atmosphere

MDPI

Article

Research on Modeling Weighted Average Temperature Based on the Machine Learning Algorithms

Kai Li [1], Li Li [1,*], Andong Hu [2], Jianping Pan [1], Yixiang Ma [1] and Mingsong Zhang [1]

[1] Research Center of Beidou Navigation and Environmental Remote Sensing, Suzhou University of Science and Technology, Suzhou 215009, China; 2213021026@post.usts.edu.cn (K.L.); 2113021101@post.usts.edu.cn (J.P.); 2113021100@post.usts.edu.cn (Y.M.); 2113021104@post.usts.edu.cn (M.Z.)

[2] Cooperative Institute for Research in Environmental Sciences, University of Colorado Boulder, Boulder, CO 80309, USA; andong.hu@colorado.edu

* Correspondence: gszl.lili@usts.edu.cn

Abstract: In response to the nonlinear fitting difficulty of the traditional weighted average temperature (T_m) modeling, this paper proposed four machine learning (ML)-based T_m models. Based on the seven radiosondes in the Yangtze River Delta region from 2014 to 2019, four forecasting ML-based T_m models were constructed using Light Gradient Boosting Machine (LightGBM), Support Vector Machine (SVM), Random Forest (RF), and Classification and Regression Tree (CART) algorithms. The surface temperature (T_s), water vapor pressure (E_s), and atmospheric pressure (P_s) were identified as crucial influencing factors after analyzing their correlations to the T_m. The ML-based T_m models were trained using seven radiosondes from 2014 to 2018. Then, the mean bias and root mean square error (*RMSE*) of the 2019 dataset were used to evaluate the accuracy of the ML-based T_m models. Experimental results show that the overall accuracy of the LightGBM-based T_m model is superior to the SVM, CART, and RF-based T_m models under different temporal variations. The mean *RMSE* of the daily LightGBM-based T_m model is reduced by 0.07 K, 0.04 K, and 0.13 K compared to the other three ML-based models, respectively. The mean *RMSE* of the monthly LightGBM-based T_m model is reduced by 0.09 K, 0.04 K, and 0.11 K, respectively. The mean *RMSE* of the quarterly LightGBM-based T_m model is reduced by 0.09 K, 0.04 K, and 0.11 K, respectively. The mean bias of the LightGBM-based T_m model is also smaller than that of the other ML-based T_m models. Therefore, the LightGBM-based T_m model can provide more accurate T_m and is more suitable for obtaining GNSS precipitable water vapor in the Yangtze River Delta region.

Keywords: machine learning; atmospheric weighted average temperature; Yangtze River Delta; radiosonde

Citation: Li, K.; Li, L.; Hu, A.; Pan, J.; Ma, Y.; Zhang, M. Research on Modeling Weighted Average Temperature Based on the Machine Learning Algorithms. *Atmosphere* **2023**, *14*, 1251. https://doi.org/10.3390/atmos14081251

Academic Editor: Stephan Havemann

Received: 11 July 2023
Revised: 28 July 2023
Accepted: 2 August 2023
Published: 7 August 2023

1. Introduction

Global Navigation Satellite System (GNSS) precipitable water vapor (PWV) can be obtained through the inversion of several meteorological parameters and the zenith total delay (ZTD) of the GNSS signal. The process of PWV calculation also involves critical parameters such as zenith hydrostatic delay (ZHD), zenith wet delay (ZWD), and weighted average temperature (T_m) [1]. It is evident that T_m is a crucial conversion parameter for obtaining high-precision PWV. T_m is the continuous integration of water vapor pressure and temperature in the atmosphere from the Earth's surface to the top of the troposphere. Water vapor pressure and temperature can be obtained from atmospheric reanalysis data or directly from radiosondes [2]. However, the low temporal resolution of atmospheric reanalysis data and radiosondes prevents users from accessing real-time information on meteorological parameters [3]. Therefore, a high-accuracy T_m model is essential to improve the accuracy and practicality of obtaining real-time GNSS-PWV.

The traditional T_m model is an empirical model depending on the periodic variation of the T_m phase and amplitude [4,5]. According to Bevis et al., based on 8718 radiosonde records covering a range of 27° to 65° latitude in North America, a global T_m model was

constructed using one-dimensional linear regression based on the surface temperature [6]. Yao et al. combined the Bevis model with the global pressure and temperature (GPT) to construct the global weighted mean temperature (GWMT) model, which provides global T_m estimates [7]. Guo, et al. constructed a local T_m model that exhibited overall higher accuracy compared to the Bevis model based on seven radiosondes data from 2015 to 2017 in the Yangtze River Delta region [8].

The abovementioned empirical T_m models are not in consideration of the nonlinear relationship between T_m and spatiotemporal meteorological factors. Researchers have created some nonlinear T_m models to address the inaccurate calculation of linear empirical T_m models [9]. Machine learning methods, renowned for their potent nonlinear fitting capabilities, have been extensively used to improve the accuracy of T_m models. Thus, numerous scholars have successfully constructed ML-based T_m models, which have demonstrated more accuracy than traditional T_m models in extensive experiments [9–14]. Ding et al. demonstrated that their back propagation neural networks (BPNN) T_m model outperformed conventional models [15]. Sun et al. constructed T_m models based on the random forest (RF), BPNN, and generalized regression neural network (GRNN) algorithms. These T_m models are able to obtain high-accuracy T_m [10]. Cai et al. used artificial neural networks to develop a high-accuracy hybrid T_m model for the China region [16].

In addition to the T_m model, machine learning algorithms have demonstrated significant success in wind power prediction, air temperature estimation, and landslide disaster prevention. It is possible to accurately predict the fluctuations and trends in wind power and prevent landslide disasters. Yun et al. combined the light gradient boosting machine (LightGBM) model with a convolutional neural network model to construct an ultra-short-term wind power prediction model, from which they obtained high-precision wind power prediction [17]. Mohamed Saber and his colleagues developed a river valley flash flood prediction model based on the LightGBM algorithms. Their multiple field tests confirmed the effectiveness of the LightGBM-based model in predicting flash floods [18]. Wind speed modeling was performed by Morshed-Bozorgdel and his team utilizing the stacking ensemble machine learning (SEML) method. The results revealed that the performance of the base algorithms was significantly affected by the incorporation of the SEML method in wind speed modeling. The highest correlation coefficient (R) achieved in wind speed modeling at the sixteen stations using the SEML was 0.89. It was observed that the implementation of the SEML method led to an increase in the accuracy of wind speed modeling by more than 43% [19]. A modified air temperature estimation model was proposed by Xu, et al. [20], which combined the temperature–vegetation index method with a multiple regression model. Statistical results demonstrated that the modified method yielded higher accuracy in estimating air temperature for the winter wheat planted area than the traditional regression method. The aforementioned achievements demonstrate the broad prospects and potential of machine learning algorithm models in forecasting applications. The machine learning methods are of great significance for T_m modeling and improving their accuracy in GNSS-PWV studies [9,17,18,21,22].

The accuracy of the traditional T_m models is limited because the nonlinear relationships between T_m and spatiotemporal meteorological factors have not been considered in previous empirical T_m models. In this paper, seven radiosondes in the Yangtze River Delta region from 2014 to 2018 will be utilized to construct the ML-based T_m models, which are based on the RF, LightGBM, support vector machine (SVM), and classification and regression tree (CART) algorithms. Then, the accuracy of ML-based T_m models will be assessed using 2019 radiosondes. Ultimately, the most appropriate ML-based T_m model will be selected for the Yangtze River Delta region.

