Computer Modeling in Engineering & Sciences

Tech Science Press

DOI: 10.32604/cmes.2022.018450

ARTICLE

Estimating Daily Dew Point Temperature Based on Local and Cross-Station Meteorological Data Using CatBoost Algorithm

Fuqi Yao[1], Jinwei Sun[1] and Jianhua Dong[2,*]

[1]School of Hydraulic Engineering, Ludong University, Yantai, 264010, China

[2]State Key Laboratory of Water Resources and Hydropower Engineering Science, Wuhan University, Wuhan, 430072, China

[*]Corresponding Author: Jianhua Dong. Email: djh0530dyz@126.com

Received: 25 July 2021 Accepted: 25 August 2021

ABSTRACT

Accurate estimation of dew point temperature (T_{dew}) plays a very important role in the fields of water resource management, agricultural engineering, climatology and energy utilization. However, there are few studies on the applicability of local T_{dew} algorithms at regional scales. This study evaluated the performance of a new machine learning algorithm, i.e., gradient boosting on decision trees with categorical features support (CatBoost) to estimate daily T_{dew} using limited local and cross-station meteorological data. The random forests (RF) algorithm was also assessed for comparison. Daily meteorological data from 2016 to 2019, including maximum, minimum and average temperature (T_{max}, T_{min} and T_{mean}), maximum, minimum and average relative humidity (RH_{max}, RH_{min} and RH_{mean}), maximum, minimum and average global solar radiation (Rs_{max}, Rs_{min} and Rs_{mean}) from three weather stations in Hunan of China were used to evaluate the CatBoost and RF algorithms. The results showed that both algorithms achieved satisfactory estimation accuracy at the target stations (on average RMSE = 1.020°C, R^2 = 0.969, MAE = 0.718°C and NRMSE = 0.087) in the absence of complete meteorological parameters (with only temperature data as input). The CatBoost algorithm (on average RMSE = 1.900°C and R^2 = 0.835) was better than the RF algorithm (on average RMSE = 2.214°C and R^2 = 0.828). The accuracy and stability of the CatBoost and RF algorithms were positively correlated with the number of input parameters, and the three-parameter algorithms achieved higher estimation accuracy than the two-parameter algorithms. The developed methodology is helpful to predict T_{dew} at regional scale.

KEYWORDS

Dew point temperature; categorical boosting; random forests; cross-station; accuracy

1 Introduction

Dew point temperature (T_{dew}) is the temperature at which water vapor in the air condenses into water droplets. Accurate estimation of T_{dew} data plays a significant role in the fields of energy utilization [1], thermal energy [2,3] and engineering [4]. T_{dew} is also an essential parameter for studying long-term climate change [5,6]. T_{dew} is usually used in conjunction with relative humidity to calculate the water content in the air [7]. It can be also combined with the wet bulb

temperature to calculate the ambient temperature to prevent crop frost in advance and reduce the risk of crop yield reduction [8]. In many fields, T_{dew} is needed to estimate reference crop evapotranspiration (ET_0) [9]. T_{dew} also affects human life safety and living environment comfort during the heatwave [10]. Compared with other meteorological variables, T_{dew} is still relatively inadequate. Because of its importance and non-linear changes, accurate estimation of T_{dew} has vital scientific significance in the above fields.

Compared with traditional meteorological parameters (such as temperature, precipitation and sunshine duration), T_{dew} is relatively more challenging to obtain. This is because some weather stations cannot measure T_{dew} normally. Scholars have mainly used traditional regression techniques to estimate T_{dew} [11], but the estimated data had significant errors and certain uncertainties. To better solve the problem of incomplete T_{dew} data, scholars have proposed various methods to estimate T_{dew}. In recent years, machine learning algorithms have been used to estimate various parameters and recognise images (including snow cover area, leaf area index (LAI), T_{dew} and image classification etc.) with excellent performance [12–17]. Kuter [14] estimated the snow cover over parts of the European Alps using remote sensing data combined with multiple adaptive regression spline (MARS), support vector regression (SVR), random forests (RF) and artificial neural network (ANN) algorithms. He concluded that MARS and RF algorithms would outperform ANN and SVR algorithms in terms of estimation performance and computational cost. Houborg et al. [15] obtained satisfactory results for estimating LAI by hybridizing the Cubist and RF algorithm. Because of their excellent performance in handling the non-linear relationships between inputs and output, e.g., ANN, MARS, RF, adaptive neural fuzzy inference system (ANFIS), support vector machine (SVM), extreme learning machine (ELM) and gene expression programming (GEP).

Among the above machine learning algorithms, ANN is the earliest and widely used algorithm. Shank et al. [18] applied the ANN algorithm with meteorological data from 20 weather stations in Georgia of USA to estimate T_{dew} within the next 1–12 h and established a general algorithm. They demonstrated that ANN had satisfactory accuracy in estimating T_{dew}. Zounemat-Kermani et al. [19] studied the potential of multilinear regression (MLR) and Levenberg-Marquardt neural network (LM-NN) in Ontario of Canada to estimate hourly T_{dew}. It was found that LM-NN had better performance than the MLR algorithm. Shiri et al. [20] coupled the ANN and GEP algorithms to estimate T_{dew} in Seoul and Incheon of South Korea and found that the ANN algorithm had excellent estimation capability. Still, the performance of the GEP algorithm was better than the ANN algorithm. Nadig et al. [21] implemented single and hybrid ANN algorithms to estimate air temperature and T_{dew}. They concluded that the hybrid algorithm could effectively improve the estimation stability, and the average error was decreased by 34.1%. The ANFIS algorithm was also often used to estimate T_{dew}. Mohammadi et al. [22] applied the ANFIS algorithm to estimate T_{dew} at Kerman and Tabas stations in Iran. They found that water vapor pressure (V_p) and RH were the most relevant and irrelevant meteorological parameters, respectively. Kisi et al. [23] evaluated several combined algorithms such as the ANFIS algorithm with sub-clustering identification (ANFIS-SC) and ANFIS algorithm with grid partitioning identification (ANFIS-GP) to estimate daily T_{dew} at three stations in South Korea. They indicated that the accuracy of these two algorithms was very close, and they were better than the other studied algorithms.

Baghban et al. [24] evaluated the genetic algorithm (GA)-optimized least squares support vector machine (LSSVM) and ANFIS algorithm to estimate T_{dew}, and found that these two algorithms had high accuracy and stability. The MARS algorithm's advantage is that it can

handle big data with high dimensions with short computational time and high prediction accuracy. Therefore, it has been used in many fields. Dong et al. [25] estimated daily diffuse solar radiation (R_d) at five stations in China using MARS and SVM. The results confirmed that the MARS algorithm had an excellent performance in estimating R_d. Wu et al. [26] applied MARS, ANFIS and SVM to estimate daily ET_0 in different climate zones of China. They found that MARS had a compelling estimation accuracy, which was superior to the other algorithms. Other scholars have used the MARS algorithm to estimate ET_0 [27] and R_s [28]. Of course, MARS has been also used to estimate T_{dew}. Shiri et al. [29] applied the MARS algorithm to estimate daily T_{dew} at six meteorological stations in northwestern Iran. They argued that the MARS algorithm had good performance in estimating T_{dew}, and its accuracy was better than the other algorithms. Many scholars have studied the MARS and GEP algorithms together. Among them, by using the meteorological data of thirteen meteorological stations in the arid region of Iran from 1960 to 2014, Attar et al. [30] coupled GEP, MARS and SVM algorithms to estimate T_{dew}. They found that the MARS algorithm obtained excellent T_{dew} estimates, which was better than SVM and GEP algorithms. In another study, the GEP, MARS and SVM algorithms were simultaneously used to estimate monthly ET_0 in Iran. The results further indicated that the MARS algorithm had better estimation accuracy than the other algorithms [31].

Scholars often compare the SVM algorithm with the ELM algorithm when estimating meteorological parameters. Because the ELM algorithm has excellent generalization performance and reduces operational time, it has been widely used in many fields. Deka et al. [32] evaluated daily T_{dew} in India's humid and semi-arid areas using SVM and ELM algorithms. They found that the ELM algorithm was better than the SVM algorithm. Some scholars also hybridized other machine learning models with ELM or SVM algorithms. Amirmojahedi et al. [33] hybridized ELM and wavelet transform (WT) (ELM-WT) to estimate T_{dew} in Bandar Abas of Iran and compared it with the SVM and ANN algorithms. The results showed that the hybrid algorithm performed better than the SVM and ANN algorithms, indicating that the hybrid algorithm was feasible in estimating T_{dew}. The kernel-based algorithm has been widely used in recent years because of its high accuracy and strong stability. The most popular ones are the SVM and ELM algorithms [34]. Wong et al. [35] compared the kernel-based ELM (K-ELM) and LS-SVM algorithms to estimate engine performance. They concluded that the estimation accuracy and stability of K-ELM and LS-SVM algorithms were very close. Feng et al. [36,37] also estimated ET_0 in China based on only temperature data.

In many cases, local meteorological data are partially or entirely missing due to various reasons. It is not easy to estimate the meteorological parameters at the target station. Therefore, it is essential to use the meteorological data from cross stations to estimate the meteorological data at the target station. Mehdizadeh et al. [38] evaluated the GEP algorithm to estimate daily T_{dew} at two stations in northwestern Iran using cross-station meteorological data. They demonstrated that estimating the target-station meteorological data using those from cross stations was highly accurate and feasible. Lu et al. [39] evaluated the feasibility of the gradient boosting decision tree (GBDT) and M5 model tree (M5Tree) algorithms to estimate daily pan evaporation (E_p) in the Poyang Lake area of China using combined local and cross-stations data. They presented that satisfactory accuracy was obtained using cross-stations meteorological data when the cross stations were less than 100 km away. Kim et al. [6] applied generalized regression neural networks (GRNN) and multilayer perceptron (MLP) algorithms to estimate daily T_{dew} with meteorological data from two stations in California of USA. The results indicated that the GRNN algorithm had better performance in estimating T_{dew}. Karimi et al. [40] coupled GEP and SVM algorithms to estimate

ET_0 with meteorological data from cross stations in South Korea's humid zones. The authors concluded that both the GEP and SVM algorithms could successfully estimate ET_0. The GEP algorithm performed better than the SVM algorithm in estimating ET_0 under the cross-station scenarios.

However, most machine learning algorithms are complex and require high computational costs during the calibration process. For example, algorithms such as SVM and ELM. Gradient boosting is an advanced intelligent technology that has been widely used due to its excellent data-processing capability and other advantages [39,41]. In the past, it mainly solved problems such as noisy data and complex parameter relationships, such as web searching [42], R_d [43] and ET_0 [44] estimation. Theoretical results of the gradient boosting provide solid explanation on how iteration combines basic predictions (weak models) through a greedy process corresponding to gradient descent in function space. CatBoost is a new machine learning algorithm using gradient boosting on decision trees with categorical features support [42]. The CatBoost algorithm has attracted much attention due to its higher computational efficiency and handling overfitting problems. RF, ANN, SVM and other machine learning algorithms have been used to estimate T_{dew}. Nevertheless, tree-based integrated algorithms, especially the CatBoost algorithm, have not been tested to estimate T_{dew} to the authors' knowledge. Compared with other tree algorithms and the utilization of local and cross-station meteorological data for estimating T_{dew}, the feasibility of extending local T_{dew} algorithms to regional scales has not been carried out. Although high prediction accuracy is the main consideration when using artificial intelligence algorithm, good stability and less computational workload should also be considered. For some regions where the meteorological data are partially or entirely lacking due to defective equipment and other reasons, estimation of T_{dew} in regions of lacking data becomes more meaningful through the regional application of local algorithms. Therefore, the purposes of this study focused on three points. First of all, different combinations of local meteorological data at three stations in different regions of Hunan Province of China were used to train and test the CatBoost and RF algorithms for estimating T_{dew}. Secondly, different data sets from cross stations were used to estimate T_{dew} at the target station. Finally, the effect of each meteorological variable for estimating daily T_{dew} at the target station was evaluated under two input scenarios, the potential of regional application of local T_{dew} algorithms was assessed, and the best algorithm and the most effective input combination were further proposed.

2 Materials and Methods

2.1 Study Area and Meteorological Data

Meteorological data at three weather stations (including Fenghuang, Huayuan and Longshan stations) in different regions of Xiangxi Tujia and Miao Autonomous Prefecture in northwestern Hunan Province of China were used to train and test the machine learning algorithms for estimating T_{dew}. This area has a subtropical monsoon humid climate with an area of 15,462 square kilometers. There are many types of crops in the area, including rice, wheat, maize, soybeans, etc. Surrounded by mountains, the local water resource is abundant.

The three stations selected in this study were cross stations. Daily meteorological data, including maximum, minimum and average temperature (T_{max}, T_{min} and T_{mean}), maximum, minimum and average relative humidity (RH_{max}, RH_{min} and RH_{mean}), maximum, minimum and average global solar radiation (Rs_{max}, Rs_{min} and Rs_{mean}) during 2016–2019 were collected to train and test the CatBoost and RF algorithms. The information and geographic locations of the selected stations are shown in Table 1 and Fig. 1. The meteorological data were provided and quality

inspected by the National Meteorological Information Center (NMIC) of the China Meteorological Administration (CMA). If the meteorological data were lost or the ratio of measured T_{dew} to actual T_{dew} was above 1, the information was further excluded. The input data were divided into three parts in sequence. The first two-thirds were used to develop and train machine learning algorithms, and the last one-third were used to test the algorithms. In terms of the number of observations, there were probably 700 for modelling and 350 for validation. In addition, there was a little bit of invalid data. All simulations were performed in a computer with Intel CPU I7 6700 @ 3.4–4.0 GHz and 16 GB of RAM memory.

