Downscaling the Midsummer Temperature-Humidity Index Based on Multiple Machine Learning Methods

Danwa Wu, Zhenhai Yao, Linlin Wu, Xichang Luo, Shuai Sun, Binfang He, and Yali Zhang

***Abstract*—To improve the finesse of the temperature-humidity index (THI), this study applies four machine learning methods in THI downscaling, including multiple linear regression, random forest (RF), support vector machine, and gradient boosting machine. The temperature data and specific humidity data of the China Meteorological Administration Land Data Assimilation System (CLDAS) are used to establish a downscaling model, and site observational data are used to test the model precision. By taking land surface temperature (LST), vegetation coverage, altitude, and slope as downscaling factors, the monthly average THI calculated by CLDAS-V2.0 data is downscaled from 6 to 1 km in Anhui Province in July and August from 2002 to 2021. The results show that the spatial resolution of THI is improved effectively by the four downscaling models, and there is a significant correlation between the downscaled values and the site values, with a correlation coefficient greater than 0.97. The downscaling effect of RF is slightly better than that of the other three algorithms and better describes the distribution of summer resort resources. Simulated results from RF are piecewise corrected by using the mean variation, and the correlation between corrected values and observations in July and August are both improved (>0.98). According to the estimation of the corrected THI (1 km × 1 km), the proportion of summer resort area in Anhui Province is 9.58% in July and 19.29% in August.**

***Index Terms*—China meteorological administration land data assimilation system (CLDAS), downscale, machine learning, moderate-resolution imaging spectroradiometer (MODIS), temperature-humidity index (THI).**

I. Introduction

THE temperature-humidity index (THI) is an important index for representing climatic suitability. It reflects the heat exchange between the human body and the surrounding environment through a combination of temperature and humidity. It is widely used in human settlement evaluation, climate resource utilization, agricultural production, and so on [1], [2], [3], [4]. The THI was proposed by Thom [5] and reflects the influence of hot and humid climate environments on human comfort in summer through the combination of dry bulb temperature and wet bulb temperature. Other influential climate comfort models include the apparent temperature (AT) model [6], [7], [8] and universal thermal climate index (UTCI) [9], [10]. AT introduces related theories of physiology and clothing material science and mainly studies the regulating effect of different humidity levels on human thermal sensation. UTCI is a mechanism model based on human thermal balance theory combined with a multinode thermo physiological model and clothing model. The research and application of climatic suitability models have developed rapidly in China since the 1990s. Large cities have established local climatic suitability models and carried out related services, such as Beijing, Shanghai, Tianjin, and Chongqing [11], [12]. The national standard "climatic suitability evaluation on human settlements" was published in China in 2011, and it pointed out that climatic suitability should be evaluated by THI in summer time [13].

Because of the spatial unevenness and discontinuity, observational data from meteorological stations to calculate THI are greatly limited. To obtain continuous THI data in larger areas, many scholars have started THI simulation research. At present, many methods, including remote sensing inversion, spatial interpolation, and statistical analysis, are used. Due to its wide observation range and good repeatability of Earth observation, the moderate-resolution imaging spectroradiometer (MODIS) data is often used to simulate THI in large area [14]. The simulating process generally includes three steps: first, invert the land surface temperature (LST) by thermal infrared spectrum channels and the whole integrated water vapor content with near-infrared channels [15], [16], [17], [18], [19], [20], [21]; second, calculate the air temperature with LST and the near-ground humidity parameters with whole integrated water vapor [22], [23], [24], [25], [26], [27], [28], [29], [30], [31]; finally, count THI with air temperature and near-ground humidity parameters by Magnus empirical formula. Xie et al. [14] proposed the THI model based on MODIS data and used the model to calculate the monthly THI in China in 2003, found that the method of THI inversion by remote sensing is feasible. Li et al. [32] replaced air temperature with LST and relative humidity with the normalized water vapor index to improve THI. Furthermore, they found that the value is higher than the traditional THI, but the overall distribution is not affected. By using station observations and MODIS data, Shi et al. [33] simulated the spatial distribution of THI in Zhejiang Province based on the GridMet model and

Manuscript received 11 May 2023; revised 2 July 2023; accepted 22 July 2023. Date of publication 27 July 2023; date of current version 7 August 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 41575155, and in part by the Innovative Development Project of Anhui Meteorological Bureau, China under Grant CXM202104. *(Corresponding author: Linlin Wu.)*

Danwa Wu, Zhenhai Yao, Xichang Luo, and Yali Zhang are with the Anhui Public Meteorological Service Center, Hefei 230031, China (e-mail: wudanwa@126.com; 1719690761@qq.com; 591047875@qq.com; zhangyaliit@163.com).