2. Data and Methods

2.1. Data

The study area is the Yangtze River Delta region of China. The region is located between 114° E and 124° E and 26° N to 36° N. Figure 1 depicts the topography of the

region. The radiosondes were obtained at http://weather.uwyo.edu/ (accessed on 28 May 2023). The data covers the period from 2014 to 2019, with a time resolution of 12 h. Table 1 provides the location information of the radiosondes. Four main variables, such as water vapor pressure (E_s), surface temperature (T_s), atmospheric pressure (P_s), and T_m, will be used in the ML-based T_m models for data training and accuracy analysis.

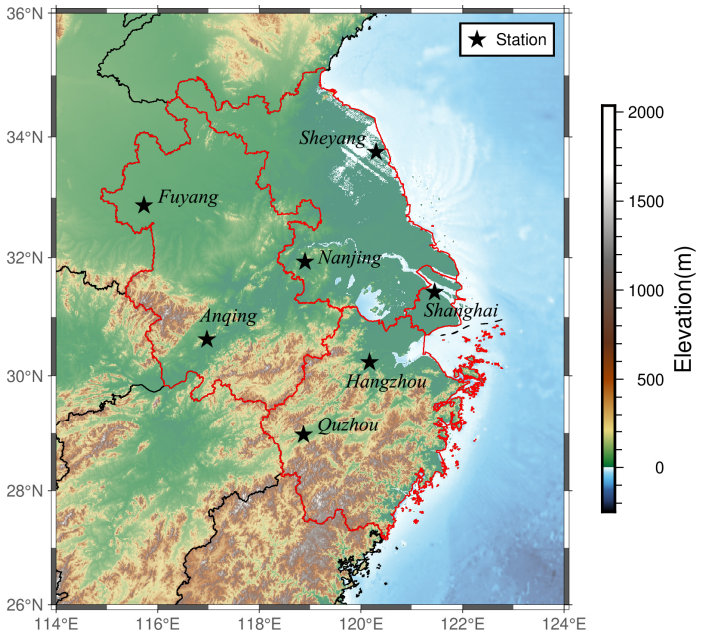

Figure 1. The radiosondes distribution in the Yangtze River Delta region.

Table 1. The location information of radiosondes in the Yangtze River Delta region.

Region	Site Number	Location [°]	Elevation [m]
Hangzhou	58457	(30.23° N, 120.16° E)	43.00
Quzhou	58633	(28.96° N, 118.86° E)	71.00
Shanghai	58362	(31.40° N, 121.46° E)	4.00
Anqing	58424	(30.53° N, 117.05° E)	20.00
Fuyang	58203	(32.86° N, 115.73° E)	33.00
Nanjing	58238	(32.00° N, 118.80° E)	7.00
Sheyang	58150	(33.76° N, 120.25° E)	7.00

2.2. Methods

2.2.1. Machine Learning Algorithms

1. LightGBM

The LightGBM is an efficient machine learning framework that utilizes gradient lifting algorithms to train scalable models quickly. The exclusive feature bundling technology is used to reduce the number of features and speed up training in the experiment. In addition, the gradient-based one-side sampling technique is used to filter out the samples with small prediction errors to minimize the training time. Compared to the traditional level-wise growth strategy, LightGBM prioritizes the selection of leaf nodes with the highest splitting gain for growth after finding an optimal split node. This approach allows for faster identification of leaf nodes that contribute to minimizing the loss function but may potentially lead to overfitting. Therefore, LightGBM provides parameters to control the growth strategy. LightGBM supports multi-threaded parallel learning and can also operate in a distributed environment to handle large-scale datasets.

Therefore, LightGBM has fast efficiency, high accuracy, low memory usage, custom loss function, and scalability advantages. It is suitable for discovering the nonlinear

relationships between T_m and other meteorological influencing factors, such as temperature, humidity, wind speed, and precipitation [23].

A predictive LightGBM-based model can be developed to forecast future meteorological variables by leveraging historical meteorological observations, geographical features, and seasonal variations. Moreover, the LightGBM-based model enables the strength and likelihood of weather phenomena prediction, such as heavy rain, typhoons, and tornadoes. The LightGBM adeptly captures intricate relationships among diverse influencing factors, encompassing meteorological conditions, geographic topography, and environmental data, thereby providing timely alerts and decision-making support [24–26].

2. RF

The Random Forest is an ensemble learning technique that utilizes bagging to create multiple distinct training datasets and employs multiple CART for prediction. The prediction outcome is determined by either the highest voting score or the average value. The core concept behind this method is that the collective judgments of multiple classifiers yield superior results compared to those of a single classifier, as illustrated in Equation (1).

$$\hat{y} = \frac{1}{B}\sum_{i=1}^{B} f_i(X) \tag{1}$$

where B represents the number of classification regression trees in the RF. The segmentation points of the regression tree in the model are determined by the minimum regression error, which is the weighted sum of the subset regression error, as shown in Equations (2) and (3).

$$K = \frac{M_L}{M} \times K(B_L) + \frac{M_R}{M} \times K(B_R) \tag{2}$$

$$M(B) = \frac{\sum_{i=1}^{M} (y_i - \overline{y})^2}{M} \tag{3}$$

where $K(B_L)$ and $K(B_R)$ represents the regression error for left and right subsets. M_L and M_R are the left and right subsets, respectively. M represents the total number of samples. The $K(B)$ refers to the regression error.

The RF provides accurate prediction by combining historical meteorological data and processing complex relationships and features between them to construct models. The RF application in meteorology fully utilizes its advantages in adaptability to complex nonlinear relationships and robustness to outliers. It can be a powerful tool in weather forecasting, model evaluation, and meteorological data analysis [27–29].

3. SVM

The SVM is a machine learning algorithm derived from statistical theory primarily applied to classification and regression problems. It maps the original data from the input space to a high-dimensional feature space through nonlinear mapping. Its purpose is to construct an optimal classification hyperplane in the feature space, separating samples of different categories. It maximizes the distance between the classification hyperplane and the support vectors, thereby achieving better generalization capability, as Equation (4) shows.

$$f(x) = w^T x + b \tag{4}$$

where w is the weight vector, b is the bias term, and x is the input variable. As a predictive model, the SVM can solve nonlinear problems between variables and demonstrates good generalization ability and robustness to outliers. The model has broad potential and offers promising prospects for forecasting meteorological variables [21,30,31].

4. CART

The CART is a common decision tree algorithm. It is frequently employed for regression and classification problems. It aims to split the input space into different regions

and assign each region a category or regression value. When the CART is applied to the regression analysis of meteorological variables, it selects a feature as the root node, and the entire training dataset is used as the node's dataset. The result of Equation (5) can be used to evaluate the splitting effectiveness of each feature in the current node.

$$MSE = \frac{\sum_{i=1}^{n}(Y_i - Y_{mean})^2}{n} \quad (5)$$

where MSE is the mean squared error, n is the sample size of the current node, Y_i is the target variable, Y_{mean} is the mean of current node sample. The MSE of each target variable at different thresholds was calculated by CART and selects the target variable and threshold corresponding to the minimum MSE as the optimal split. Then, repeat the process continuously until the stopping criteria are met. Finally, the mean or median is chosen as its regression value in the node.

The CART demonstrates robust performance and flexibility in regression tasks, rendering it extensively applied for predictions across diverse domains. It solves nonlinear problems and disposes of outliers present within multiple features [32–34].

While the ML-based T_m model can address the issue of nonlinear fitting models, it also exhibits certain limitations. For instance, machine learning algorithms are susceptible to the quality of input data. The presence of noise, missing values, or outliers in the input data can affect the model's performance and accuracy. There are various ML algorithms, and selecting an appropriate model is crucial for predictive performance. Additionally, model parameter tuning is necessary to obtain the best prediction results. The problems of overfitting and underfitting may arise in ML-based models. Overfitting occurs when the model performs well on the training data but poorly on unseen data, while underfitting indicates that the model fails to capture the complex relationships within the data, leading to suboptimal performance. To mitigate the impact of such issues on model accuracy, rigorous data preprocessing and model optimization are required.