Table 1: Average values of each basic information during 2016–2019 at the three stations selected in this study

Station name	Latitude (N)	Longitude (E)	Altitude (m)	T_{max} (°C)	T_{min} (°C)	T_{dew} (°C)	RH (%)	Rs (MJ m^{-2} day^{-1})	N (h)
Fenghuang	27.9	109.6	343	20.7	12.9	13.2	84.2	0.4	3.8
Huayuan	28.6	109.5	324	14.2	8.2	57.2	9.2	7.9	4.8
Longshan	29.5	109.4	456	22.9	14.7	6.7	1060.9	3.7	27.2

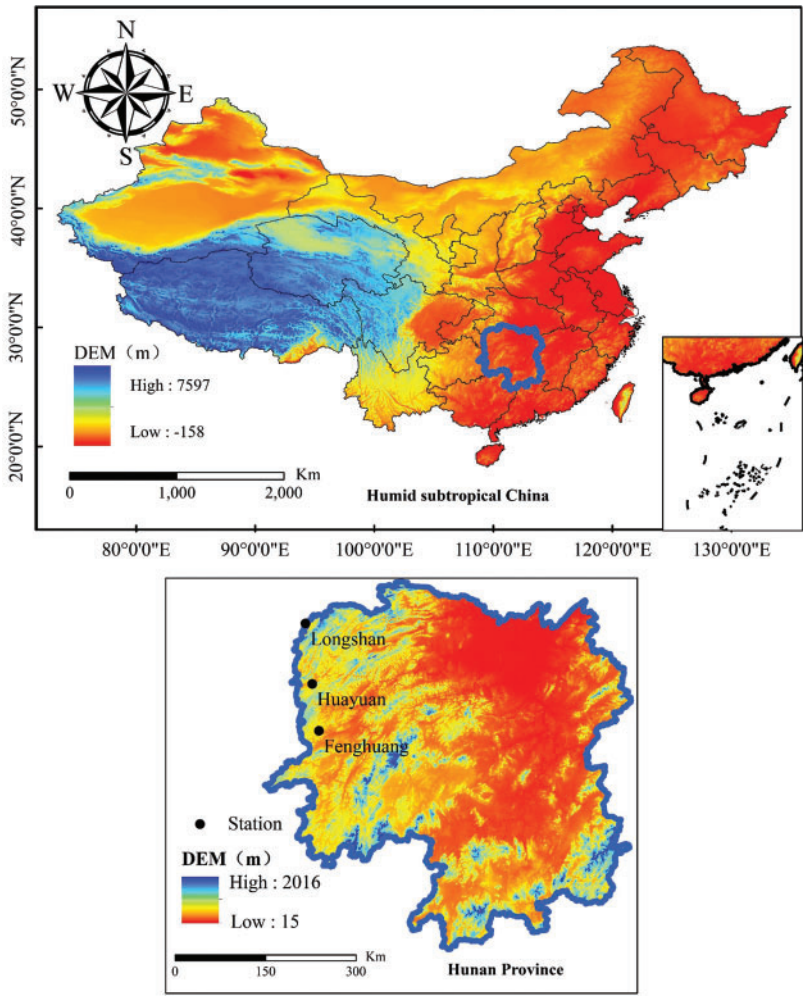

Figure 1: Spatial distributions of the three weather stations used in this study

2.2 Random Forests (RF)

The RF algorithm was proposed by Breiman [45], which was designed and developed using the classification and regression trees (CART) and the concept of "bagging". The RF algorithm has been widely used in regression and estimation studies. Because it is a machine learning algorithm that can effectively solves high-dimensional regression problems, the RF algorithm can use a subset of the data through bootstrap to process random binary trees. By repeatedly selecting random T ($T < N$) sample sets, a new training sample set is generated from the N original training samples. In the whole process of selecting samples, the same part of samples may be collected repeatedly. Therefore, a random subset of the training data set needs to be randomly extracted from the original data set for the development and training of the algorithm (the flowchart is shown in Fig. 2). Data sets that are not used in the algorithm are often called out-of-bag data ("out-of-bag" (OOB)). The algorithm's unused data sets will not be used to fit but will be used to test the algorithm's estimation ability.

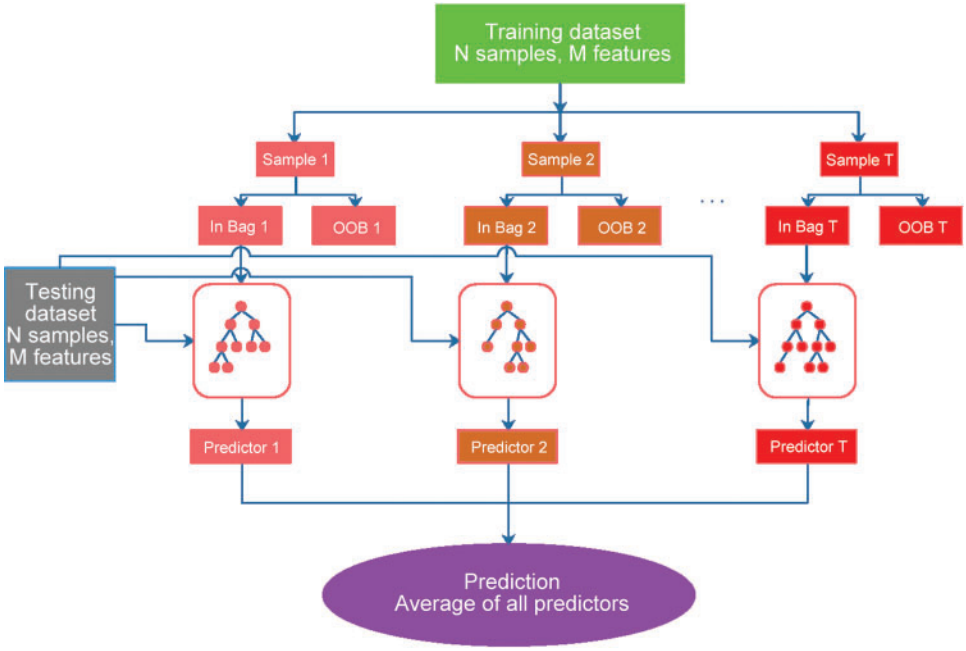

Figure 2: Flow chart of the RF algorithm

The CART algorithm in the RF algorithm differs from other traditional algorithms, which is based on feature selection according to the Gini coefficient. The criterion for selecting the Gini coefficient is that each child node needs to achieve the highest purity. At this time, the smaller the Gini coefficient, the higher the stability of the algorithm and the higher the purity. CART is a binary tree, which means that each non-leaf node can only produce two branches. If multiple (higher than two) discrete variables are generated on a non-leaf node, the variable may be reused multiple times. Each feature selected from the RF tree is randomly generated from all the features, which reduces the risk of overfitting. Unlike other decision trees, each RF tree is part of the selected feature. Among the selected features in this part, the best feature is selected to divide the left and right subtrees of the decision tree, thereby increasing the randomness and

further enhancing the algorithm's generalization ability. Finally, the average of all predictors can be derived.

In short, the final estimation of the RF algorithm is the average of all factors. More detailed information and methods on the RF algorithm can be found in the paper of Breiman [45].

2.3 Categorical Boosting (Catboost)

Gradient boosting on decision trees with categorical features support (such as CatBoost) is a new gradient enhanced decision tree (GBDT) algorithm. The traditional algorithm is pre-processed during the training process, while the CatBoost algorithm performs classification feature processing. Moreover, it can successfully handle classification features. Therefore, compared to the conventional GBDT algorithm, the CatBoost algorithm has more advantages. Specifically, for each example, the CatBoost algorithm randomly arranges and combines the data sets and calculates the average label value of the sample, which is the same as the replacement category value before the given category value.

If a permutation is $\theta = [\sigma_1, \ldots, \sigma_n]_n^T$, it is substituted with:

$$x_{\sigma_{p,\mathrm{k}}} = \frac{\sum_{j=1}^{p-1} [x_{\sigma_{\mathrm{j,k}}} = x_{\sigma_{p,\mathrm{k}}}] \cdot \mathrm{Y}_{\sigma j} + \beta \cdot P}{\sum_{j=1}^{p-1} [x_{\sigma_{\mathrm{j,k}}} = x_{\sigma_{p,\mathrm{k}}}] + \beta} \tag{1}$$

where P is a prior value and β is the weight of the prior.

Another advantage of this algorithm is that it uses a new algorithm to calculate the leaf value when selecting the tree structure, which helps reduce the problem of overfitting [42]. The CatBoost algorithm can combine all classification features as a new classification feature. The CatBoost algorithm will recombine it for extensive use when constructing a new segmentation for the tree. Another advantage of the CatBoost algorithm is that it uses the forget tree as a predictor. In addition, the length of each leaf index of the tree is equal to the binary vector of the tree depth. This makes the CatBoost algorithm widely used. First of all, all the floating-point number features, statistical features, and single-hot encoding features are binarized, which are used to calculate the algorithm prediction.

Usually, the prediction offset is the main problem that plagues modeling. In each iteration of GDBT, the loss function uses the same data set to obtain the gradient of the algorithm, and then trains to obtain a basic learner, which will cause the gradient estimation deviation, which will lead to the problem of overfitting the algorithm. The CatBoost algorithm uses ordered boosting to replace gradient estimation in traditional algorithms, reducing gradient estimation bias and improving algorithm capabilities [42]. The structural flow chart of the CatBoost algorithm is shown in Fig. 3, with more detailed information and methods related to the RF algorithm, which can be found in the research of Dorogush et al. [42].

The main parameters of both algorithms were optimized using the grid search method. The best performing parameters were used for model training and testing. The main parameters of the RF algorithm are the maximum depth and the number of trees. Trees are more prone to be overfitting if they have larger maximum depths. In this study, the upper and lower limits on the parameters were first determined by trial and error methods. A grid was then created and the best combination of parameters was found by setting different step sizes. CatBoost is also a tree-based algorithm. Although it has many parameters, the main parameters affecting model's accuracy and stability are the same as RF.

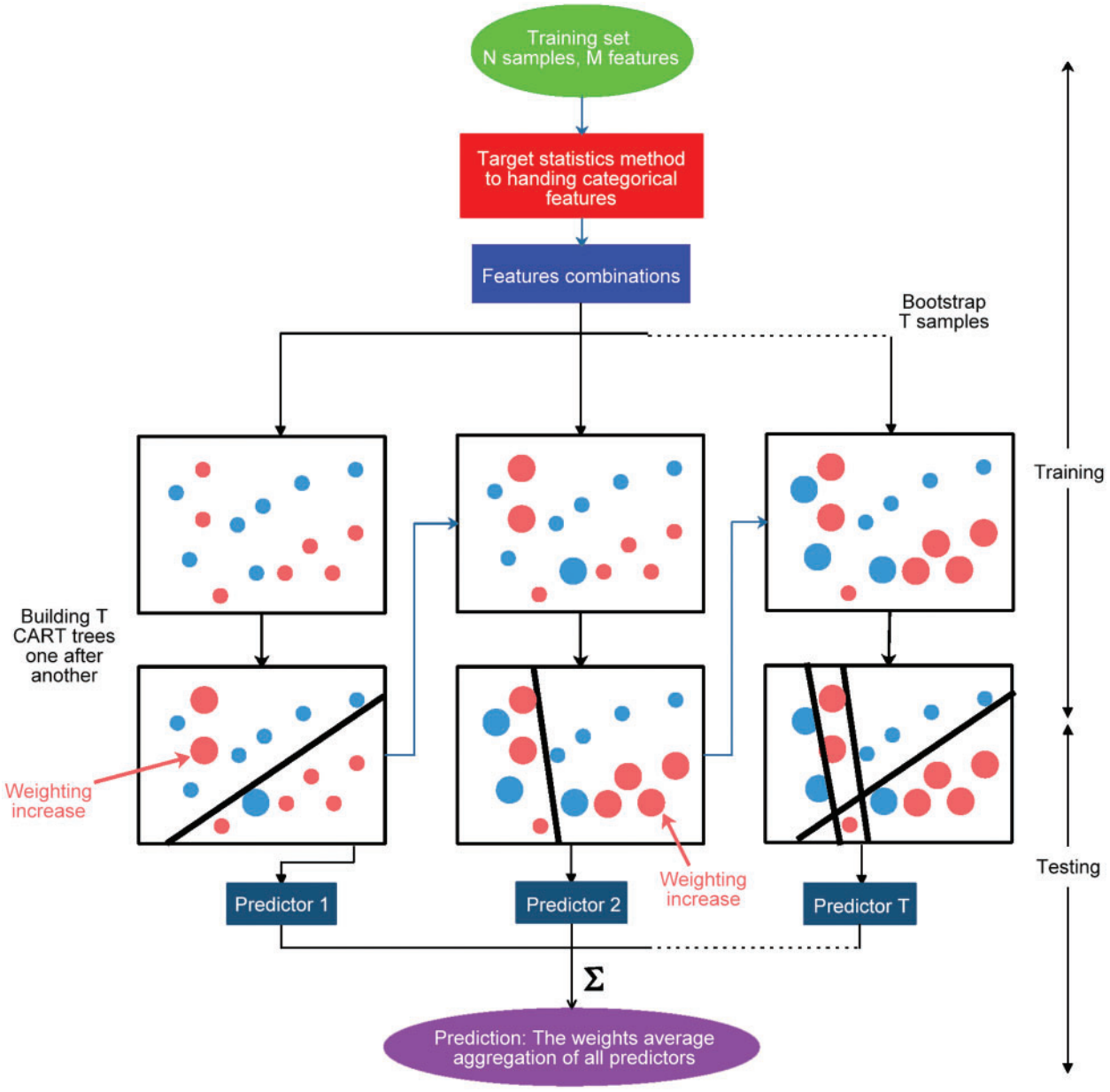

Figure 3: Flow chart of the CatBoost algorithm

2.4 Algorithm Comparison and Statistical Analysis

In this study, four commonly used statistical indicators were selected to analyze and compare the algorithm performance in T_{dew} estimation under two input scenarios. These statistical indicators were coefficient of determination (R^2), root mean square error (RMSE), mean absolute error

