	$4.05 < T \leq 9.95$
$T_{dew} = 0.8883*T + 0.4574*RH + 0.0347*EA + 0.0041*W - 0.0601*S - 60.2489$	$RH \leq 39.5$	$9.95 < T \leq 15.35$
$T_{dew} = 0.8800*T + 0.3631*RH + 0.0346*EA + 0.0041*W - 0.0337*S - 56.861$	$39.5 < RH \leq 46.5$	$9.95 < T \leq 15.35$
$T_{dew} = 0.9395*T + 0.6002*RH + 0.0095*EA + 0.0041*W - 0.0105*S - 43.7243$	$RH < 28.5$	$15.35 < T \leq 20.85$
$T_{dew} = 0.9251*T + 0.4472*RH + 0.0400*EA + 0.0041*W - 0.0454*S - 65.1975$	$28.5 < RH \leq 35.5$	$15.35 < T \leq 20.85$
$T_{dew} = 0.9034*T + 0.3602*RH + 0.0358*EA + 0.0334*W - 0.0234*S - 58.3174$	$35.5 < RH \leq 46.5$	$15.35 < T \leq 20.85$
$T_{dew} = 0.9004*T + 0.2864*RH + 0.0449*EA + 0.0053*W - 0.0361*S - 62.6688$	$46.5 < RH \leq 59.5$	$T \leq 14.45$
$T_{dew} = 0.9328*T + 0.2231*RH + 0.0151*EA + 0.0053*W - 0.0233*S - 33.5626$	$RH > 59.5$	$T \leq 14.45$
$T_{dew} = 0.9333*T + 0.2794*RH + 0.0292*EA + 0.0649*W - 0.0196*S - 49.493$	All values	$14.45 < T \leq 20.85$
$T_{dew} = 0.8695*T + 0.6839*RH + 0.0749*EA + 0.0808*W - 0.0566*S - 100.5357$	$RH \leq 25.5$	$T > 20.85$
$T_{dew} = 0.8514*T + 0.4688*RH + 0.0167*EA + 0.1031*W - 0.0543*S - 44.4641$	$25.5 < RH \leq 36.5$	$T > 20.85$
$T_{dew} = 0.8677*T + 0.3523*RH - 0.0016*EA + 0.0998*W + 0.0019*S - 25.4284$	$RH > 36.5$	$T > 20.85$

As previously mentioned, Deka et al. [14] used SVM, ANN, and ELM and implemented meteorological parameters of minimum, maximum, and average temperatures, relative humidity, atmospheric pressure, water vapor pressure, sunny hours, and solar radiation in order to estimate the DPT in two cities of Kerman province, Iran. The minimum RMSE reported in the mentioned study was 0.49, related to the ELM method by using minimum temperature and water vapor pressure data as input parameters. In the present study, with the application of the M5 and applying meteorological parameters of average temperature, relative humidity, actual vapor pressure, wind speed, and sunny hours, the RMSE decreased to 0.37, indicating the high accuracy of the M5 model tree for estimation DPT values. The output of the M5 was a simple linear relationship that can be used to calculate the

DPT values easily, while the ELM does not have such a capability. Additionally, in another study by Baghban et al. [35] the maximum estimation accuracy of the DPT was reported using an SVM model with an RMSE value of 0.4.

Furthermore, the precision of the M5-15 in the current study was more than the accuracy of the proposed SVM method by Baghban et al. [35]. To investigate the influence of input parameters on the DPT estimation, the RMSE and  $R^2$  were utilized for different groupings of input variables. For this purpose, all utilized models, including GEP, M5, and SVR were selected for sensitivity analysis (Table 6). Each model confirmed the extent to which the eliminated variable would affect the model accuracy. As shown in Table 6, the precision of all models decreased if each of T, RH,  $V_p$ , W, and S input parameters were removed from the modeling. Furthermore, it can be comprehended that T had the greatest effect in increasing the prediction accuracy. In other words, eliminating T caused a sharp increase in RMSE values in all studied models.

**Table 6.** Effect of removing input variables on the utilized model's accuracy for predicting DPT values.

Model	Input Parameters	GEP		M5		SVR	
		RMSE (Degree)	$R^2$	RMSE (Degree)	$R^2$	RMSE (Degree)	$R^2$
1	All	2.60	0.835	0.37	0.996	0.55	0.989
2	Remove T	5.43	0.227	4.18	0.173	4.65	0.169
3	Remove S	2.58	0.847	2.73	0.753	3.16	0.980
4	Remove RH	3.18	0.752	3.72	0.689	3.83	0.342
5	Remove W	2.58	0.930	2.68	0.843	2.63	0.863
6	Remove $V_p$	2.78	0.642	2.91	0.541	2.31	0.763