2.2.2. Evaluations

The bias and root mean square error (*RMSE*) are used as the accuracy evaluation metrics in this experiment.

$$X_{BIAS} = \frac{\sum_{i=1}^{N}\left|x_{obs,i} - x_{model,i}\right|}{n} \quad (6)$$

The bias measures the gap between predicted values ($X_{model,i}$) and the observed values ($X_{obs,i}$). It represents the average error between the observed value ($X_{obs,i}$) and the predicted values ($X_{model,i}$), as Equation (6) shows.

$$X_{RMSE} = \sqrt{\frac{\sum_{i=1}^{N}(x_{obs,i} - x_{model,i})^2}{N}} \quad (7)$$

The *RMSE* serves as a metric for assessing the accuracy of regression models. It measures the bias between the predicted values ($X_{model,i}$) and the observed values ($X_{obs,i}$). By computing the square root of bias, *RMSE* provides a single value to represent the model's overall performance. As depicted in Equation (7), a lower *RMSE* signifies higher accuracy, while a higher *RMSE* implies more significant prediction errors.

3. Results and Discussion

Figure 2 illustrates the modeling workflow of ML-based algorithms. The data was first preprocessed, and then modeling was carried out using four distinct ML algorithms. Subsequently, modeling optimization is performed. Finally, the accuracy of the models is evaluated using the *RMSE* and bias.

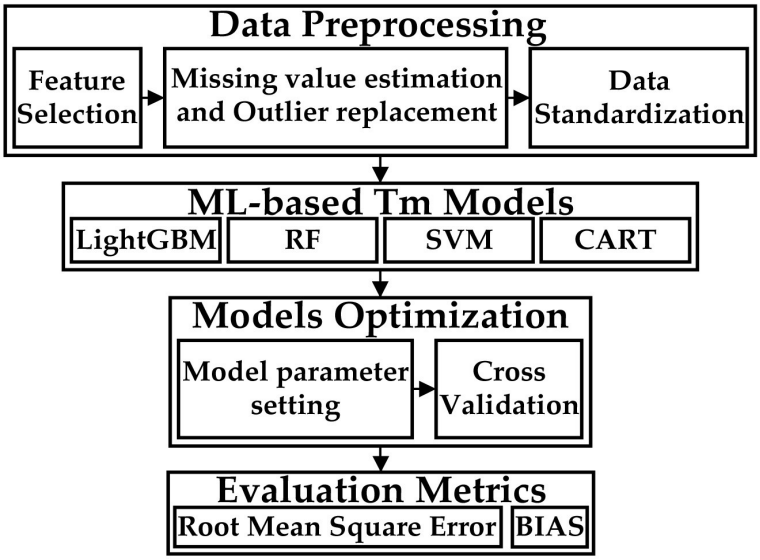

Figure 2. Modeling Procedures of ML-based T_m models.

3.1. Feature Selection

3.1.1. Correlation

The T_S, P_s, and E_s serve as indispensable input data in the T_m calculation process. Their spatiotemporal variations significantly influence the weighted distribution of the data. The T_S variability determines the weighting temperature from different locations, with higher T_S exerting a more significant influence on the result. Moreover, when considering the vertical temperature distribution, the changes in P_s are also employed as essential weights in the T_m computation, assigning different weights to the temperature at various pressure levels. Additionally, the impact of water vapor pressure and humidity on temperature is particularly notable under humid climatic conditions, potentially assigning higher weights to T_m from moist air.

To select input features of models, the Pearson coefficient was used for the correlation analysis between T_m and T_s, E_s, and P_s [35–38].

$$R = \frac{\sum_{i=1}^{n}(X_i - X) \times (Y_i - Y)}{\sqrt{\sum_{i=1}^{n}(X_i - X)^2} \times \sqrt{\sum_{i=1}^{n}(Y_i - Y)^2}} \quad (8)$$

where n represent the sample number, and X and Y represent different variables. The correlation criteria are as follows. $|R| \geq 0.81$ represents an extremely strong correlation; 0.61–0.80 represents a strong correlation; 0.41–0.60 represents a moderately correlated; 0.21–0.40 represents a weakly correlated; and 0.0–0.20 represents a minimal correlated or no correlation. Figure 3a–c shows the R of T_m and E_s, T_s, P_s are 0.96, 0.95, and -0.87, indicating an extremely strong correlation between T_m and E_s, T_s, P_s.

3.1.2. Collinearity

As shown in Table 2, there is a strong correlation among the independent variables, which implies severe multicollinearity. The multicollinearity can introduce problems in the model, such as unstable coefficient estimates and difficulties in interpreting the impact of individual variables.

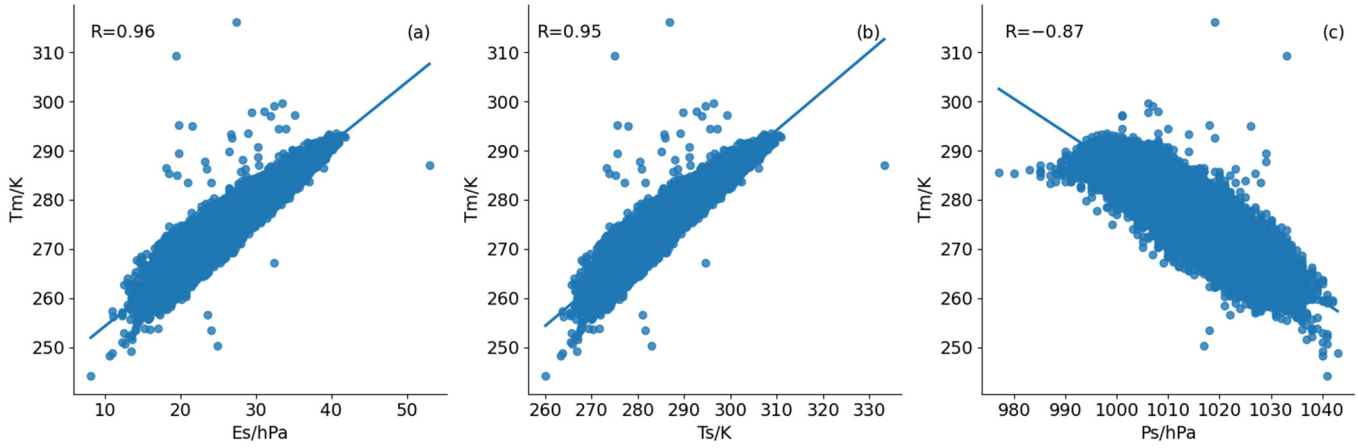

Figure 3. The linear correlations between T_m and E_s, T_s, and P_s, where (**a**) is the correlation analysis diagram of T_m and E_s where (**b**) is the correlation analysis diagram of T_m and T_s where (**c**) is the correlation analysis diagram of T_m and P_s.

Table 2. Correlations, tolerances, and variance inflation factors (VIF) between variables.

Dependent Variable	Independent Variable	R	Independent Variable	Independent Variable	R	R^2	Tol	VIF
T_m	T_s	0.95	T_s	P_s	−0.85	0.68	0.32	3.1
T_m	P_s	−0.87	T_s	E_s	0.91	0.83	0.17	5.9
T_m	E_s	0.96	P_s	E_s	−0.85	0.70	0.30	3.3

The tolerance ($Tol = 1 - R^2$) was used to evaluate multicollinearity among independent variables in the experiment. The R^2 represents the square of the correlation coefficient between two different variables and is referred to the determination coefficient. Table 2 shows that the R^2 and Tol between the T_s and P_s are 0.68 and 0.32. Also, the R^2 and Tol between T_s and E_s are 0.83 and 0.17, and the R^2 and Tol between P_s and Es are 0.70 and 0.30. If the Tol is greater than 0.2, we can conclude that there is no multicollinearity among the independent variables. However, the Tol between T_s and E_s is less than 0.2. It suggests the possible presence of severe multicollinearity. The VIF can be used to further verification of their multicollinearity, which is calculated by $VIF = 1/(1 - R^2)$. From Table 2, the VIF between T_s and E_s is 5.9. It can be concluded that there is no severe multicollinearity between T_s and E_s since their VIF is less than 10 [39,40].