(MAE) and normalized root mean square error (NRMSE). The mathematical equations of each statistical indicator are described as follows:

$$R^2 = \frac{\left[\sum_{i=1}^{n}(O_{i,m} - \overline{O}_{i,\mathrm{m}})(O_{i,e} - \overline{O}_{i,\mathrm{e}})\right]^2}{\sum_{i=1}^{n}(O_{i,m} - \bar{O}_{\mathrm{i,m}})^2 \sum_{i=1}^{n}(O_{i,e} - \overline{O}_{i,\mathrm{e}})^2} \tag{2}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(O_{i,m} - O_{\mathrm{i,e}})^2} \tag{3}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|O_{i,m} - O_{\mathrm{i,e}}| \tag{4}$$

$$NRMSE = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(O_{i,m} - O_{\mathrm{i,e}})^2}}{\overline{O}_{i,m}} \tag{5}$$

where $O_{i,m}$, $O_{i,\mathrm{e}}$, $\overline{O}_{\mathrm{i,m}}$, $\overline{O}_{i,\mathrm{e}}$ and n are the measured, estimated, mean of measured, mean of estimated T_{dew} and the number of observations, respectively. The higher the R^2 value, that is, the closer it is to 1, the higher the accuracy and the better the regression line matches the data. In contrast, algorithm performance is inversely related to RMSE, MAE, and NRMSE.

3 Results and Discussion

3.1 Comparison of Algorithm Accuracy under Various Local Input Combinations

This section evaluates the applicability of RF and CatBoost algorithms for estimating T_{dew} under the local input scenario, using meteorological data from Fenghuang, Huayuan and Longshan stations in China. The daily meteorological data were maximum, minimum and average temperature (T_{max}, T_{min} and T_{mean}), maximum, minimum and average relative humidity (RH_{max}, RH_{min} and RH_{mean}), maximum, minimum and average global solar radiation (Rs_{max}, Rs_{min} and Rs_{mean}). Table 2 presents the statistical results of the RF and CatBoost algorithms in estimating T_{dew} during the testing phase at the three stations under nine single-parameter inputs. Overall, the CatBoost algorithm (on average RMSE = 6.304°C and $R^2 = 0.328$) had better performance than the RF algorithm (on average RMSE = 7.014°C and $R^2 = 0.307$). It can be seen from Table 2 that the most relevant parameters for estimating T_{dew} in Fenghuang, Huayuan and Longshan stations were T_{min}, T_{mean} and T_{min}, respectively. The importance of T (on average RMSE = 2.415°C and $R^2 = 0.891$) was greater than that of RH (on average RMSE = 8.889°C and $R^2 = 0.024$) and Rs (on average RMSE = 8.673°C and $R^2 = 0.038$). Therefore, T was the most effective meteorological variable among the single factors, and the estimation accuracy and stability of the CatBoost algorithm were better than those of the RF algorithm.

Table 2: Statistical results of the two machine learning algorithms during the testing phase with single local parameters at three stations

Station	RF					CatBoost				
	Input	RMSE (°C)	R^2	MAE (°C)	NRMSE	Input	RMSE (°C)	R^2	MAE (°C)	NRMSE
Fenghuang										
	T_{min}	2.436	0.905	1.887	0.163	T_{min}	2.183	0.922	1.699	0.146
	T_{mean}	2.444	0.902	1.894	0.164	T_{mean}	2.312	0.910	1.799	0.155
	T_{max}	2.827	0.868	2.208	0.189	T_{max}	2.502	0.896	1.945	0.168
	RH_{mean}	7.771	0.065	6.506	0.520	Rs_{min}	6.753	0.155	5.685	0.453
	Rs_{min}	7.841	0.054	6.362	0.525	Rs_{mean}	6.960	0.102	5.857	0.467
	RH_{max}	8.137	0.028	6.730	0.545	Rs_{max}	7.011	0.087	6.014	0.470
	Rs_{mean}	8.228	0.031	6.632	0.551	RH_{mean}	7.108	0.052	6.246	0.477
	RH_{min}	8.243	0.017	6.883	0.552	RH_{min}	7.110	0.054	6.268	0.477
	Rs_{max}	8.581	0.007	7.100	0.575	RH_{max}	7.133	0.049	6.240	0.478
Huayuan										
	T_{mean}	2.261	0.904	1.784	0.213	T_{min}	2.132	0.918	1.645	0.201
	T_{min}	2.298	0.900	1.775	0.217	T_{mean}	2.202	0.912	1.773	0.208
	T_{max}	2.687	0.861	2.175	0.253	T_{max}	2.353	0.898	1.913	0.222
	RH_{min}	9.798	0.016	8.015	0.924	RH_{min}	9.185	0.023	7.581	0.866
	Rs_{mean}	10.086	0.004	7.998	0.951	RH_{max}	9.226	0.015	7.659	0.870
	RH_{mean}	10.176	0.001	8.302	0.959	Rs_{max}	9.252	0.000	7.438	0.872
	Rs_{min}	10.355	0.001	8.181	0.976	Rs_{mean}	9.426	0.000	7.480	0.889
	RH_{max}	10.554	0.003	8.742	0.995	RH_{mean}	9.455	0.004	7.871	0.891
	Rs_{max}	11.012	0.008	8.840	1.038	Rs_{min}	9.531	0.000	7.542	0.898
Longshan										
	T_{min}	2.471	0.872	1.954	0.225	T_{min}	2.198	0.903	1.774	0.200
	T_{mean}	2.555	0.863	2.017	0.232	T_{mean}	2.341	0.887	1.917	0.213
	T_{max}	2.731	0.845	2.157	0.248	T_{max}	2.528	0.869	2.060	0.230
	Rs_{mean}	8.611	0.041	6.589	0.783	Rs_{mean}	8.076	0.045	6.230	0.734
	Rs_{min}	8.986	0.023	7.153	0.817	Rs_{min}	8.403	0.019	6.594	0.764
	Rs_{max}	9.037	0.025	7.065	0.822	Rs_{max}	7.966	0.056	6.093	0.724
	RH_{max}	9.314	0.015	7.641	0.847	RH_{max}	8.843	0.048	7.428	0.804
	RH_{mean}	9.811	0.005	8.115	0.892	RH_{mean}	8.961	0.019	7.477	0.815
	RH_{min}	10.127	0.021	8.374	0.921	RH_{min}	9.056	0.001	7.512	0.823

To explore the effect of the two-parameter input combination on T_{dew} estimation, we randomly combined nine single parameters and determined the top five accurate algorithms with the two parameter inputs. The statistical results during the testing phase are shown in Table 3. It can be seen from Table 3 that under the two-parameter combinations, the CatBoost algorithm (on average RMSE = 0.499°C and $R^2 = 0.995$) had slightly better performance than the RF algorithm (on average RMSE = 0.746°C and $R^2 = 0.990$). When the input parameters were T and RH, the estimation accuracy of each algorithm was highest, which were the most effective meteorological factors under the two-parameter combinations. At Fenghuang station, the optimal parameter combinations of the RF and CatBoost algorithms were T_{min}, RH_{min} (RMSE = 0.422°C

and $R^2 = 0.997$) and T_{mean}, RH_{max} (RMSE = 0.228°C and $R^2 = 0.999$), respectively. In terms of accuracy, the difference between the CatBoost algorithm and the RF algorithm was small, but the former was more stable than the latter. The performance of the algorithms at different stations was also different. The performance of CatBoost algorithm at Fenghuang station (RMSE = 0.270°C and $R^2 = 0.999$) was better than that at Huayuan (RMSE = 0.537°C and $R^2 = 0.995$) and Longshan (RMSE = 0.715°C and $R^2 = 0.990$) stations when the input combination was T_{mean} and RH_{mean}. Therefore, the station's geographic location, terrain and climate also affected the accuracy of the algorithm's performance in estimating T_{dew}. Fan et al. [46] and Feng et al. [37] also confirmed that climate and geographical conditions would significantly impact the algorithm's performance. Under this input combination of T_{mean} and RH_{mean}, the algorithm's performance was good. Therefore, in terms of two-parameter combination, T and RH were the most effective meteorological factors. Under the parameter combination of T_{mean} and RH_{mean}, the estimation accuracy and stability of the CatBoost algorithm were better than those of the RF algorithm.

Table 3: Statistical results of the two machine learning algorithms during the testing phase with two-parameter local data at three stations

		RF					CatBoost			
Station	Input	RMSE (°C)	R^2	MAE (°C)	NRMSE	Input	RMSE (°C)	R^2	MAE (°C)	NRMSE
Fenghuang										
	T_{min}, RH_{min}	0.422	0.997	0.277	0.028	T_{mean}, RH_{max}	0.228	0.999	0.166	0.015
	T_{mean}, RH_{min}	0.423	0.997	0.270	0.028	T_{min}, RH_{mean}	0.257	0.999	0.190	0.017
	T_{min}, RH_{mean}	0.424	0.997	0.267	0.028	T_{max}, RH_{min}	0.269	0.999	0.193	0.018
	T_{min}, RH_{max}	0.426	0.997	0.277	0.029	T_{mean}, RH_{mean}	0.270	0.999	0.175	0.018
	T_{mean}, RH_{mean}	0.451	0.996	0.293	0.030	T_{mean}, RH_{min}	0.299	0.998	0.204	0.020
Huayuan										
	T_{mean}, RH_{max}	0.823	0.989	0.471	0.078	T_{max}, RH_{min}	0.529	0.995	0.352	0.050
	T_{mean}, RH_{mean}	0.840	0.989	0.478	0.079	T_{mean}, RH_{mean}	0.537	0.995	0.303	0.051
	T_{mean}, RH_{min}	0.840	0.989	0.475	0.079	T_{mean}, RH_{min}	0.550	0.995	0.299	0.052
	T_{min}, RH_{max}	0.879	0.988	0.512	0.083	T_{max}, RH_{mean}	0.573	0.994	0.388	0.054
	T_{min}, RH_{mean}	0.895	0.987	0.515	0.084	T_{mean}, RH_{max}	0.575	0.994	0.340	0.054
Longshan										
	T_{min}, RH_{max}	0.933	0.985	0.536	0.085	T_{min}, RH_{min}	0.657	0.992	0.360	0.060
	T_{min}, RH_{mean}	0.938	0.985	0.520	0.085	T_{min}, RH_{mean}	0.658	0.992	0.357	0.060
	T_{min}, RH_{min}	0.942	0.984	0.531	0.086	T_{min}, RH_{max}	0.677	0.991	0.365	0.062
	T_{mean}, RH_{min}	0.964	0.984	0.556	0.088	T_{mean}, RH_{min}	0.698	0.990	0.376	0.063
	T_{mean}, RH_{mean}	0.983	0.983	0.579	0.089	T_{mean}, RH_{mean}	0.715	0.990	0.403	0.065

To evaluate the effect of the three-parameter combination on T_{dew} estimation, we randomly combined nine single parameters and determined the top five accurate algorithms. The statistical results during the testing phase are shown in Table 4. Tables 3 and 4 showed that the trends were almost the same. The CatBoost algorithm was slightly better than the RF algorithm in estimating T_{dew} (on average RMSE decreased by 40.3%, R^2 increased by 0.6%, MAE decreased by 34.7% and NRMSE decreased by 40.6%). The two algorithms were best applied at Fenghuang station, both outperforming the other two stations. The RF and Catboost algorithms had some differences in parameter combinations. The parameter combination contained T and RH in the RF algorithm, while the CatBoost algorithm had a combination of T, RH and Rs. Therefore, T and RH were still the most effective meteorological factors. Algorithms with T_{mean} and RH_{mean} generally had better stability. In terms of MAE at the Longshan station, the RF and CatBoost algorithms were overfitted and seriously overestimated T_{dew} (on average MAE = 0.535°C and 0.358°C, respectively). This conclusion was consistent with that obtained by Shiri [29] when estimating daily T_{dew} using the RF algorithm, which suffered from over-fitting. In the study of Fan et al. [34] for estimating ET_0 through a machine learning algorithm, the RF algorithm had a poor-fitting effect relative to the GBDT algorithm. In terms of the three parameters, T and RH were still the most effective meteorological factors for estimating T_{dew}. The parameter combination of T_{mean} and RH_{mean} had relatively good accuracy. The estimation accuracy and stability of the CatBoost algorithm were better than those of the RF algorithm.