Linlin Wu is with the CMA Cloud-Precipitation Physics and Weather Modification Key Laboratory, Beijing 100081, China (e-mail: wulinlin@foxmail.com).

Shuai Sun is with the National Meteorological Information Center of China, Beijing 100081, China (e-mail: sunshuai@cma.gov.cn).

Binfang He is with the Anhui Institute of Meteorological Sciences, Hefei 230031, China (e-mail: he_binfang@sina.com).

Digital Object Identifier 10.1109/JSTARS.2023.3299459

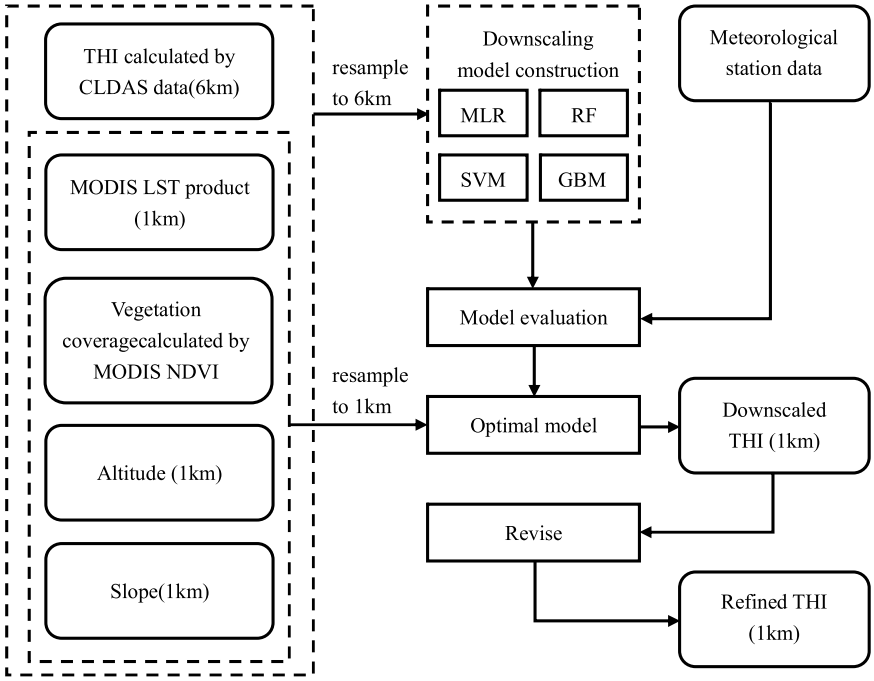

Fig. 1. Production procedure for fine-scale THI dataset.

found that the THI variations are closely related to terrain. Many methods mainly rely on satellite remote sensing data and geographic information data [34], [35]. However, satellite data quality is limited by its temporal resolution and cloud layers. Data processing is also complicated. The universality of spatial interpolation and statistical analysis needs to be considered in different places; therefore, their application in a large-scale area is limited.

In recent years, the China Meteorological Administration Land Data Assimilation System (CLDAS), providing high-quality grid point data such as temperature, pressure, humidity, and wind speed, has been operationally applied. The THI with a spatial grid distance of 6 km can be calculated by the temperature and humidity from CLDAS. However, its spatial resolution still has difficulty meeting the needs of refined summer tourism resource evaluation and tourism services. Therefore, this study analyzes the downscaling method of THI by using MODIS data and geographic data based on multiple machine learning methods. Taking Anhui Province as an example, the grid distance is reduced from 6 to 1 km. The accuracy of the downscaling results is evaluated by meteorological station data, the optimal model is selected, and the results are revised. Thus, the THI dataset with both refinement and accuracy can be obtained. The production procedure for fine-scale THI dataset is shown in Fig. 1.