Based on the correlation analysis, it is evident that the key input variables for the ML-based models are T_s, P_s, and E_s in the study, whereas the output variable is T_m.

3.2. The ML-Based T_m Modeling

3.2.1. Data Preprocessing

After determining the modeling variables, the dataset mentioned in Section 2.1 must be partitioned accordingly. The training dataset consists of radiosondes collected in the Yangtze River Delta region during 2014–2018. The validation dataset consists of radiosondes in 2019.

To ensure the modeling accuracy, the preprocessing of outliers and missing data in the dataset is necessary. Two approaches were used to replace the outliers in the study. One method is interpolation when the outliers are moderately outliers [39]. Another one is the k-nearest neighbor (KNN) algorithm used for extreme outliers. The KNN relies on the distances between samples in the feature space to perform classification and prediction, which can improve data reliability. Interpolation was employed based on the numerical range to fill in the gaps of missing data.

Considering the different value ranges among variables. To avoid focusing more on features with a larger range during the training procedure of modeling, data standardization

is necessary. It can improve the convergence speed, robustness, and interpretability of models [41,42]. The linear normalization was chosen for standardization, which maps the original dataset to the range of [0, 1] using the following formula.

$$x' = \frac{(x - \mathrm{MIN}(x))}{(\mathrm{MAX}(x) - \mathrm{MIN}(x))} \quad (9)$$

3.2.2. The Model Optimization

Model optimization is the most crucial step in the modeling procedure. It includes the selections of cross-validation and model parameters, such as learning rate, maximum depth, and iteration number.

The GridSearch algorithm is employed to automatically search for the best training parameters. It iteratively trains to select the optimal parameters of ML-based T_m models. After a series of iterations, the optimal learning rate, maximum depth, and iteration number for the LightGBM-based T_m model are determined to be 0.03, 30, and 700, respectively. The optimal kernel function and penalty factor for the SVM-based T_m model are determined to be radial basis function (RBF) kernel and 1, respectively. The optimal number of classifiers and maximum depth for the RF and CART-based T_m models are 100 and 30, respectively.

The main step of K-fold cross-validation is dividing the dataset into 15 groups. Subsequently, one group is randomly chosen as the validation dataset, while the remaining *K-1* groups would be the training dataset. The training dataset was used for the model's training, and the validation dataset was used to evaluate the model's accuracy. The procedure will repeat *K* times; the average accuracy rate is the final evaluation metric of the models. After completing optimization, the optimized models will be employed to predict T_m during 2019 in the Yangtze River Delta region.

3.3. Accuracy Analysis

In this section, to ensure the rigors of accuracy analysis, the ML-based T_m models were evaluated at various temporal resolutions (daily, monthly, and quarterly) to assess their performances comprehensively.

3.3.1. The Daily Models

Figure 4 shows the daily bias and *RMSE* of the four ML-based T_m models. As shown in Figure 4, the average *RMSE* of the LightGBM-based T_m model is 1.85 K, which is 0.07 K, 0.04 K, and 0.13 K lower than the SVM, CART, and RF-based T_m models, respectively. Moreover, the maximum *RMSE* of the LightGBM-based T_m model is lower than the SVM, CART, and RF-based T_m models by 0.15 K, 0.13 K, and 0.15 K. The LightGBM-based T_m model shares the same minimum *RMSE* with the SVM and CART-based T_m models but outperforms the RF model. As for the bias, the LightGBM, CART, and RF-based models demonstrate relatively small variations, while the SVM-based model exhibits more significant fluctuations. Clearly, the LightGBM, CART, and RF-based models are more stable and suitable for obtaining T_m in the Yangtze River Delta region.

The left side of Figure 5 illustrates that the predicted ML-based T_m does not exactly match the true T_m values. However, they exhibit a similar overall trend, with the predicted T_m from the LightGBM-based T_m model being closer to the true value. On the right side of Figure 5, it is evident that the LightGBM-based T_m model demonstrates a smaller overall deviation compared to other ML-based T_m models and is more proximate to zero.

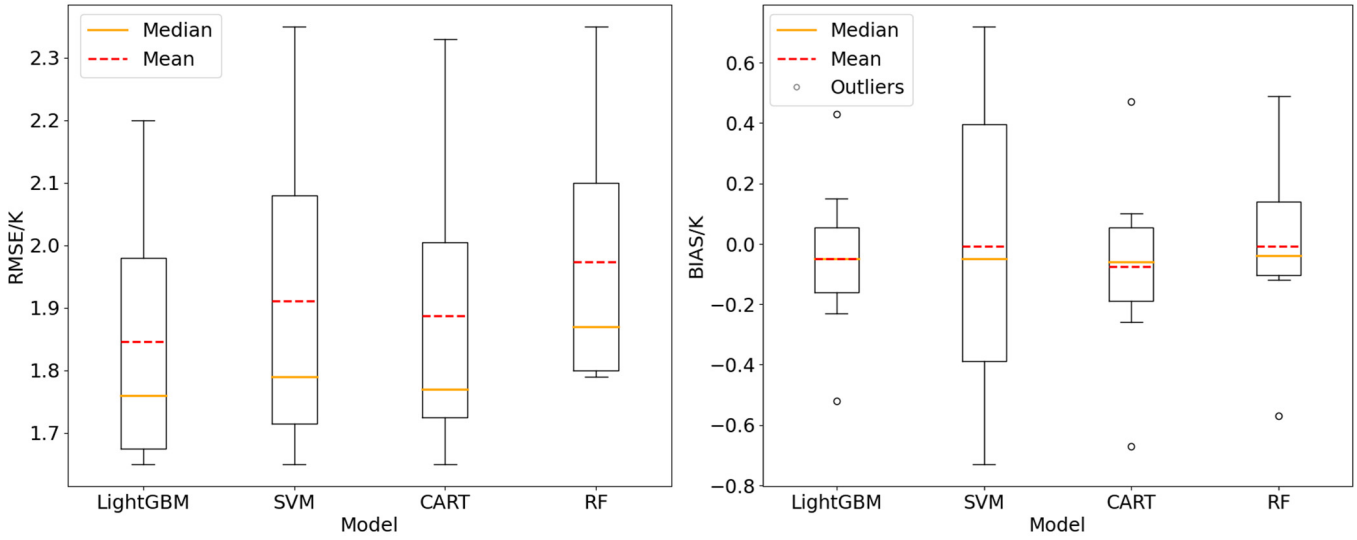

Figure 4. Accuracy analysis of daily ML-based T_m models.