Table 4: Statistical results of the two machine learning algorithms during the testing phase with three-parameter local data at three stations

		RF					CatBoost			
Station	Input	RMSE (°C)	R^2	MAE (°C)	NRMSE	Input	RMSE (°C)	R^2	MAE (°C)	NRMSE
Fenghuang										
	T_{mean}, RH_{min}, Rs_{mean}	0.418	0.997	0.278	0.028	T_{mean}, T_{min}, RH_{max}	0.228	0.999	0.163	0.015
	T_{mean}, T_{min}, RH_{min}	0.420	0.997	0.269	0.028	T_{mean}, RH_{max}, RH_{mean}	0.236	0.999	0.165	0.016
	T_{min}, RH_{max}, RH_{min}	0.424	0.997	0.270	0.028	T_{mean}, T_{min}, RH_{mean}	0.241	0.999	0.167	0.016
	T_{mean}, RH_{mean}, RH_{min}	0.426	0.997	0.264	0.029	T_{max}, RH_{mean}, RH_{min}	0.257	0.999	0.183	0.017
	T_{min}, RH_{mean}, RH_{min}	0.448	0.996	0.281	0.030	T_{max}, T_{min}, RH_{min}	0.266	0.999	0.190	0.018
Huayuan										
	T_{mean}, RH_{mean}, RH_{min}	0.814	0.989	0.467	0.077	T_{mean}, RH_{mean}, Rs_{mean}	0.465	0.996	0.302	0.044
	T_{mean}, RH_{max}, RH_{min}	0.817	0.990	0.455	0.077	T_{mean}, RH_{min}, Rs_{min}	0.472	0.996	0.307	0.044

(Continued)

Table 4 (Continued)

Station	RF					CatBoost				
	Input	RMSE (°C)	R^2	MAE (°C)	NRMSE	Input	RMSE (°C)	R^2	MAE (°C)	NRMSE
Longshan	T_{mean}, RH_{max}, RH_{mean}	0.820	0.990	0.462	0.077	T_{mean}, RH_{mean}, Rs_{max}	0.474	0.996	0.300	0.045
	T_{max}, T_{mean}, RH_{min}	0.831	0.989	0.478	0.078	T_{mean}, RH_{mean}, Rs_{min}	0.497	0.996	0.304	0.047
	T_{max}, T_{min}, RH_{min}	0.838	0.988	0.502	0.079	T_{mean}, RH_{min}, Rs_{max}	0.500	0.996	0.315	0.047
	T_{min}, RH_{max}, RH_{mean}	0.927	0.985	0.533	0.084	T_{min}, RH_{mean}, Rs_{max}	0.563	0.994	0.356	0.051
	T_{min}, RH_{max}, RH_{min}	0.928	0.985	0.524	0.084	T_{min}, RH_{max}, Rs_{mean}	0.573	0.994	0.357	0.052
	T_{mean}, T_{min}, RH_{min}	0.931	0.985	0.534	0.085	T_{min}, RH_{mean}, Rs_{min}	0.576	0.993	0.350	0.052
	T_{min}, RH_{mean}, RH_{min}	0.935	0.985	0.539	0.085	T_{min}, RH_{max}, Rs_{max}	0.576	0.994	0.355	0.052
	T_{mean}, T_{min}, RH_{mean}	0.949	0.984	0.547	0.086	T_{min}, RH_{min}, Rs_{max}	0.584	0.993	0.373	0.053

To better compare the effect of each parameter combination on T_{dew} estimation, we plotted the estimated T_{dew} and measured values of some parameter combinations during the testing phase in Fig. 4 (with the CatBoost algorithm at Fenghuang station as an example). It can be seen from Fig. 4 that when the input combination was any single parameter (especially the meteorological factors RH_{mean}, RH_{max}, RH_{min}), the scatter diagram had no clear trends, and the scatter was distributed on both sides of the standard line, showing poor accuracy. Adding RH or Rs to T, the algorithm accuracy was significantly improved compared with the single parameter algorithm. Under two- and three-parameter combinations, the scatter points obtained by the algorithm were closer to the standard line and more uniformly distributed when the single parameter was used. It showed that in the study of estimating daily T_{dew}, the increase of meteorological parameters could improve the estimation performance of the algorithm. Dong et al. [47] also confirmed that the increase in effective meteorological parameters could improve the algorithm's estimation accuracy. However, the difference in accuracy between the two and three-parameter combinations was not significant. It can be seen that the most effective meteorological variables were T and RH. The additional incorporation of Rs to the algorithm failed to improve the algorithm accuracy, and some even declined. It indicated that the Rs was not a necessary parameter to estimate T_{dew}. Dong et al. [25] also showed that adding extra parameters would reduce the estimation accuracy of the algorithm.

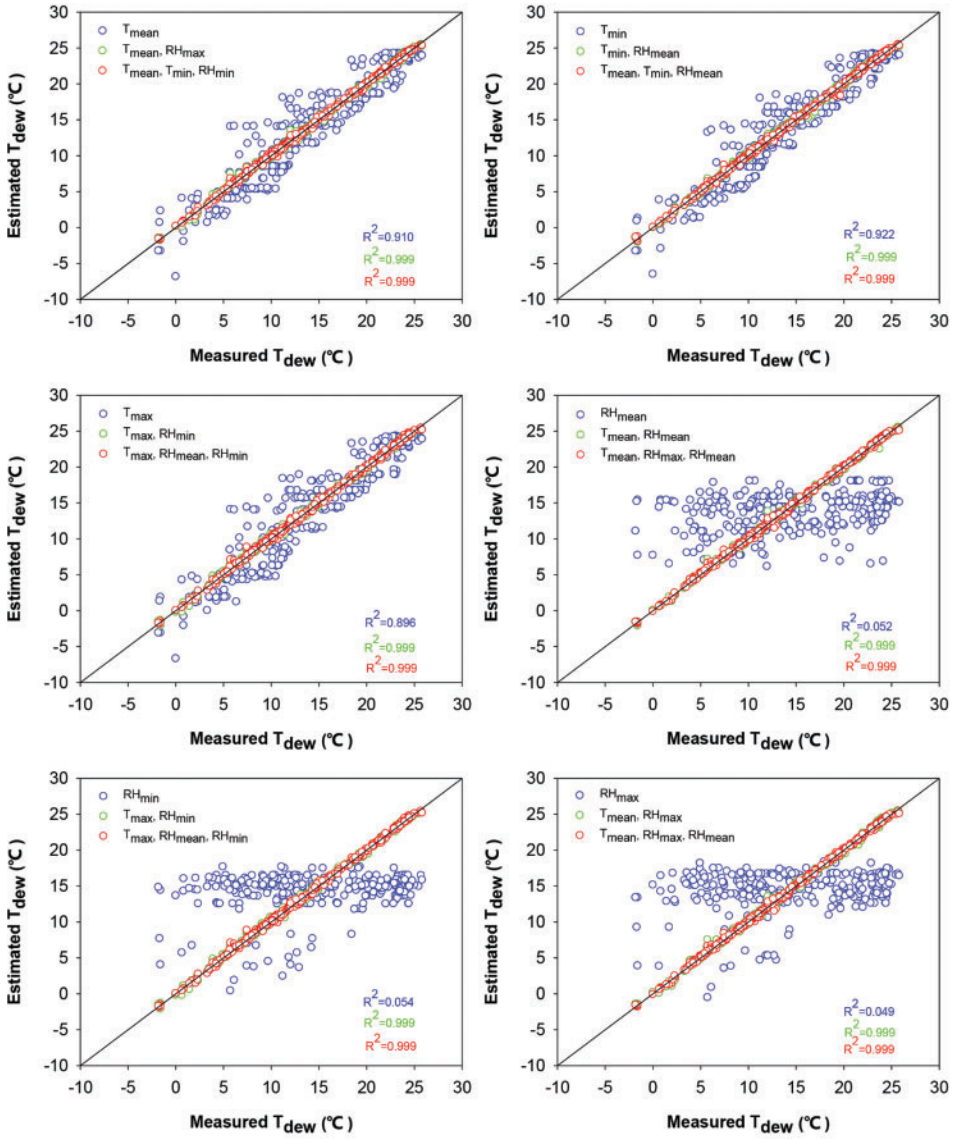

Figure 4: Scatter plots of the daily T_{dew} estimated by the algorithm and the corresponding measured values during the testing phase under local data input conditions (note: fine line is the best-fitted line)

To better compare the performance of the two algorithms, Fig. 5 compares the statistical results obtained by combining some parameters of the two algorithms during the testing phase at the Fenghuang station. Because the three- and two-parameter combinations of the two algorithms had very similar performances, the single- and two-parameter combinations were compared. It can be seen from the figure that the performance of the RF and CatBoost algorithms was very close. Still, the CatBoost algorithm was slightly better than the RF algorithm under various combinations. For the single-parameter combination, T was more important than RH and Rs. Therefore, the performance of the algorithm with only T data was slightly worse than that of algorithms with two parameters. The required meteorological data was also smallest, which showed the advantage of this input combination. The RMSE values of the two-parameter algorithms

were close to 0, showing extremely high stability. Regarding MAE values, algorithms with the single-parameter Rs (on average MAE = 6.275°C) or RH (on average MAE = 6.275°C) showed overfitting. In the RF algorithm, the performance of the input combination of T_{mean} and RH_{min} was better than that with RH_{min} (RMSE decreased by 94.5%, R^2 increased by 5758.2%, MAE decreased by 95.7% and NRMSE decreased by 94.6%). It showed that increasing the input number of meteorological parameters effectively improved the estimation accuracy of the algorithm. This conclusion was consistent with previous studies. Therefore, the meteorological factors T and RH were the most influential parameters for estimating T_{dew}, and the CatBoost algorithm had better performance than the RF algorithm.

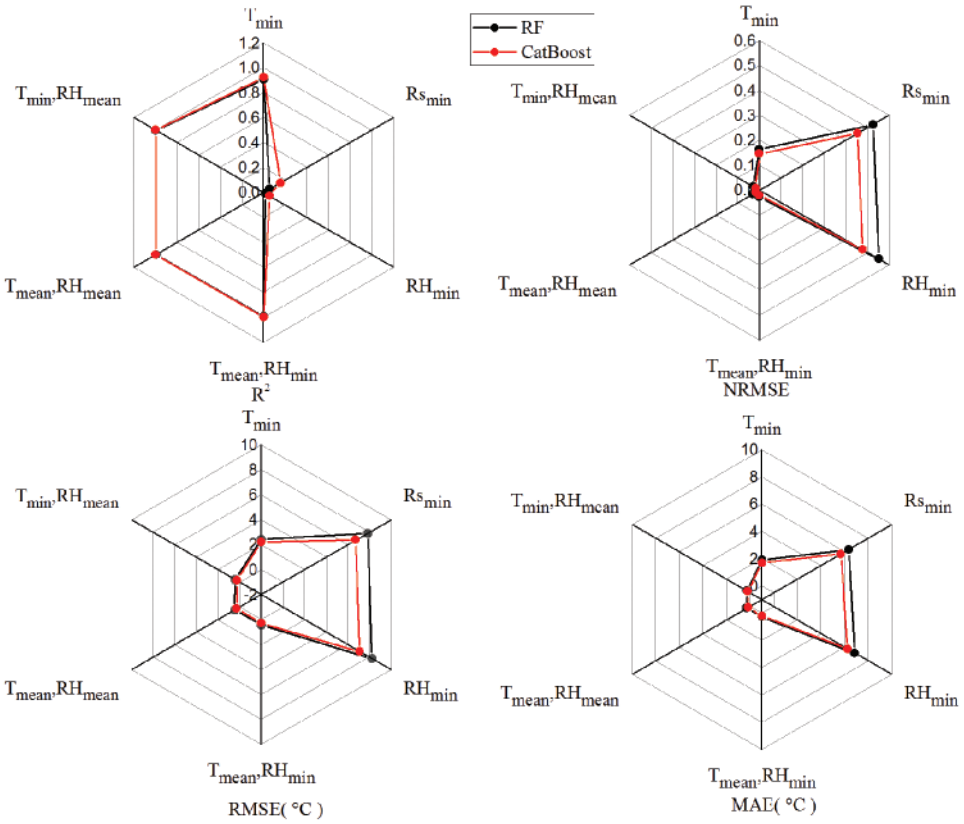

Figure 5: Radar chart of the statistical restuls during the testing phase under local input conditions at the Fenghuang station