## 5. Conclusions

In the current study, three data-driven methods including GEP, M5, and SVR were used to estimate DPT values at Tabriz synoptic station, Iran. For this purpose, the meteorological parameters were collected from the Meteorological Organization of East Azerbaijan province in the period 1998 to 2016. Also, 15 different input combinations were defined to study the effect of meteorological parameters on the estimation of DPT values. The results of this study revealed that the SVR-6, using two input parameters of T and RH, and GEP-10 using three parameters of T, RH, and S, had appropriate performance in the estimation of DPT values. Furthermore, the overall analysis of the studied methods showed that the M5-15 using five parameters of T, S, RH, W, and  $V_p$  had the best performance in the estimation of DPT values at Tabriz station in comparison with all considered models with different input combinations. To conclude, M5-15 is proposed as the most accurate method for the estimation of DPT values at the Tabriz synoptic station, Iran.

**Author Contributions:** The authors have made equal contributions.

**Funding:** This research received no external funding.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Famiglietti, C.A.; Fisher, J.B.; Halverson, G.; Borbas, E.E. Global Validation of MODIS Near-Surface Air and Dew Point Temperatures. *Geophys. Res. Lett.* **2018**, *45*, 7772–7780. [[CrossRef](#)]
2. Shiri, J. Prediction vs. estimation of dewpoint temperature: Assessing GEP, MARS and RF models. *Hydrol. Res.* **2018**. [[CrossRef](#)]
3. Ali, H.; Fowler, H.J.; Mishra, V. Global observational evidence of strong linkage between dew point temperature and precipitation extremes. *Geophys. Res. Lett.* **2018**, *45*, 12320–12330. [[CrossRef](#)]

4. Samadianfard, S.; Delirhasannia, R.; Kisi, O.; Agirre-Basurko, E. Comparative analysis of ozone level prediction models using gene expression programming and multiple linear regression. *GEOFIZIKA* **2013**, *30*, 43–74.
5. Mosavi, A.; Ozturk, P.; Chau, K.W. Flood prediction using machine learning models: Literature review. *Water* **2018**, *10*, 1536. [[CrossRef](#)]
6. Samadianfard, S.; Nazemi, A.H.; Sadraddini, A.A. M5 model tree and gene expression programming based modeling of sandy soil water movement under surface drip irrigation. *Agric. Sci. Dev.* **2014**, *3*, 178–190.
7. Samadianfard, S.; Sattari, M.T.; Kisi, O.; Kazemi, H. Determining flow friction factor in irrigation pipes using data mining and artificial intelligence approaches. *Appl. Artif. Intell.* **2014**, *28*, 793–813. [[CrossRef](#)]
8. Dehghani, M.; Riahi-Madvar, H.; Hooshyaripor, F.; Mosavi, A.; Shamshirband, S.; Zavadskas, E.K.; Chau, K.W. Prediction of Hydropower Generation Using Grey Wolf Optimization Adaptive Neuro-Fuzzy Inference System. *Energies* **2019**, *12*, 289. [[CrossRef](#)]
9. Lee, O.; Kim, S. Estimation of Future Probable Maximum Precipitation in Korea Using Multiple Regional Climate Models. *Water* **2018**, *10*, 637. [[CrossRef](#)]
10. Jabbari, A.; Bae, D.H. Application of Artificial Neural Networks for Accuracy Enhancements of Real-Time Flood Forecasting in the Imjin Basin. *Water* **2018**, *10*, 1626. [[CrossRef](#)]
11. Samadianfard, S.; Asadi, E.; Jarhan, S.; Kazemi, H.; Kheshtgar, S.; Kisi, O.; Sajjadi, S.; Abdul Manaf, A. Wavelet neural networks and gene expression programming models to estimate short-term soil temperature at different depths. *Soil Tillage Res.* **2018**, *175*, 37–50. [[CrossRef](#)]
12. Nie, J.; Liu, J.; Li, N.; Meng, X. Dew point measurement using dual quartz crystal resonator sensor. *Sens. Actuators B Chem.* **2017**, *246*, 792–799. [[CrossRef](#)]
13. Shamshirband, S.; Jafari Nodoushan, E.; Adolf, J.E.; Abdul Manaf, A.; Mosavi, A.; Chau, K.W. Ensemble models with uncertainty analysis for multi-day ahead forecasting of chlorophyll a concentration in coastal waters. *Eng. Appl. Comput. Fluid Mech.* **2019**, *13*, 91–101. [[CrossRef](#)]
14. Deka, P.C.; Patil, A.P.; Yeswanth Kumar, P.; Naganna, S.R. Estimation of dew point temperature using SVM and ELM for humid and semi-arid regions of India. *ISH J. Hydraul. Eng.* **2018**, *24*, 190–197. [[CrossRef](#)]
15. Zounemat-Kermani, M. Hourly predictive Levenberg–Marquardt ANN and multi linear regression models for predicting of dew point temperature. *Meteorol. Atmos. Phys.* **2012**, *117*, 181–192. [[CrossRef](#)]
16. Jia, Z.; Wang, Z.; Wang, H. Characteristics of Dew Formation in the Semi-Arid Loess Plateau of Central Shaanxi Province, China. *Water* **2019**, *11*, 126. [[CrossRef](#)]
17. Attar, N.; Khalili, K.; Behmanesh, J.; Khanmohammadi, N. On the reliability of soft computing methods in the estimation of dew point temperature: The case of arid regions of Iran. *Comput. Electron. Agric.* **2018**, *153*, 334–346. [[CrossRef](#)]
18. Mehdizadeh, S.; Behmanesh, J.; Khalili, K. Application of gene expression programming to estimate daily dew point temperature. *Appl. Therm. Eng.* **2017**, *112*, 1097–1107. [[CrossRef](#)]
19. Google Earth Data. Available online: <https://earth.google.com/web/@32.00526119,53.69582824,2905.05963411a,3214415.88153899d,35y,0h,0t,0r> (accessed on 3 February 2018).
20. Vapnik, V.; Golowich, S.; Smola, A. Support vector method for function approximation regression estimation, and signal processing. In *Advances in Neural Information Processing Systems*; MIT Press: Cambridge, MA, USA, 1996; Volume 9.
21. Suykens, J.A.K.; Van Gestel, T.; Brabanter, J.; De Moor, B.; Vandewalle, J. *Least Squares Support Vector Machines*; World Scientific: Singapore, 2002.
22. Sihag, P.; Jain, P.; Kumar, M. Modelling of impact of water quality on recharging rate of storm water filter system using various kernel function based regression. *Model. Earth Syst. Environ.* **2018**, *4*, 61–68. [[CrossRef](#)]
23. Sette, S.; Boullart, L. Genetic programming: Principles and applications. *Eng. Appl. Artif. Intell.* **2001**, *14*, 727–736. [[CrossRef](#)]
24. Aytok, A.; Kisi, O. A genetic programming approach to suspended sediment modeling. *J. Hydrol.* **2008**, *351*, 288–298. [[CrossRef](#)]
25. Quinlan, J.R. Learning with continuous classes. In *Proceedings of the 5th Australian Joint Conference on Artificial Intelligence*, Hobart, Tasmania, 16–18 November 1992; World Scientific: Singapore, 1992; pp. 343–348.
26. Pal, M. M5 model tree for land cover classification. *Int. J. Remote Sens.* **2006**, *27*, 825–831. [[CrossRef](#)]