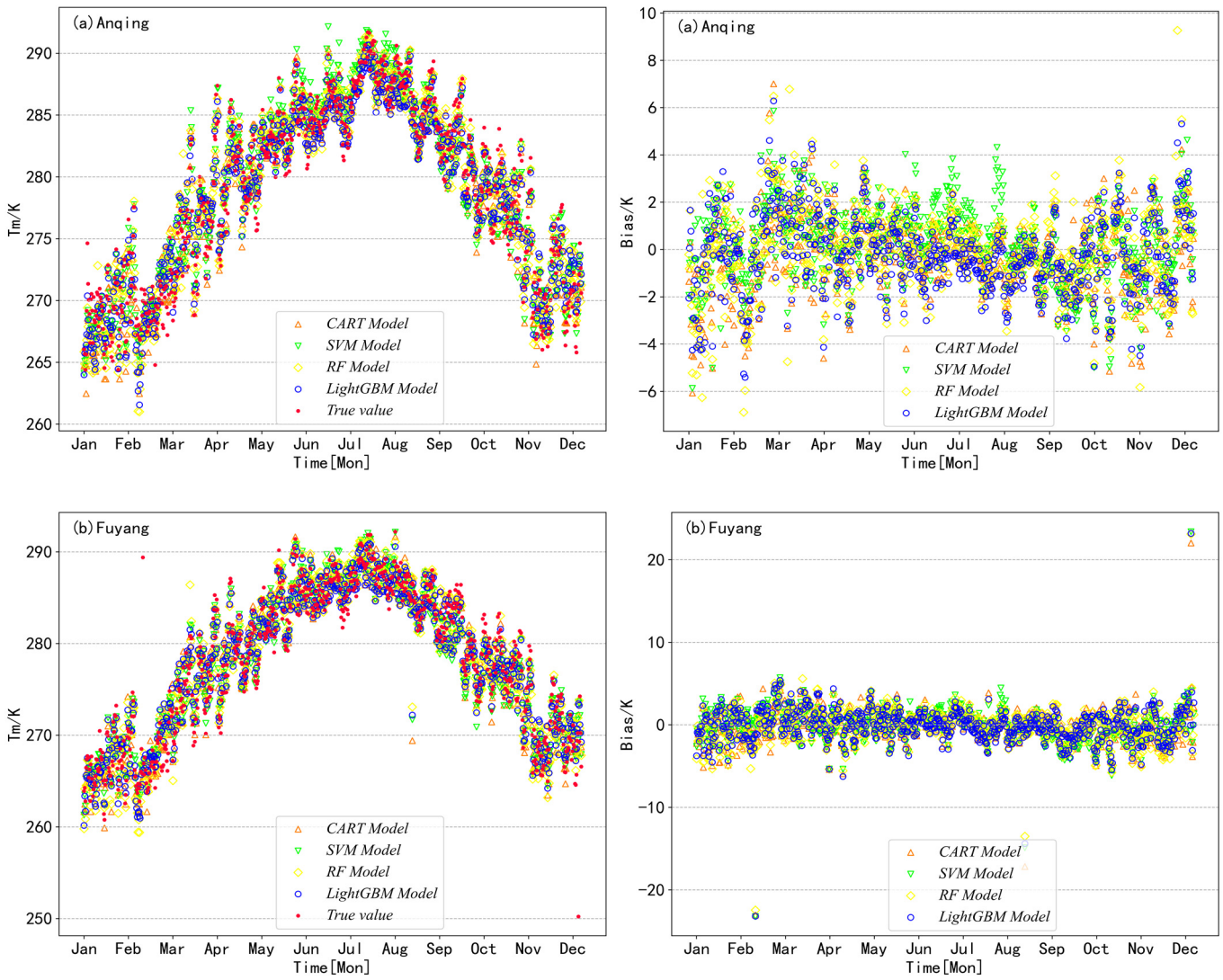

Figure 5. *Cont.*

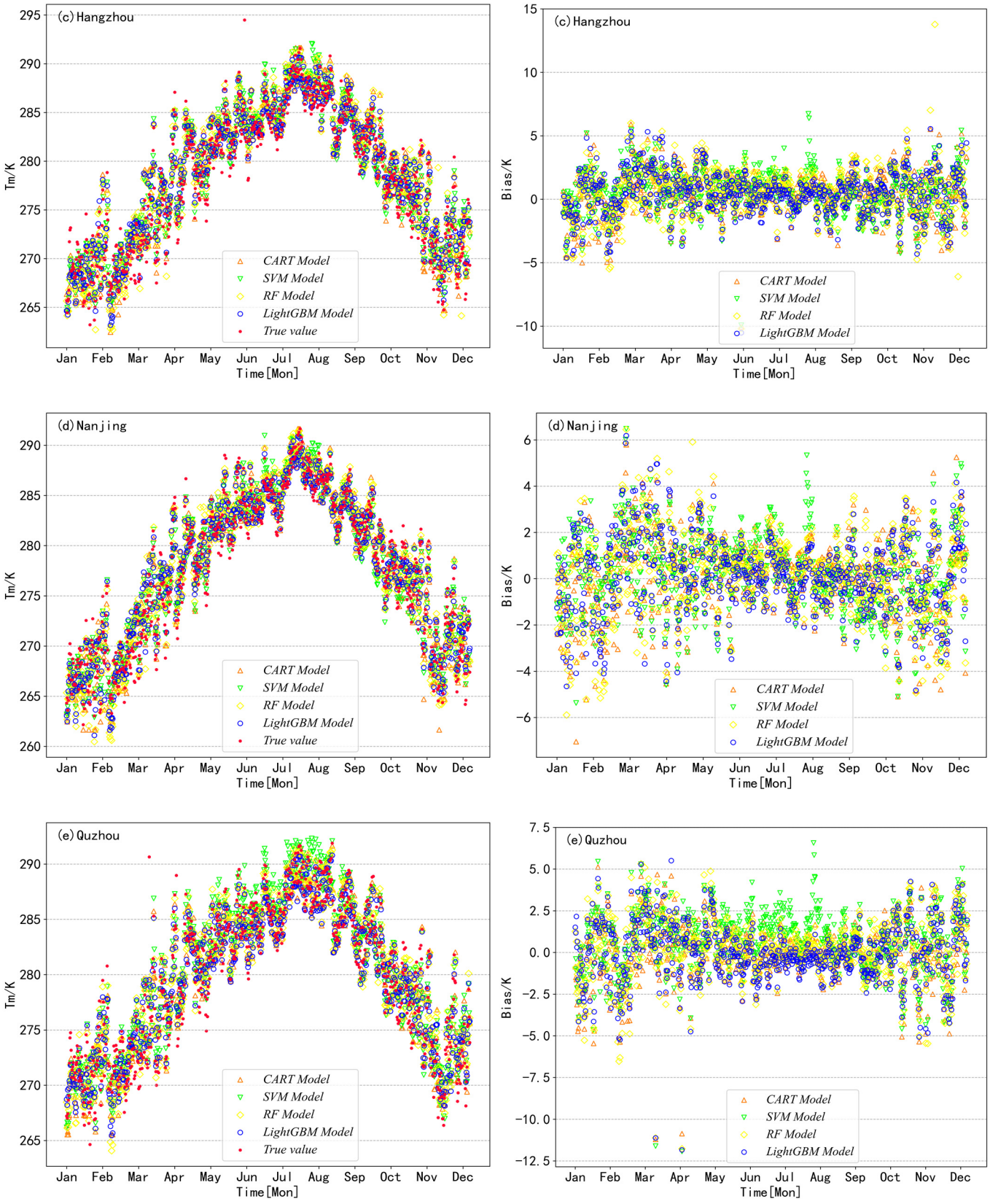

Figure 5. *Cont.*

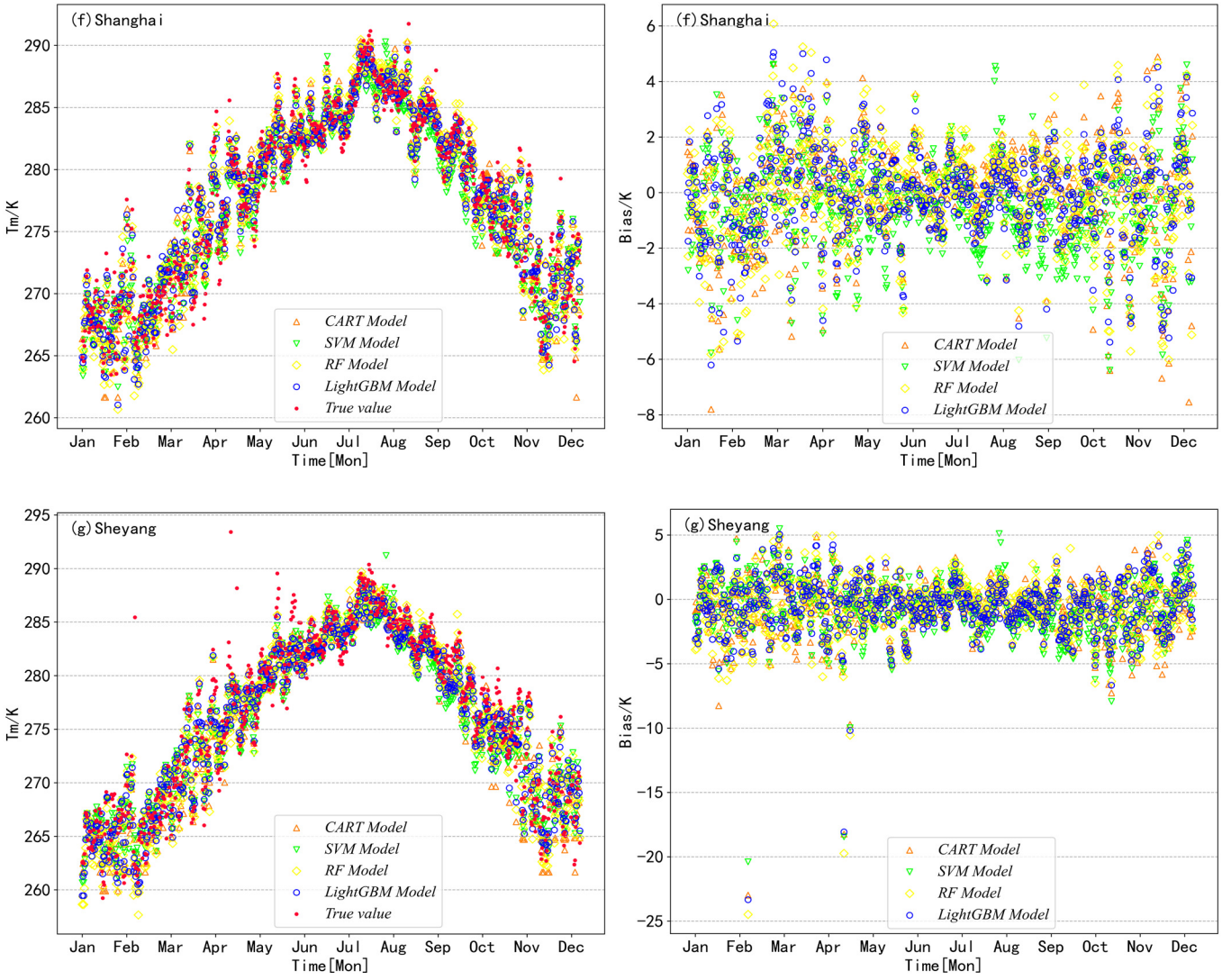

Figure 5. Comparison of T_m and day-by-day deviation of different ML-based T_m models, the figures in the left column show the changing trends of T_m values among the models, whereas the figures in the right column show the biases among the models.

3.3.2. Quarterly and Monthly Models

To further investigate the impact of temporal changes on model accuracy. The accuracy of the monthly and quarterly ML-based T_m models was analyzed in this section. The radiosondes in 2019 were divided into 12 validation datasets to analyze the accuracy of monthly ML-based T_m models. Similarly, when the temporal resolution is quarterly, the 2019 radiosonde data were divided into four validation datasets to analyze the accuracy of quarterly ML-based T_m models.