3.2 Comparing Algorithm Accuracy by Replacing Local Data with Cross-Station Data

It is important to use meteorological data from cross stations in different regions to estimate T_{dew} at the target station. In some developing countries with incomplete measurement equipment, some meteorological data cannot be obtained normally due to various reasons. Therefore, it is necessary to use this method for T_{dew} estimation. The method also reflected the modeling ability of the regional application of the algorithm [37]. This section evaluated the applicability of RF and CatBoost algorithms to estimate the target-station T_{dew} using meteorological data from cross stations during the testing phase. Daily meteorological data at Fenghuang (Fh), Huayuan (Hy) and Longshan (Ls) stations (different regions) in China were used. The daily meteorological data were

maximum, minimum and average temperature (T_{max}, T_{min} and T_{mean}), maximum, minimum and average relative humidity (RH_{max}, RH_{min} and RH_{mean}), maximum, minimum and average global solar radiation (Rs_{max}, Rs_{min} and Rs_{mean}). For example, Fh-Hy meant that the meteorological data of Huayuan station were applied to Fenghuang station, and so on. To explore the effect of the two-parameter combination on T_{dew} after the station exchange, we randomly combined nine single parameters and explored the top five accurate algorithms. The statistical results during the testing phase are shown in Table 5. It can be seen from Table 5 that the performance of most CatBoost algorithms was better than that of RF algorithms. Except for the parameter combination of T_{mean} and RH_{max} in Fh-Hy and Fh-Ls, the performance of the CatBoost algorithm was poor (on average RMSE = 0.512°C, R^2 = 0.914, MAE = 1.820°C and NRMSE = 0.212). On the whole, before and after the station change, the best five parameter combinations did not change at each station, but the performance order changed. For example, the best two-parameter combination of the CatBoost algorithm at Fenghuang station before the change was T_{mean}, RH_{max} (RMSE = 0.228°C, R^2 = 0.999, MAE = 0.166°C and NRMSE = 0.015). The best two-parameter combination after the station change was T_{min}, RH_{mean} (RMSE = 0.307°C, R^2 = 0.998, MAE = 0.224°C and NRMSE = 0.028). The performance difference of each parameter combination was small, which further confirmed that the CatBoost algorithm had high stability. In Fh-Ls, the optimal parameter combination of the RF algorithm before switching stations was T_{min}, RH_{min}. The ranking remained unchanged after switching stations, and RMSE was decreased by 12.8%. It showed that in Fh-Ls, when the input parameter combination was T_{min}, RH_{min}, the algorithm had higher stability and the stability was improved after the station was changed. The feasibility of using the meteorological data from cross stations to estimate the target-station T_{dew} was confirmed. In Ls-Hy, the overall RMSE of the CatBoost algorithm was less than that of the RF algorithm, and the R^2 was greater than that of the RF algorithm. The accuracy and stability of the CatBoost algorithm after station replacement were better than those of the RF algorithm. This conclusion was also consistent with the finding of Lu et al. [39]. However, compared with the algorithm without station exchange, the CatBoost (on average RMSE increased by 39.5%, R^2 decreased by 0.6%) and RF (on average RMSE increased by 29.3%, R^2 decreased by 1.0%) algorithms lowered the performance. It showed that the application of data from Huayuan station instead of Longshan station had instability, but other applications of station exchange had good stability.

Table 5: Statistical results of the two machine learning algorithms during the testing phase using two-parameter cross-station data at three stations

		RF					CatBoost			
Station	Input	RMSE (°C)	R^2	MAE (°C)	NRMSE	Input	RMSE (°C)	R^2	MAE (°C)	NRMSE
Fh-Hy										
	T_{min}, RH_{mean}	0.413	0.997	0.312	0.039	T_{min}, RH_{mean}	0.302	0.998	0.218	0.028
	T_{min}, RH_{min}	0.425	0.997	0.314	0.040	T_{mean}, RH_{min}	0.331	0.998	0.236	0.031
	T_{mean}, RH_{min}	0.464	0.996	0.348	0.044	T_{mean}, RH_{mean}	0.371	0.998	0.257	0.035
	T_{min}, RH_{max}	0.506	0.995	0.376	0.048	T_{max}, RH_{mean}	0.457	0.997	0.323	0.043
	T_{mean}, RH_{mean}	0.523	0.995	0.387	0.049	T_{mean}, RH_{max}	0.597	0.931	1.818	0.213

(Continued)

Table 5 (Continued)

		RF					CatBoost			
Station	Input	RMSE (°C)	R^2	MAE (°C)	NRMSE	Input	RMSE (°C)	R^2	MAE (°C)	NRMSE
Fh-Ls										
	T_{min}, RH_{min}	0.368	0.997	0.269	0.033	T_{min}, RH_{mean}	0.311	0.998	0.229	0.028
	T_{min}, RH_{mean}	0.388	0.997	0.279	0.035	T_{mean}, RH_{min}	0.311	0.998	0.231	0.028
	T_{min}, RH_{max}	0.414	0.997	0.310	0.038	T_{mean}, RH_{mean}	0.357	0.997	0.249	0.032
	T_{mean}, RH_{min}	0.437	0.996	0.308	0.038	T_{max}, RH_{min}	0.380	0.997	0.266	0.035
	T_{mean}, RH_{mean}	0.468	0.996	0.334	0.043	T_{mean}, RH_{max}	0.426	0.896	1.822	0.211
Hy-Fh										
	T_{mean}, RH_{min}	0.724	0.991	0.392	0.049	T_{mean}, RH_{max}	0.560	0.995	0.367	0.038
	T_{mean}, RH_{mean}	0.732	0.990	0.398	0.049	T_{max}, RH_{mean}	0.617	0.993	0.404	0.041
	T_{mean}, RH_{max}	0.740	0.990	0.426	0.050	T_{max}, RH_{min}	0.622	0.993	0.402	0.042
	T_{min}, RH_{max}	0.799	0.988	0.434	0.054	T_{mean}, RH_{min}	0.645	0.992	0.489	0.043
	T_{min}, RH_{mean}	0.810	0.989	0.451	0.054	T_{mean}, RH_{mean}	0.656	0.992	0.471	0.044
Hy-Ls										
	T_{mean}, RH_{mean}	0.749	0.990	0.421	0.068	T_{mean}, RH_{max}	0.479	0.995	0.323	0.044
	T_{mean}, RH_{max}	0.761	0.989	0.453	0.069	T_{max}, RH_{mean}	0.525	0.995	0.335	0.048
	T_{mean}, RH_{min}	0.767	0.990	0.441	0.070	T_{max}, RH_{min}	0.608	0.993	0.414	0.055
	T_{min}, RH_{max}	0.798	0.989	0.462	0.073	T_{mean}, RH_{mean}	0.710	0.991	0.519	0.065
	T_{min}, RH_{mean}	0.811	0.989	0.469	0.074	T_{mean}, RH_{min}	0.731	0.990	0.538	0.066
Ls-Hy										
	T_{min}, RH_{mean}	1.219	0.975	0.677	0.115	T_{min}, RH_{min}	0.916	0.985	0.467	0.086
	T_{min}, RH_{min}	1.220	0.975	0.681	0.115	T_{min}, RH_{mean}	0.931	0.985	0.463	0.088
	T_{mean}, RH_{min}	1.229	0.975	0.687	0.116	T_{min}, RH_{max}	0.957	0.985	0.484	0.090
	T_{min}, RH_{max}	1.234	0.975	0.689	0.116	T_{mean}, RH_{min}	0.957	0.984	0.484	0.090
	T_{mean}, RH_{mean}	1.252	0.975	0.714	0.118	T_{mean}, RH_{mean}	0.991	0.984	0.517	0.093

(Continued)

Table 5 (Continued)

Station	RF					CatBoost				
	Input	RMSE (°C)	R^2	MAE (°C)	NRMSE	Input	RMSE (°C)	R^2	MAE (°C)	NRMSE
Ls-Fh										
	T_{mean}, RH_{min}	0.755	0.990	0.380	0.051	T_{min}, RH_{min}	0.651	0.992	0.305	0.044
	T_{min}, RH_{mean}	0.758	0.990	0.360	0.051	T_{min}, RH_{mean}	0.654	0.992	0.301	0.044
	T_{min}, RH_{min}	0.766	0.989	0.372	0.051	T_{mean}, RH_{min}	0.662	0.992	0.299	0.044
	T_{min}, RH_{max}	0.773	0.989	0.373	0.052	T_{min}, RH_{max}	0.663	0.992	0.306	0.044
	T_{mean}, RH_{mean}	0.784	0.989	0.388	0.053	T_{mean}, RH_{mean}	0.686	0.991	0.317	0.046

To explore the impact of the three-parameter combination on T_{dew} after the station exchange, we randomly combined nine single parameters and determined the top five accurate algorithms. The statistical results during the test period are shown in Table 6. Overall, the performance of the CatBoost algorithm was slightly better than that of the RF algorithm in most cases. Consistent with the trends in Table 5, before and after the station exchange, the top five parameter combinations did not change. Only the performance order changed, indicating that the two algorithms had certain stability in T_{dew} estimation. In Fh-Hy, the RF algorithm's performance after the exchange was improved compared to the that before the exchange (on average RMSE decreased by 22.5%, R^2 increased by 1.0%). In contrast, the performance of the CatBoost algorithm was slightly reduced (on average RMSE increased by 49.2%, R^2 decreased by 0.1%). The RF algorithm's estimation performance was slightly better than that of the CatBoost algorithm, indicating that the RF algorithm was more suitable for Fh-Hy and the station replacement may also effectively improve the algorithm performance. After changing stations, when the parameter combination was T_{mean}, RH_{min}, Rs_{mean}, the performance of the RF algorithm was best (RMSE = 0.292°C, R^2 = 1.000, MAE = 0.221°C and NRMSE = 0.026). In Ls-Hy, the CatBoost algorithm (on average RMSE = 0.810°C and R^2 = 0.989) was better than the RF algorithm (on average RMSE = 1.198°C and R^2 = 0.976). However, compared with the Longshan station before the exchange, the performance of RF (on average RMSE increased by 28.3%, R^2 decreased by 0.8%) and CatBoost algorithm (on average RMSE increased by 41.2%, R^2 decreased by 0.5%) were both reduced. It showed that the meteorological data of Huayuan Station could not be applied to Longshan Station, consistent with the performance of each algorithm of the two-parameter combinations. However, in the processing of other station exchanges, each algorithm showed good performance, confirming the feasibility of using the meteorological data of cross stations for T_{dew} estimation. It can be known from the three-parameter combinations that both had meteorological factors of T and RH. It showed that T and RH were the most effective meteorological factors for estimating T_{dew}, which was highly consistent with the conclusions of Mehdizadeh et al. [37]. Rs was more suitable for the CatBoost algorithm at Huayuan and Longshan stations.

Table 6: Statistical results of the two machine learning algorithms during the testing phase using three-parameter cross-station data at three stations

		RF					CatBoost			
Station	Input	RMSE (°C)	R^2	MAE (°C)	NRMSE	Input	RMSE (°C)	R^2	MAE (°C)	NRMSE
Fh-Hy										
	T_{mean}, RH_{min}, Rs_{mean}	0.292	1.000	0.221	0.026	T_{mean}, T_{min}, RH_{mean}	0.292	0.998	0.222	0.027
	T_{min}, RH_{mean}, RH_{min}	0.304	0.998	0.234	0.029	T_{max}, T_{min}, RH_{min}	0.325	0.998	0.246	0.031
	T_{mean}, T_{min}, RH_{min}	0.330	0.998	0.235	0.031	T_{mean}, RH_{max}, RH_{mean}	0.370	0.998	0.256	0.035
	T_{min}, RH_{max}, RH_{min}	0.356	0.998	0.260	0.034	T_{mean}, T_{min}, RH_{max}	0.388	0.998	0.272	0.037
	T_{mean}, RH_{mean}, RH_{min}	0.371	0.998	0.257	0.035	T_{max}, RH_{mean}, RH_{min}	0.460	0.997	0.321	0.043
Fh-Ls										
	T_{mean}, RH_{min}, Rs_{mean}	0.366	0.997	0.266	0.033	T_{mean}, T_{min}, RH_{mean}	0.289	0.998	0.201	0.026
	T_{min}, RH_{mean}, RH_{min}	0.373	0.997	0.272	0.034	T_{max}, T_{min}, RH_{min}	0.293	0.998	0.213	0.027
	T_{mean}, T_{min}, RH_{min}	0.391	0.997	0.290	0.036	T_{mean}, T_{min}, RH_{max}	0.348	0.998	0.238	0.032
	T_{min}, RH_{max}, RH_{min}	0.423	0.996	0.317	0.040	T_{mean}, RH_{max}, RH_{mean}	0.353	0.997	0.250	0.032
	T_{mean}, RH_{mean}, RH_{min}	0.451	0.996	0.329	0.041	T_{max}, RH_{mean}, RH_{min}	0.379	0.997	0.271	0.035
Hy-Fh										
	T_{mean}, RH_{max}, RH_{mean}	0.716	0.990	0.394	0.048	T_{mean}, RH_{mean}, Rs_{min}	0.451	0.996	0.253	0.030
	T_{mean}, RH_{mean}, RH_{min}	0.716	0.991	0.405	0.048	T_{mean}, RH_{mean}, Rs_{max}	0.493	0.995	0.294	0.033
	T_{mean}, RH_{max}, RH_{min}	0.727	0.990	0.399	0.049	T_{mean}, RH_{min}, Rs_{min}	0.510	0.995	0.317	0.034
	T_{max}, T_{mean}, RH_{min}	0.743	0.990	0.409	0.050	T_{mean}, RH_{min}, Rs_{max}	0.595	0.994	0.395	0.040
	T_{max}, T_{min}, RH_{min}	0.753	0.990	0.438	0.051	T_{mean}, RH_{mean}, Rs_{mean}	0.625	0.993	0.465	0.042

(Continued)

Table 6 (Continued)