27. Najafzadeh, M.; Shiri, J.; Sadeghi, G.; Ghaemi, A. Prediction of the friction factor in pipes using model tree. *ISH J. Hydraul. Eng.* **2018**, *24*, 9–15. [[CrossRef](#)]
28. Witten, I.H.; Frank, E. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*; Morgan Kaufmann: San Francisco, CA, USA, 2005.
29. Alizadeh, Z.; Yazdi, J.; Kim, J.H.; Al-Shamiri, A.K. Assessment of Machine Learning Techniques for Monthly Flow Prediction. *Water* **2018**, *10*, 1676. [[CrossRef](#)]
30. Sudheer, K.P.; Gosain, A.K.; Rangan, D.M.; Saheb, S.M. Modeling evaporation using an artificial neural network algorithm. *Hydrol. Process.* **2003**, *16*, 3189–3202. [[CrossRef](#)]
31. Taylor, K.E. Summarizing multiple aspects of model performance in a single diagram. *J. Geophys. Res. Atmos.* **2001**, *106*, 7183–7192. [[CrossRef](#)]
32. Choubin, B.; Moradi, E.; Golshan, M.; Adamowski, J.; Sajedi-Hosseini, F.; Mosavi, A. An Ensemble prediction of flood susceptibility using multivariate discriminant analysis, classification and regression trees, and support vector machines. *Sci. Total Environ.* **2019**, *651*, 2087–2096. [[CrossRef](#)] [[PubMed](#)]
33. Kurup, P.U.; Dudani, N.K. Neural networks for profiling stress history of clays from PCPT data. *J. Geotech. Geoenviron. Eng.* **2014**, *128*, 569–579. [[CrossRef](#)]
34. Deo, R.C.; Ghorbani, M.A.; Samadianfard, S.; Maraseni, T.; Bilgili, M.; Biazar, M. Multi-layer perceptron hybrid model integrated with the firefly optimizer algorithm for windspeed prediction of target site using a limited set of neighboring reference station data. *Renew. Energy* **2018**, *116*, 309–323. [[CrossRef](#)]
35. Baghban, A.; Bahadori, M.; Rozyn, J.; Lee, M.; Abbas, A.; Bahadori, A. Estimation of air dew point temperature using computational intelligence schemes. *Appl. Therm. Eng.* **2016**, *93*, 1043–1052. [[CrossRef](#)]



© 2019 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).