Figures 6 and 7 show the average bias and *RMSE* for the quarterly and monthly ML-based T_m models. It shows that the *RMSE* of four quarterly ML-based T_m models fluctuates between 1.46 K and 2.58 K, while the bias varies between 0.05 K and 0.77 K. The *RMSE* of four monthly ML-based T_m models fluctuates between 1.27 K and 2.85 K, while their bias varies from 0.01 K to 1.29 K.

The result shows that the LightGBM-based T_m model is more accurate and stable than the other three ML-based T_m models in any spatiotemporal conditions. In summary, the LightGBM-based T_m model is more stable in providing accurate T_m and more suitable for GNSS-PWV research in the Yangtze River Delta region.

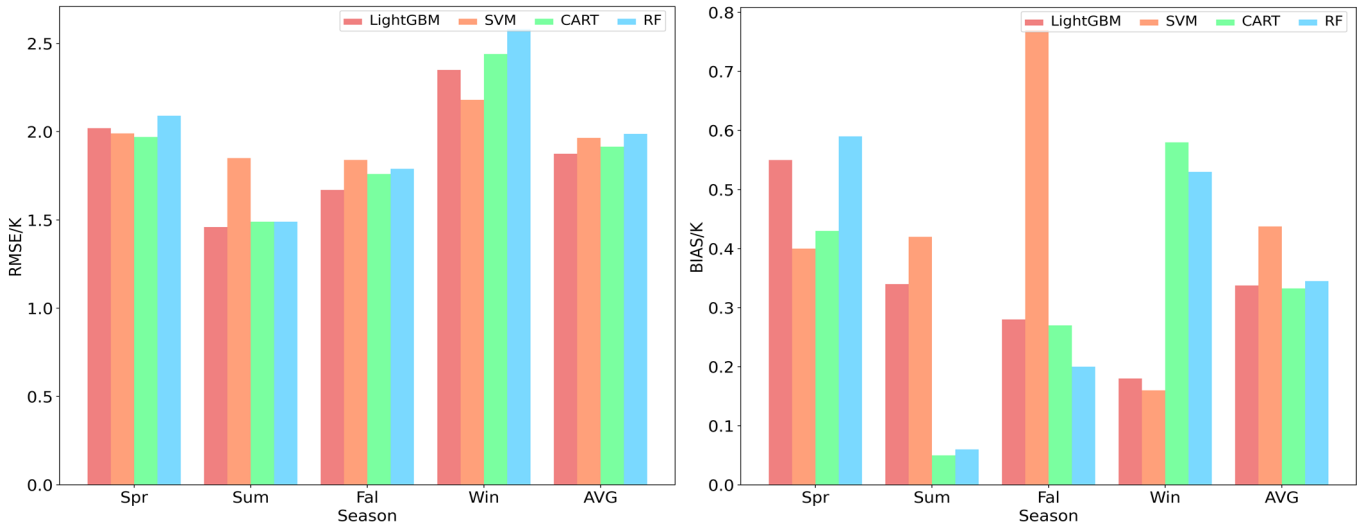

Figure 6. Accuracy analysis of quarterly ML-based T_m models.