		RF					CatBoost			
Station	Input	RMSE (°C)	R^2	MAE (°C)	NRMSE	Input	RMSE (°C)	R^2	MAE (°C)	NRMSE
Hy-Ls										
	T_{mean}, RH_{max}, RH_{mean}	0.733	0.990	0.427	0.067	T_{mean}, RH_{mean}, Rs_{max}	0.420	0.996	0.275	0.038
	T_{mean}, RH_{mean}, RH_{min}	0.735	0.990	0.431	0.067	T_{mean}, RH_{mean}, Rs_{min}	0.429	0.996	0.274	0.039
	T_{mean}, RH_{max}, RH_{min}	0.756	0.990	0.427	0.069	T_{mean}, RH_{min}, Rs_{min}	0.522	0.994	0.349	0.047
	T_{max}, T_{mean}, RH_{min}	0.759	0.990	0.438	0.069	T_{mean}, RH_{min}, Rs_{max}	0.549	0.994	0.394	0.050
	T_{max}, T_{min}, RH_{min}	0.825	0.989	0.510	0.075	T_{mean}, RH_{mean}, Rs_{mean}	0.726	0.990	0.534	0.066
Ls-Hy										
	T_{mean}, T_{min}, RH_{min}	1.182	0.976	0.659	0.111	T_{min}, RH_{max}, Rs_{max}	0.800	0.990	0.407	0.075
	T_{min}, RH_{mean}, RH_{min}	1.186	0.976	0.657	0.112	T_{min}, RH_{mean}, Rs_{min}	0.804	0.989	0.406	0.076
	T_{min}, RH_{max}, RH_{mean}	1.200	0.976	0.668	0.113	T_{min}, RH_{min}, Rs_{max}	0.809	0.988	0.435	0.076
	T_{mean}, T_{min}, RH_{mean}	1.204	0.976	0.671	0.114	T_{min}, RH_{mean}, Rs_{max}	0.817	0.989	0.416	0.077
	T_{min}, RH_{max}, RH_{min}	1.219	0.975	0.677	0.115	T_{min}, RH_{max}, Rs_{mean}	0.819	0.989	0.417	0.077
Ls-Fh										
	T_{min}, RH_{max}, RH_{min}	0.751	0.990	0.363	0.050	T_{min}, RH_{mean}, Rs_{min}	0.613	0.993	0.323	0.041
	T_{mean}, T_{min}, RH_{min}	0.752	0.990	0.366	0.050	T_{min}, RH_{max}, Rs_{max}	0.626	0.992	0.332	0.042
	T_{mean}, T_{min}, RH_{mean}	0.755	0.990	0.361	0.051	T_{min}, RH_{mean}, Rs_{max}	0.632	0.992	0.346	0.042
	T_{min}, RH_{mean}, RH_{min}	0.771	0.989	0.375	0.052	T_{min}, RH_{max}, Rs_{mean}	0.638	0.992	0.338	0.043
	T_{min}, RH_{max}, RH_{mean}	0.773	0.989	0.370	0.052	T_{min}, RH_{min}, Rs_{max}	0.640	0.992	0.360	0.043

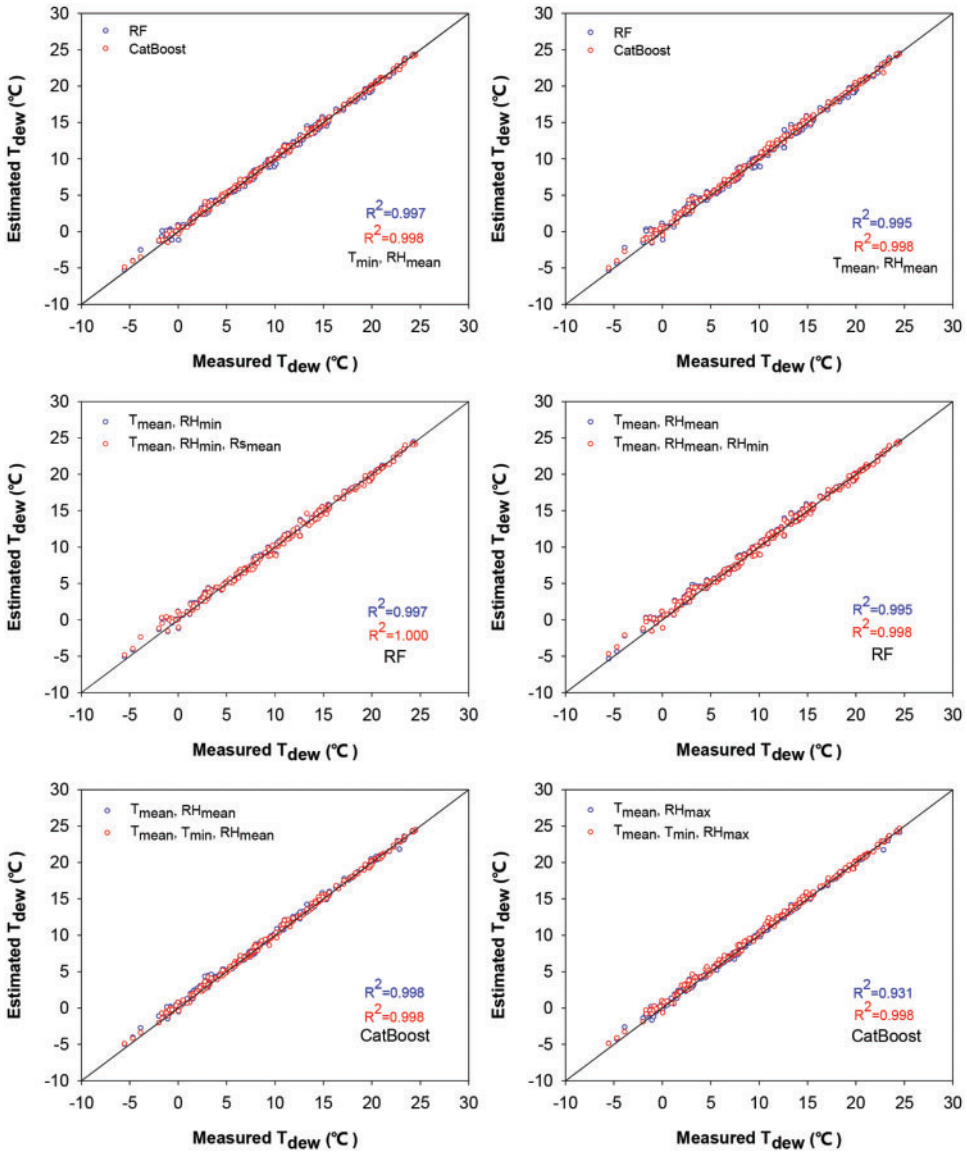

Figure 6: Scatter plots of the daily T_{dew} estimated by the algorithm and the corresponding measured values during the testing phase using cross-station data

To better compare each algorithm's impact to estimate T_{dew} at the target station using meteorological data from cross stations. We plotted estimated T_{dew} of some parameter combinations and measured T_{dew} during the test period in the case of station exchange and draw it in Fig. 6 (using Fh-Hy as an example). Fig. 6 showed that the performance of each algorithm in the case of two- and three-parameter combinations after the station exchange was very close. The difference in estimation accuracy was small, and the resulting scattered points were evenly distributed, all very close to the standard line. It showed that in the case of changing stations, the combination of T and RH parameters achieved high accuracy. Increasing the number of parameters made the algorithm more stable, but there were also a small number of algorithms with reduced accuracy. The CatBoost algorithm was slightly better than the RF algorithm. When the input parameter

combination was T_{mean}, RH_{mean}, the estimation accuracy of the CatBoost algorithm ($R^2 = 0.998$) was slightly better than that of the RF algorithm ($R^2 = 0.995$). Fig. 7 showed the distribution of the RMSE values of each algorithm in Tables 5 and 6 after each station exchange into a bar graph. As shown in Fig. 7, the CatBoost algorithm was more stable than the RF algorithm, and the distribution was more uniform. Among them, the stability of the algorithm was the most uniform under the conditions of Hy-Fh, Hy-Ls and Ls-Hy. The highest stability was obtained by the CatBoost algorithm in the case of Fh-Hy and Fh-Ls. The worst stability was found in the RF algorithm in the case of Ls-Hy (on average RMSE > 1.1°C). The comprehensive description confirmed the feasibility of using the meteorological data from cross stations to estimate T_{dew} at the target station. Each algorithm had good performance. It also showed that the modeling ability of the regional application of the algorithm had great potential. However, the CatBoost algorithm performed better than the RF algorithm and was more suitable for regional application in estimating daily T_{dew}. The best parameter combination was the two-parameter combination of T and RH, which had the highest cost performance.

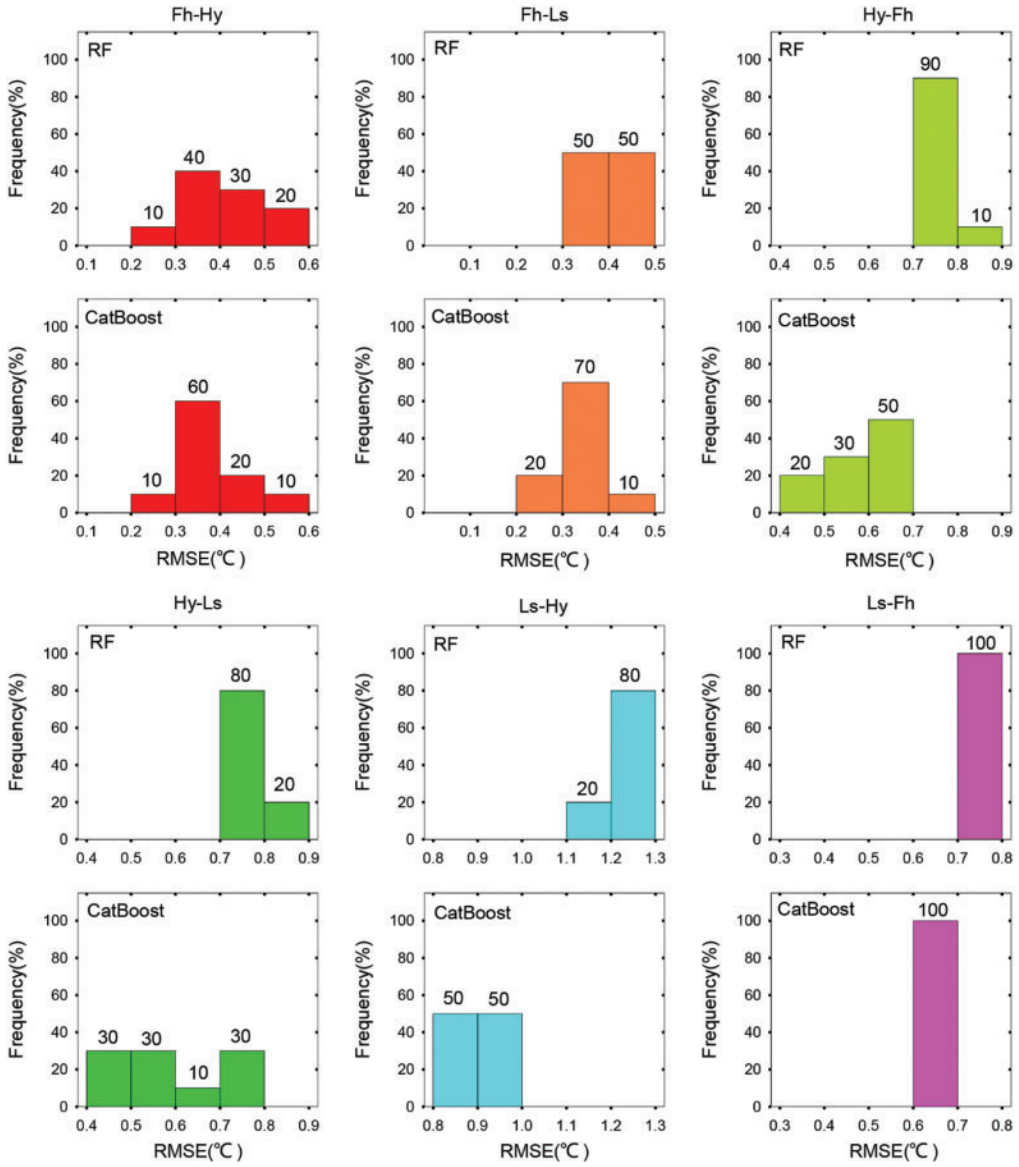

Figure 7: The RMSE distribution of the two algorithms after changing stations

3.3 Comprehensive Comparison of the Algorithm with Local Input and Cross-Station Input

Tables 7 and 8 show the average statistical results of the first five parameter combinations obtained when each algorithm using local and cross-station input data in the case of two-parameter and three-parameter combinations, respectively. It can be seen from Table 7 that in the two-parameter combinations, the CatBoost algorithm (on average RMSE = 0.568°C, R^2 = 0.990, MAE = 0.407°C and NRMSE = 0.056) was better than the RF algorithm (on average RMSE = 0.739°C, R^2 = 0.989, MAE = 0.432°C and NRMSE = 0.063). In Ls-Hy, each algorithm showed poor performance. Conversely, the best performance of each algorithm at Fh-Ls was better than the performance of the algorithms with local inputs. This indicated the feasibility of estimating daily T_{dew} at the target station using data from cross stations in different regions. This also provided the possibility in some developing countries to obtain unmeasured T_{dew} through regional modeling. As concluded by Shiri et al. [20]. They selected ANN and GEP algorithms for cross-station processing at two stations, Incheon and Seoul, Korea, to estimate T_{dew} values. The research provides guidance for regionalized modeling in Korea. As shown in Table 7, the average RMSE of the RF and CatBoost algorithms varied from 0.415 to 1.231°C and 0.265 to 0.950°C. It can be seen that the CatBoost algorithm was also superior to the RF algorithm in terms of instability. Tables 7 and 8 had very similar trends. In the three-parameter combinations, the average RMSE of RF and CatBoost algorithms ranged from 0.331 to 1.198°C and 0.246 to 0.810°C. The three-parameter combinations had higher stability than the two-parameter combinations, which was also consistent with the conclusions above. Under the combination of the three parameters of the local and cross-station inputs, the CatBoost algorithm exhibited the best accuracy and stability at Fenghuang station (on average RMSE = 0.315°C and R^2 = 0.998).

Table 7: Average statistical results of top five accurate algorithms of the two-parameter combinations under local and cross-station scenarios during the testing phase

	RF				CatBoost			
Station	RMSE (°C)	R^2	MAE (°C)	NRMSE	RMSE (°C)	R^2	MAE (°C)	NRMSE
Fh	0.429	0.997	0.277	0.029	0.265	0.999	0.186	0.018
Hy	0.855	0.988	0.490	0.081	0.553	0.995	0.336	0.052
Ls	0.952	0.984	0.544	0.087	0.681	0.991	0.372	0.062
Fh-Hy	0.466	0.996	0.347	0.044	0.412	0.984	0.570	0.070
Fh-Ls	0.415	0.997	0.300	0.037	0.357	0.977	0.559	0.067
Hy-Fh	0.761	0.990	0.420	0.051	0.620	0.993	0.427	0.042
Hy-Ls	0.777	0.989	0.449	0.071	0.611	0.993	0.426	0.056
Ls-Hy	1.231	0.975	0.690	0.116	0.950	0.985	0.483	0.089
Ls-Fh	0.767	0.989	0.375	0.052	0.663	0.992	0.306	0.044

Table 8: Average statistical results of top five accurate algorithms of the three-parameter combinations under local and cross-station scenarios during the testing phase

	RF				CatBoost			
Station	RMSE (°C)	R^2	MAE (°C)	NRMSE	RMSE (°C)	R^2	MAE (°C)	NRMSE
Fh	0.427	0.997	0.272	0.029	0.246	0.999	0.174	0.016
Hy	0.824	0.989	0.473	0.078	0.482	0.996	0.306	0.045
Ls	0.934	0.985	0.535	0.085	0.574	0.994	0.358	0.052

(Continued)

Table 8: Average statistical results of top five accurate algorithms of the three-parameter combinations under local and cross-station scenarios during the testing phase

Station	RF				CatBoost			
	RMSE (°C)	R^2	MAE (°C)	NRMSE	RMSE (°C)	R^2	MAE (°C)	NRMSE
Fh-Hy	0.331	0.998	0.241	0.031	0.367	0.998	0.263	0.035
Fh-Ls	0.401	0.997	0.295	0.037	0.332	0.998	0.235	0.030
Hy-Fh	0.731	0.990	0.409	0.049	0.535	0.995	0.345	0.036
Hy-Ls	0.762	0.990	0.447	0.069	0.529	0.994	0.365	0.048
Ls-Hy	1.198	0.976	0.666	0.113	0.810	0.989	0.416	0.076
Ls-Fh	0.760	0.990	0.367	0.051	0.630	0.992	0.340	0.042

Fig. 8 shows the average statistical values corresponding to each algorithm. As can be seen in Fig. 8, the three-parameter combinations were more stable. Because there were many parameters in the three-parameter combinations, the algorithm's performance to estimate T_{dew} was more stable. Still, the accuracy of each algorithm under the two-parameter and three-parameter combinations was not much different. Combining the data in Tables 2–6, adding meteorological parameters could improve the estimation accuracy of the algorithm, but adding extra parameters would also reduce the accuracy of the algorithm in estimating T_{dew}. This result also confirmed the previous conclusions. From another point of view, the correct choice of parameters is very important. Moreover, the experience of the scholar determines the amount of work involved in the research [48–50]. In Fig. 8, the average statistical results in the CatBoost algorithm varied more than the RF algorithm. For example, in Fh-Ls, the average MAE of the CatBoost algorithm was decreased by 58.1%, and the stability of the algorithm was improved under the three-parameter combinations. The estimation accuracy and stability of the CatBoost algorithm were better than the RF algorithm at the corresponding stations. The CatBoost algorithm was more suitable for regional application in estimating T_{dew}.

To better illustrate the importance of each meteorological factor in estimating T_{dew}, we have plotted the number of occurrences of each meteorological factor listed in Tables 2–6 in Fig. 9. The most frequently occurring meteorological factors in the RF algorithm were T_{min} (54 times) and RH_{min} (54 times). The most frequently occurring meteorological factor in the CatBoost algorithm were T_{min} (51 times), followed by RH_{mean} (45 times). According to the parameter combinations in Tables 2–6, Rs were rarely present under the three-parameter combinations. It can be seen that the meteorological factors T and RH were the most effective parameters. This result was consistent with the previous conclusion. Moreover, the meteorological factors T and RH were better obtained than Rs, which was also the advantage of choosing the combination algorithm of factors T and RH, which was consistent with the conclusion of Dong et al. [47,51]. Rs can be used as an alternative parameter to estimate T_{dew}. It can be seen from Table 8 that the minimum values and average values appeared more frequently in the meteorological factors T and RH. This reason may be because, in daily meteorological data changes, the minimum values and average values were closer to the values, while the maximum values deviated larger. Mehdizadeh et al. [38] also discovered this scenario. Moreover, the data processing speed of the CatBoost algorithm was better than that of the RF algorithm. The calculation CPU time required of both the CatBoost and RF algorithms was less than 1 s. The CatBoost algorithm has a very small advantage over the RF algorithm, which was also consistent with the results of Huang et al. [52]. The accuracy and stability of each algorithm in estimating T_{dew} were based on the performance

of scatter plots, radar charts, bar charts, and line charts. As a result, the CatBoost algorithm was the best and the most effective meteorological factors for the two input scenarios were T and RH. The most cost-effective parameter combination was the two-parameter combination of T and RH. For both algorithms, it was feasible to estimate T_{dew} at the target station using different regions' cross-station data. This conclusion also confirmed the modeling and estimation capabilities of the regional application of the algorithm.

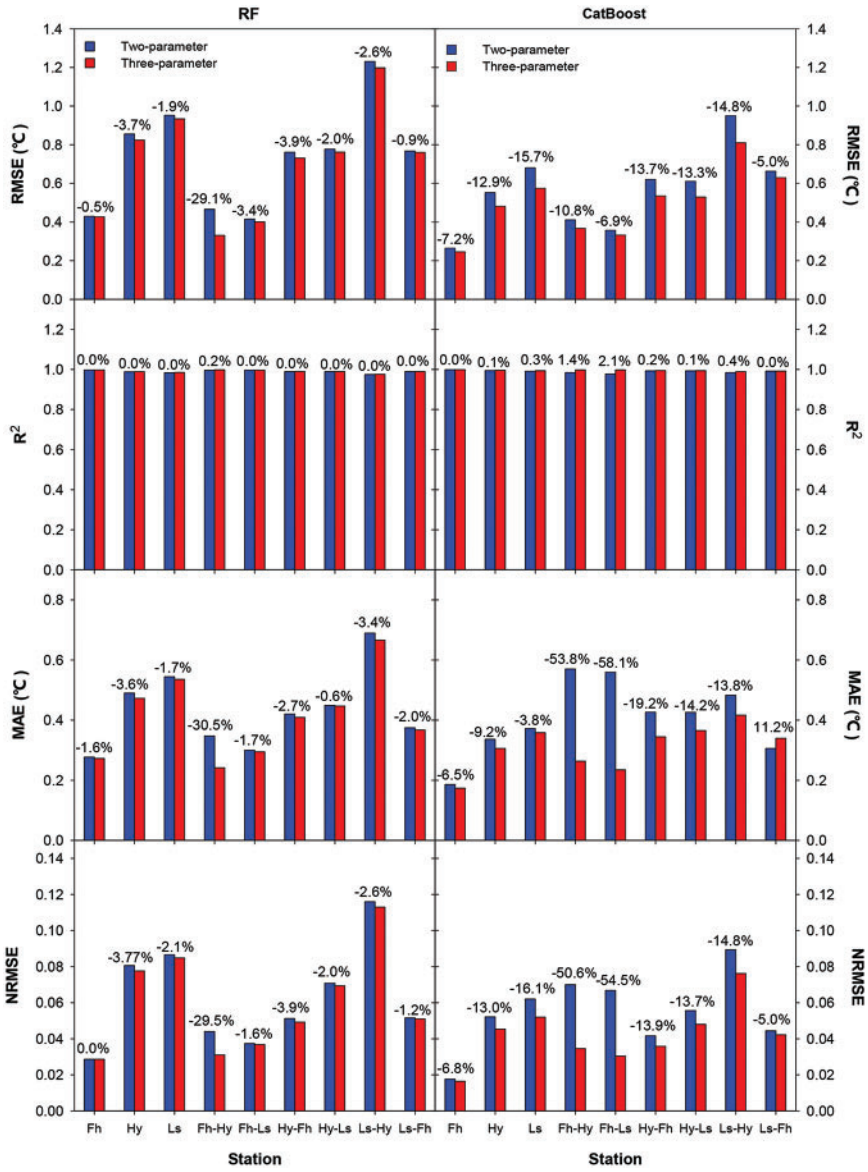

Figure 8: Percentage growth in statistical results of the two algorithms of the three-parameter combinations under local and cross-station scenarios relative to the average statistical indicators under the two-parameter combinations

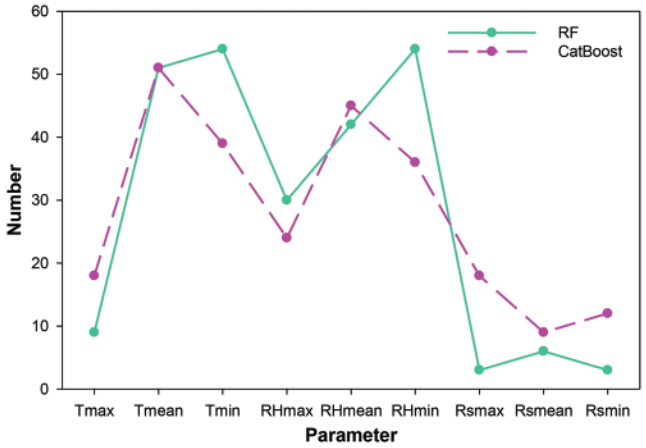

Figure 9: The number of occurrences of each parameter in all listed parameter combinations

4 Conclusions

This paper evaluated the applicability of a new algorithm (CatBoost) under two input scenarios (local and cross-station data) combined with limited meteorological data from different regional stations in China to accurately estimate daily T_{dew} and extend it to regional applications. The RF algorithm was also assessed for comparison. The daily routine meteorological data (including T_{max}, T_{min}, T_{mean}, RH_{max}, RH_{min}, RH_{mean}, Rs_{max}, Rs_{min} and Rs_{mean}) at three weather stations of Hunan from 2016 to 2019 were used to train and test the algorithms. The results showed that in the absence of complete meteorological parameters (with meteorological factor T), each machine learning algorithm achieved satisfactory estimation accuracy at the target station. During the testing phase of the two-input scenarios, the CatBoost algorithm was better than the RF algorithm. The accuracy and stability of most machine learning algorithms were positively correlated with the number of input parameters, and the performance of the algorithm was significantly better than that of the single-parameter when the two parameters were used as inputs. The algorithm performance difference was minuscule when two and three parameters were used as inputs. The top five accurate algorithms of the two-parameter combinations included T and RH, whose importance were greater than that of Rs, and the meteorological data were easier to obtain relative to Rs. The main meteorological factors were the minimum and average T and RH. Incorporation of Rs when estimating T_{dew} may reduce the algorithm performance. Therefore, the increase of parameters sometimes caused the increase of influencing factors, so that the performance of the algorithm may decline. When normal meteorological data are partially or wholly lacking in certain areas, meteorological data from cross-stations in different regions can be used to form various combinations of parameters as input to estimate T_{dew} at the target station. This conclusion confirmed the potential of both algorithms to extend local modeling to regional applications. Considering factors such as accuracy and stability, the CatBoost algorithm implemented regional modeling in China and even similar climate regions worldwide and estimates that T_{dew} had excellent potential. The most practical input parameter combination in the two input scenarios were T and RH. Differences in the study area and climate will also lead to different regional applicability of the algorithm and incomplete meteorological data. Some new hybrid machine learning algorithms have also been developed to obtain higher accuracy in estimating T_{dew}.

Acknowledgement: This study was jointly supported by the Shandong Provincial Natural Science Fund (ZR2020ME254 and ZR2020QD061). Thanks to the National Meteorological Information Center of China Meteorological Administration for offering the meteorological data.

Funding Statement: This research was supported by the Shandong Provincial Natural Science Fund (ZR2020ME254 and ZR2020QD061).