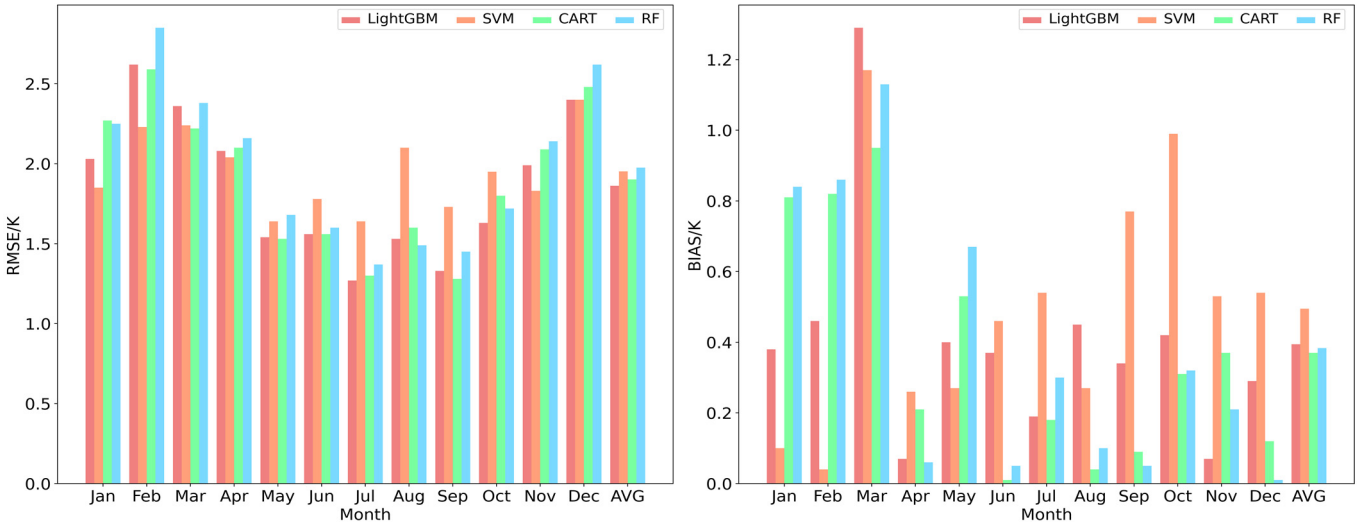

Figure 7. Accuracy analysis of the monthly ML-based T_m model.

4. Conclusions

The purpose of the paper was to solve the nonlinear fitting problem of traditional T_m models based on the four machine learning algorithms. It aimed to select the most suitable ML-based T_m model in the Yangtze River Delta region.

Due to the strong correlations between T_m and T_s, E_s, and P_s, which were selected as the features of ML-based models. The optimization procedure of ML-based models includes the cross-validation and optimal parameters identification, such as learning rate, maximum depth, and iteration counts, to ensure the reliability and accuracy of models.

The comparisons among the LightGBM, SVM, CART, and RF-based T_m models revealed that the *RMSE* of the daily LightGBM-based T_m model decreased by 0.07 K, 0.04 K, and 0.13 K than the other three ML-based T_m models. Similarly, the *RMSE* of the monthly LightGBM-based T_m model decreased 0.09 K, 0.04 K, and 0.11 K, while the *RMSE* of the quarterly LightGBM-based T_m model decreased 0.09 K, 0.04 K, and 0.11 K. It was evident that the LightGBM-based T_m model was more stable and superior to other ML-based T_m models in different temporal variations. In summary, the LightGBM-based T_m model can be used to get more high-precision T_m and GNSS-PWV in the Yangtze River Delta region.

This study cannot demonstrate the suitability of the ML-based T_m models for other regions because it only validates the applicability in the Yangtze River Delta region. In

order to further validate the applicability of ML-based T_m models, future research will involve constructing a comprehensive T_m model for the China region by algorithm fusion incorporating more ML-based algorithms.

Author Contributions: Conceptualization, K.L.; Data curation, Y.M. and M.Z.; Formal analysis, K.L., L.L. and Y.M.; Funding acquisition, L.L., and A.H.; Methodology, L.L. and A.H.; Project administration, L.L.; Resources, L.L., and J.P.; Software, K.L., J.P., Y.M. and M.Z.; Supervision, L.L.; Validation, K.L., J.P. and M.Z.; Writing—original draft, K.L. and J.P.; Writing—review & editing, L.L. and A.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported in part by the China Natural Science Funds under Grant 42204037, 41904033 and 42204014, the Graduate Practical Innovation Project of Jiangsu Province under Grant SJCX23_1718.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The authors would like to express their sincere gratitude to the University of Wyoming and the Jiangsu Institute of Meteorological Sciences for the provision of radiosondes and GNSS observations. We also thank the Reviewers for their constructive comments and suggestions, which resulted in a significant improvement in the quality of the paper. Lastly, I thank the National Natural Science Foundation of China (No. 42204037, 41904033 and 42204014) for their financial support for this research.

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

Abbreviations

The following abbreviations are used in this manuscript:

T_m	Weighted average temperature
T_s	Surface temperature
E_s	Water vapor pressure
P_s	Atmospheric pressure
ML	Machine learning
LightGBM	Light Gradient Boosting Machine
SVM	Support Vector Machine
RF	Random Forest
CART	Classification and Regression Tree
GNSS	Classification and Regression Tree
RMSE	Root mean square error
PWV	Precipitable water vapor
ZTD	Zenith total delay
ZHD	Zenith hydrostatic delay
ZWD	Zenith wet delay
GPT	Global Pressure and Temperature Model
GWMT	Global Weighted Mean Temperature
BPNN	Back Propagation Neural Networks
GRNN	Generalized Regression Neural Network
EFB	Exclusive feature bundling
GOSS	Gradient-based One-Side Sampling
VIF	Variance Inflation Factor

References

1. Askne, J.; Nordius, H. Estimation of tropospheric delay for microwaves from surface weather data. *Radio Sci.* **1987**, *22*, 379–386. [CrossRef]
2. Davis, J.; Herring, T.; Shapiro, I.; Rogers, A.; Elgered, G. Geodesy by radio interferometry: Effects of atmospheric modeling errors on estimates of baseline length. *Radio Sci.* **1985**, *20*, 1593–1607. [CrossRef]