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

1. Blanco, J. M., Pena, F. (2008). Increase in the boiler's performance in terms of the acid dew point temperature: Environmental advantages of replacing fuels. *Applied Thermal Engineering, 28(7),* 777–784. DOI 10.1016/j.applthermaleng.2007.06.024.
2. Jradi, M., Riffat, S. (2014). Experimental and numerical investigation of a dew-point cooling system for thermal comfort in buildings. *Applied Energy, 132,* 524–535. DOI 10.1016/j.apenergy.2014.07.040.
3. Yang, Y., Ren, C., Yang, C., Tu, M., Luo, B. et al. (2021). Energy and exergy performance comparison of conventional, dew point and new external-cooling indirect evaporative coolers. *Energy Conversion and Management, 230,* 113824. DOI 10.1016/j.enconman.2021.113824.
4. Ali, M., Ahmad, W., Sheikh, N. A., Ali, H., Kousar, R. et al. (2021). Performance enhancement of a cross flow dew point indirect evaporative cooler with circular finned channel geometry. *Journal of Building Engineering, 35,* 101980. DOI 10.1016/j.jobe.2020.101980.
5. Robinson, P. J. (2000). Temporal trends in United States dew point temperatures. *International Journal of Climatology, 20(9),* 985–1002. DOI 10.1002/1097-0088(200007)20:9&<985::AID-JOC513>3.0.CO;2-W.
6. Kim, S., Singh, V. P., Lee, C., Seo, Y. (2015). Modeling the physical dynamics of daily dew point temperature using soft computing techniques. *Journal of Civil Engineering, 19(6),* 1930–1940. DOI 10.1007/s12205-014-1197-4.
7. Lawrence, M. G. (2005). The relationship between relative humidity and the dewpoint temperature in moist air: A simple conversion and applications. *Bulletin of the American Meteorological Society, 86(2),* 225–234. DOI 10.1175/BAMS-86-2-225.
8. Drezner, T. D. (2007). An analysis of winter temperature and dew point under the canopy of a common sonoran desert nurse and the implications for positive plant interactions. *Journal of Arid Environments, 69(4),* 554–568. DOI 10.1016/j.jaridenv.2006.11.003.
9. Hubbard, K. G., Mahmood, R., Carlson, C. (2003). Estimating daily dew point temperature for the northern great plains using maximum and minimum temperature. *Agronomy Journal, 95(2),* 323–328. DOI 10.2134/agronj2003.3230.
10. Sandstrom, M. A., Lauritsen, R. G., Changnon, D. (2004). A central-US summer extreme dew-point climatology (1949–2000). *Physical Geography, 25(3),* 191–207. DOI 10.2747/0272-3646.25.3.191.
11. Parlange, M. B., Katz, R. W. (2000). An extended version of the richardson model for simulating daily weather variables. *Journal of Applied Meteorology, 39(5),* 610–622. DOI 10.1175/1520-0450-39.5.610.
12. Hou, J., Huang, C., Zhang, Y., Guo, J. (2020). On the value of available MODIS and landsat8 OLI image pairs for MODIS fractional snow cover mapping based on an artificial neural network. *IEEE Transactions on Geoscience and Remote Sensing, 58(6),* 4319–4334. DOI 10.1109/TGRS.2019.2963075.
13. Kuter, S., Akyurek, Z., Weber, G. (2018). Retrieval of fractional snow covered area from MODIS data by multivariate adaptive regression splines. *Remote Sensing of Environment, 205,* 236–252. DOI 10.1016/j.rse.2017.11.021.
14. Kuter, S. (2021). Completing the machine learning saga in fractional snow cover estimation from MODIS terra reflectance data: Random forests versus support vector regression. *Remote Sensing of Environment, 255,* 112294. DOI 10.1016/j.rse.2021.112294.

15. Houborg, R., McCabe, M. F. (2018). A hybrid training approach for leaf area index estimation via cubist and random forests machine-learning. *ISPRS Journal of Photogrammetry and Remote Sensing, 135,* 173–188. DOI 10.1016/j.isprsjprs.2017.10.004.
16. Czyzowska-Wisniewski, E. H., van Leeuwen, W. J., Hirschboeck, K. K., Marsh, S. E., Wisniewski, W. T. (2015). Fractional snow cover estimation in complex alpine-forested environments using an artificial neural network. *Remote Sensing of Environment, 156,* 403–417. DOI 10.1016/j.rse.2014.09.026.
17. García-Gutiérrez, J., Mateos-García, D., Garcia, M., Riquelme-Santos, J. C. (2015). An evolutionary-weighted majority voting and support vector machines applied to contextual classification of LiDAR and imagery data fusion. *Neurocomputing, 163,* 17–24. DOI 10.1016/j.neucom.2014.08.086.
18. Shank, D. B., Hoogenboom, G., McClendon, R. W. (2008). Dewpoint temperature prediction using artificial neural networks. *Journal of Applied Meteorology and Climatology, 47(6),* 1757–1769. DOI 10.1175/2007JAMC1693.1.
19. Zounemat-Kermani, M. (2012). Hourly predictive levenberg–Marquardt ANN and multi linear regression models for predicting of dew point temperature. *Meteorology and Atmospheric Physics, 117(3),* 181–192. DOI 10.1007/s00703-012-0192-x.
20. 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. DOI 10.2166/nh.2013.229.
21. Nadig, K., Potter, W., Hoogenboom, G., McClendon, R. (2013). Comparison of individual and combined ANN models for prediction of air and dew point temperature. *Applied Intelligence, 39(2),* 354–366. DOI 10.1007/s10489-012-0417-1.
22. Mohammadi, K., Shamshirband, S., Petković, D., Yee, L., Mansor, Z. (2016). Using ANFIS for selection of more relevant parameters to predict dew point temperature. *Applied Thermal Engineering, 96,* 311–319. DOI 10.1016/j.applthermaleng.2015.11.081.
23. Kisi, O., Kim, S., Shiri, J. (2013). Estimation of dew point temperature using neuro-fuzzy and neural network techniques. *Theoretical and Applied Climatology, 114(3),* 365–373. DOI 10.1007/s00704-013-0845-9.
24. Baghban, A., Bahadori, M., Rozyn, J., Lee, M., Abbas, A. et al. (2016). Estimation of air dew point temperature using computational intelligence schemes. *Applied Thermal Engineering, 93,* 1043–1052. DOI 10.1016/j.applthermaleng.2015.10.056.
25. Dong, J., Wu, L., Liu, X., Fan, C., Leng, M. et al. (2020). Simulation of daily diffuse solar radiation based on three machine learning models. *Computer Modeling in Engineering & Sciences, 123(1),* 49–73. DOI 10.32604/cmes.2020.09014.
26. Wu, L., Fan, J. (2019). Comparison of neuron-based, kernel-based, tree-based and curve-based machine learning models for predicting daily reference evapotranspiration. *PLoS One, 14(5),* e217520. DOI 10.1371/journal.pone.0217520.
27. Han, Y., Wu, J., Zhai, B., Pan, Y., Huang, G. et al. (2019). Coupling a bat algorithm with xgboost to estimate reference evapotranspiration in the arid and semiarid regions of China. *Advances in Meteorology, 2019,* 9575782. DOI 10.1155/2019/9575782.
28. Wu, L., Huang, G., Fan, J., Zhang, F., Wang, X. et al. (2019). Potential of kernel-based nonlinear extension of arps decline model and gradient boosting with categorical features support for predicting daily global solar radiation in humid regions. *Energy Conversion and Management, 183,* 280–295. DOI 10.1016/j.enconman.2018.12.103.
29. Shiri, J. (2019). Prediction *vs.* estimation of dewpoint temperature: Assessing GEP, MARS and RF models. *Hydrology Research, 50(2),* 633–643. DOI 10.2166/nh.2018.104.
30. Attar, N. F., Khalili, K., Behmanesh, J., Khanmohammadi, N. (2018). On the reliability of soft computing methods in the estimation of dew point temperature: The case of arid regions of Iran. *Computers and Electronics in Agriculture, 153,* 334–346. DOI 10.1016/j.compag.2018.08.029.
31. Mehdizadeh, S., Behmanesh, J., Khalili, K. (2017). Using MARS, SVM, GEP and empirical equations for estimation of monthly mean reference evapotranspiration. *Computers and Electronics in Agriculture, 139,* 103–114. DOI 10.1016/j.compag.2017.05.002.

32. Deka, P. C., Patil, A. P., Yeswanth Kumar, P., Naganna, S. R. (2018). Estimation of dew point temperature using SVM and ELM for humid and semi-arid regions of India. *ISH Journal of Hydraulic Engineering, 24(2),* 190–197. DOI 10.1080/09715010.2017.1408037.
33. Amirmojahedi, M., Mohammadi, K., Shamshirband, S., Danesh, A. S., Mostafaeipour, A. et al. (2016). A hybrid computational intelligence method for predicting dew point temperature. *Environmental Earth Sciences, 75(5),* 415. DOI 10.1007/s12665-015-5135-7.
34. Fan, J., Yue, W., Wu, L., Zhang, F., Cai, H. et al. (2018). Evaluation of SVM, ELM and four tree-based ensemble models for predicting daily reference evapotranspiration using limited meteorological data in different climates of China. *Agricultural and Forest Meteorology, 263,* 225–241. DOI 10.1016/j.agrformet.2018.08.019.
35. Wong, P. K., Wong, K. I., Vong, C. M., Cheung, C. S. (2015). Modeling and optimization of biodiesel engine performance using kernel-based extreme learning machine and cuckoo search. *Renewable Energy, 74,* 640–647. DOI 10.1016/j.renene.2014.08.075.
36. Feng, Y., Jia, Y., Cui, N., Zhao, L., Li, C. et al. (2017). Calibration of hargreaves model for reference evapotranspiration estimation in sichuan basin of Southwest China. *Agricultural Water Management, 181,* 1–9. DOI 10.1016/j.agwat.2016.11.010.
37. Feng, Y., Peng, Y., Cui, N., Gong, D., Zhang, K. (2017). Modeling reference evapotranspiration using extreme learning machine and generalized regression neural network only with temperature data. *Computers and Electronics in Agriculture, 136,* 71–78. DOI 10.1016/j.compag.2017.01.027.
38. Mehdizadeh, S., Behmanesh, J., Khalili, K. (2017). Application of gene expression programming to predict daily dew point temperature. *Applied Thermal Engineering, 112,* 1097–1107. DOI 10.1016/j.applthermaleng.2016.10.181.
39. Lu, X., Ju, Y., Wu, L., Fan, J., Zhang, F. et al. (2018). Daily pan evaporation modeling from local and cross-station data using three tree-based machine learning models. *Journal of Hydrology, 566,* 668–684. DOI 10.1016/j.jhydrol.2018.09.055.
40. Karimi, S., Kisi, O., Kim, S., Nazemi, A. H., Shiri, J. (2017). Modelling daily reference evapotranspiration in humid locations of South Korea using local and cross-station data management scenarios. *International Journal of Climatology, 37(7),* 3238–3246. DOI 10.1002/joc.4911.
41. Fan, J., Wu, L., Zhang, F., Cai, H., Zeng, W. et al. (2019). Empirical and machine learning models for predicting daily global solar radiation from sunshine duration: A review and case study in China. *Renewable and Sustainable Energy Reviews, 100,* 186–212. DOI 10.1016/j.rser.2018.10.018.
42. Dorogush, A. V., Ershov, V., Gulin, A. (2018). CatBoost: gradient boosting with categorical features support. *Machine Learning,* 1–7. https://arxiv.org/abs/1810.11363.
43. Aler, R., Galván, I. M., Ruiz-Arias, J. A., Gueymard, C. A. (2017). Improving the separation of direct and diffuse solar radiation components using machine learning by gradient boosting. *Solar Energy, 150,* 558–569. DOI 10.1016/j.solener.2017.05.018.
44. Fan, J., Wang, X., Wu, L., Zhou, H., Zhang, F. et al. (2018). Comparison of support vector machine and extreme gradient boosting for predicting daily global solar radiation using temperature and precipitation in humid subtropical climates: A case study in China. *Energy Conversion and Management, 164,* 102–111. DOI 10.1016/j.enconman.2018.02.087.
45. Breiman, L. (2001). Random forests. *Machine Learning, 45(1),* 5–32. DOI 10.1023/A:1010933404324.
46. Fan, J., Wu, L., Ma, X., Zhou, H., Zhang, F. (2020). Hybrid support vector machines with heuristic algorithms for prediction of daily diffuse solar radiation in air-polluted regions. *Renewable Energy, 145,* 2034–2045. DOI 10.1016/j.renene.2019.07.104.
47. Dong, J., Wu, L., Liu, X., Li, Z., Gao, Y. et al. (2020). Estimation of daily dew point temperature by using bat algorithm optimization based extreme learning machine. *Applied Thermal Engineering, 165,* 114569. DOI 10.1016/j.applthermaleng.2019.114569.
48. Owolabi, K. M., Baleanu, D. (2021). Emergent patterns in diffusive Turing-like systems with fractional-order operator. *Neural Computing and Applications, 1,* 1–18. DOI 10.1007/s00521-021-05917-8.

49. Momigliano, P., Florin, A., Merilä, J. (2021). Biases in demographic modeling affect our understanding of recent divergence. *Molecular Biology and Evolution, 38(7),* 2967–2985. DOI 10.1093/molbev/msab047.
50. Rybak, I., Schwarzmeier, C., Eggenweiler, E., Rüde, U. (2021). Validation and calibration of coupled porous-medium and free-flow problems using pore-scale resolved models. *Computational Geosciences, 25(2),* 621–635. DOI 10.1007/s10596-020-09994-x.
51. Dong, J., Liu, X., Huang, G., Fan, J., Wu, L. et al. (2020). Comparison of four bio-inspired algorithms to optimize KNEA for predicting monthly reference evapotranspiration in different climate zones of China. *Computers and Electronics in Agriculture, 186,* 106211. DOI 10.1016/j.compag.2021.106211.
52. Huang, G., Wu, L., Ma, X., Zhang, W., Fan, J. et al. (2019). Evaluation of CatBoost method for prediction of reference evapotranspiration in humid regions. *Journal of Hydrology, 574,* 1029–1041. DOI 10.1016/j.jhydrol.2019.04.085.