3. Zhao, Q.; Liu, Y.; Yao, W.; Yao, Y. Hourly rainfall forecast model using supervised learning algorithm. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–9. [CrossRef]
4. Mircheva, B.; Tsekov, M.; Meyer, U.; Guerova, G. Anomalies of hydrological cycle components during the 2007 heat wave in Bulgaria. *J. Atmos. Sol.-Terr. Phys.* **2017**, *165*, 1–9. [CrossRef]
5. Lan, Z.; Zhang, B.; Geng, Y. Establishment and analysis of global gridded Tm−Ts relationship model. *Geod. Geodyn.* **2016**, *7*, 101–107. [CrossRef]
6. Bevis, M.; Businger, S.; Chiswell, S.; Herring, T.A.; Anthes, R.A.; Rocken, C.; Ware, R.H. GPS meteorology: Mapping zenith wet delays onto precipitable water. *J. Appl. Meteorol.* **1994**, *33*, 379–386. [CrossRef]
7. Yao, Y.; Zhu, S.; Yue, S. A globally applicable, season-specific model for estimating the weighted mean temperature of the atmosphere. *J. Geod.* **2012**, *86*, 1125–1135. [CrossRef]
8. Guo, B.; Li, L.; Xie, W.; Zhou, J.; Li, Y.; Gu, J.; Zhang, Z. Localized model fitting of weighted average temperature in the Yangtze River Delta. *J. Navig. Position* **2019**, *7*, 61–67. [CrossRef]
9. Ma, Y.; Chen, P.; Liu, T.; Xu, G.; Lu, Z. Development and Assessment of an ALLSSA-Based Atmospheric Weighted Mean Temperature Model with High Time Resolution for GNSS Precipitable Water Retrieval. *Earth Space Sci.* **2022**, *9*, e2021EA002089. [CrossRef]
10. Sun, Z.; Zhang, B.; Yao, Y. Improving the estimation of weighted mean temperature in China using machine learning methods. *Remote Sens.* **2021**, *13*, 1016. [CrossRef]
11. Huang, L.; Jiang, W.; Liu, L.; Chen, H.; Ye, S. A new global grid model for the determination of atmospheric weighted mean temperature in GPS precipitable water vapor. *J. Geod.* **2019**, *93*, 159–176. [CrossRef]
12. Umakanth, N.; Satyanarayana, G.C.; Simon, B.; Rao, M.; Babu, N.R. Long-term analysis of thunderstorm-related parameters over Visakhapatnam and Machilipatnam, India. *Acta Geophys.* **2020**, *68*, 921–932. [CrossRef]
13. Tran, T.T.K.; Lee, T.; Kim, J.-S. Increasing neurons or deepening layers in forecasting maximum temperature time series? *Atmosphere* **2020**, *11*, 1072. [CrossRef]
14. Ding, W.; Qie, X. Prediction of Air Pollutant Concentrations via RANDOM Forest Regressor Coupled with Uncertainty Analysis—A Case Study in Ningxia. *Atmosphere* **2022**, *13*, 960. [CrossRef]
15. Ding, M. A neural network model for predicting weighted mean temperature. *J. Geod.* **2018**, *92*, 1187–1198. [CrossRef]
16. Cai, M.; Li, J.; Liu, L.; Huang, L.; Zhou, L.; Huang, L.; He, H. Weighted Mean Temperature Hybrid Models in China Based on Artificial Neural Network Methods. *Remote Sens.* **2022**, *14*, 3762. [CrossRef]
17. Ju, Y.; Sun, G.; Chen, Q.; Zhang, M.; Zhu, H.; Rehman, M.U. A model combining convolutional neural network and LightGBM algorithm for ultra-short-term wind power forecasting. *IEEE Access* **2019**, *7*, 28309–28318. [CrossRef]
18. Saber, M.; Boulmaiz, T.; Guermoui, M.; Abdrabo, K.I.; Kantoush, S.A.; Sumi, T.; Boutaghane, H.; Nohara, D.; Mabrouk, E. Examining LightGBM and CatBoost models for wadi flash flood susceptibility prediction. *Geocarto Int.* **2022**, *37*, 7462–7487. [CrossRef]
19. Morshed-Bozorgdel, A.; Kadkhodazadeh, M.; Valikhan Anaraki, M.; Farzin, S. A novel framework based on the stacking ensemble machine learning (SEML) method: Application in wind speed modeling. *Atmosphere* **2022**, *13*, 758. [CrossRef]
20. Xu, C.; Lin, M.; Fang, Q.; Chen, J.; Yue, Q.; Xia, J. Air temperature estimation over winter wheat fields by integrating machine learning and remote sensing techniques. *Int. J. Appl. Earth Obs. Geoinf.* **2023**, *122*, 103416. [CrossRef]
21. Radhika, Y.; Shashi, M. Atmospheric temperature prediction using support vector machines. *Int. J. Comput. Theory Eng.* **2009**, *1*, 55. [CrossRef]
22. Lathifah, S.N.; Nhita, F.; Aditsania, A.; Saepudin, D. Rainfall Forecasting using the Classification and Regression Tree (CART) Algorithm and Adaptive Synthetic Sampling (Study Case: Bandung Regency). In Proceedings of the 2019 7th International Conference on Information and Communication Technology (ICoICT), Kuala Lumpur, Malaysia, 24–26 July 2019; pp. 1–5.
23. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.-Y. Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems 30, Proceedings of the NIPS 2017, Long Beach, CA, USA, 4–9 December 2017*; Curran Associates: Montreal, QC, Canada, 2017.
24. Liu, X.; Duan, H.; Huang, W.; Guo, R.; Duan, B. Classified early warning and forecast of severe convective weather based on LightGBM algorithm. *Atmos. Clim. Sci.* **2021**, *11*, 284–301. [CrossRef]
25. Tang, R.; Ning, Y.; Li, C.; Feng, W.; Chen, Y.; Xie, X. Numerical forecast correction of temperature and wind using a single-station single-time spatial LightGBM method. *Sensors* **2022**, *22*, 193. [CrossRef]
26. Xu, T.; Yu, Y.; Yan, J.; Xu, H. Long-Term Rainfall Forecast Model Based on The TabNet and LightGbm Algorithm. 2020. Available online: https://web.archive.org/web/20201126204621id_/https://assets.researchsquare.com/files/rs-107107/v1_stamped.pdf (accessed on 25 May 2023).
27. Singh, N.; Chaturvedi, S.; Akhter, S. Weather forecasting using machine learning algorithm. In Proceedings of the 2019 International Conference on Signal Processing and Communication (ICSC), Dalian, China, 20–23 September 2019; pp. 171–174.
28. Wang, A.; Xu, L.; Li, Y.; Xing, J.; Chen, X.; Liu, K.; Liang, Y.; Zhou, Z. Random-forest based adjusting method for wind forecast of WRF model. *Comput. Geosci.* **2021**, *155*, 104842. [CrossRef]
29. Jiang, N.; Fu, F.; Zuo, H.; Zheng, X.; Zheng, Q. A Municipal PM2.5 Forecasting Method Based on Random Forest and WRF Model. *Eng. Lett.* **2020**, *28*, 312–321.

30. Zhang, J.; Qiu, X.; Li, X.; Huang, Z.; Wu, M.; Dong, Y. Support vector machine weather prediction technology based on the improved quantum optimization algorithm. *Comput. Intell. Neurosci.* **2021**, *2021*, 6653659. [CrossRef] [PubMed]
31. Nayak, M.A.; Ghosh, S. Prediction of extreme rainfall event using weather pattern recognition and support vector machine classifier. *Theor. Appl. Climatol.* **2013**, *114*, 583–603. [CrossRef]
32. Kumar, R. Decision tree for the weather forecasting. *Int. J. Comput. Appl.* **2013**, *76*, 31–34. [CrossRef]
33. Geetha, A.; Nasira, G. Data mining for meteorological applications: Decision trees for modeling rainfall prediction. In Proceedings of the 2014 IEEE International Conference on Computational Intelligence and Computing Research, Coimbatore, India, 18–20 December 2014; pp. 1–4.
34. Gupta, D.; Ghose, U. A comparative study of classification algorithms for forecasting rainfall. In Proceedings of the 2015 4th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO) (Trends and Future Directions), Noida, India, 2–4 September 2015; pp. 1–6.
35. Li, J.; Zhang, B.; Yao, Y.; Liu, L.; Sun, Z.; Yan, X. A refined regional model for estimating pressure, temperature, and water vapor pressure for geodetic applications in China. *Remote Sens.* **2020**, *12*, 1713. [CrossRef]
36. Yao, Y.; Zhang, B.; Xu, C.; Yan, F. Improved one/multi-parameter models that consider seasonal and geographic variations for estimating weighted mean temperature in ground-based GPS meteorology. *J. Geod.* **2014**, *88*, 273–282. [CrossRef]
37. Li, L.; Wu, S.; Wang, X.; Tian, Y.; He, C.; Zhang, K. Seasonal multifactor modelling of weighted-mean temperature for ground-based GNSS meteorology in Hunan, China. *Adv. Meteorol.* **2017**, *2017*, 3782687. [CrossRef]
38. Isioye, O.A.; Combrinck, L.; Botai, J. Modelling weighted mean temperature in the West African region: Implications for GNSS meteorology. *Meteorol. Appl.* **2016**, *23*, 614–632. [CrossRef]
39. Miles, J. Tolerance and Variance Inflation Factor. Wiley Statsref: Statistics Reference Online. 2014. Available online: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118445112.stat06593 (accessed on 25 May 2023).
40. García, C.; García, J.; López Martín, M.; Salmerón, R. Collinearity: Revisiting the variance inflation factor in ridge regression. *J. Appl. Stat.* **2015**, *42*, 648–661. [CrossRef]
41. Yu, Z.; Qu, Y.; Wang, Y.; Ma, J.; Cao, Y. Application of machine-learning-based fusion model in visibility forecast: A case study of Shanghai, China. *Remote Sens.* **2021**, *13*, 2096. [CrossRef]
42. Yong, Z.; Youwen, L.; Shixiong, X. An improved KNN text classification algorithm based on clustering. *J. Comput.* **2009**, *4*, 230–237.