II. Study Area and Data

A. Overview of the Study Area

Anhui Province is located in Middle China (29°41'-34°38'N,114°54'-119°37'E), with a length of approximately 570 km from north to south and a width of 450 km from east to west. The total area of the province is 140100 km^2. The terrain of Anhui Province slopes from southwest to northeast, with diverse landforms. The Yangtze River and the Huaihe River traverse the whole province, running through Anhui Province for 416 km and 430 km, respectively, which divides the province into three parts: the Huaibei Plain, Jianghuai Hills, and Wannan Mountains. There are many lakes in the Yangtze River basin, the largest of which is Chaohu Lake, which covers an area of 800 km^2. The main mountains are the Dabie Mountains, Huangshan Mountains, Jiuhua Mountains, and Tianzhu Mountains. The highest peak is the Huangshan Lianhua Peak (1864 m, above sea level).

B. Research Data

1) CLDAS Data: CLDAS is the only real-time operating system in the field of land data assimilation in China. It uses fusion and assimilation technology to integrate data from ground-based observations, satellite data, numerical model products, and other sources and outputs real land products with high spatial and temporal resolutions, including air temperature, barometric pressure, specific humidity, wind speed, precipitation, solar shortwave radiation, soil moisture, and so on [36], [37]. Its spatial resolution is 0.0625° × 0.0625°, and its temporal resolution is 1 h. In this study, the daily average 2 m air temperature and 2 m specific humidity data from CLDAS in July and August from 2002 to 2021 were collected, and the daily THI was calculated according to (1). Then, the average monthly and annual THI of July and August are calculated.

$$I_c = t_c - 0.55 \times (t_c - 1.44) \times \left(1 - \frac{\frac{q_c \times p}{0.378 q_c + 0.622}}{6.112 \exp\left(\frac{17.67 t_c}{t_c + 243.5}\right)}\right) \quad (1)$$

where I_c is the CLDAS THI index, t_c is the 2 m air temperature, q_c is the 2 m specific humidity, and p is the standard atmospheric pressure and is a constant. $p = 1013.25$ hpa.

According to the national standard "Climatic suitability evaluation on human settlements (GB/T 27963-2011)," when the THI value ranges in [17.0, 25.4], the human body feels comfortable, so this section is regarded as an appropriate area for summer tourism.

2) Site Observation Data: The daily THI is calculated by the observational data of 81 stations in Anhui Province according to (2), and then the average monthly and annual THI of July and August are calculated for the accuracy test of the downscaling method.

$$I_z = t_z \ - 0.55 \times (t_z - 14.4) \times (1 - Rh_z) \tag{2}$$

where I_z is the THI of the national station, t_z is the average temperature, and $\mathrm{Rh_z}$ is the relative humidity.

3) Geographic Information Data: The geographic information data include DEM (1 km × 1 km) raster data (unit: m) and administrative boundary vector data of Anhui Province, which are obtained from the local geospatial database of the Anhui Institute of Meteorological Sciences. Based on the ArcGIS platform, slope data were obtained using DEM data, and the spatial resolution was consistent with the DEM raster data (unit: °).

4) Satellite Remote Sensing Data: The EOS/MODIS satellite remote sensing data are used in this study, obtained from https://ladsweb.modaps.eosdis.nasa.gov/, including MOD11A2 surface temperature products (1 km ×1 km) and MOD13Q1 normalized difference vegetation index (1 km×1 km) products. Based on the maximum synthesis method, 8 d data of MOD11A2 and 16 d data of MOD13Q1 were used to obtain the monthly LST and normalized difference vegetation index (NDVI), and the NDVI was converted into vegetation coverage based on binary pixel model [38]. The data period was from 2002 to 2021 in July and August. To maintain unity with the spatial resolution of CLDAS data and facilitate spatial analysis, altitude, slope, LST, and vegetation coverage were resampled to the 0.0625° × 0.0625° grid.

III. Research Methods

A. Selection of Downscaling Factors

The spatial distribution of the summer THI in Anhui is closely related to the landform and surface environment. In the western Dabie Mountains and southern Wannan Mountains, the altitude is high, and the temperature decreases vertically with altitude. At the same time, the forest coverage is high, which has a cooling effect on the near surface, so the THI in this area presents a low value range. In the Huaibei Plain area, the altitude is low, and the land-use type is mainly farmland and urban buildings. The surface absorbs and stores a large amount of solar radiation and the heat energy released by human activities. When it spreads to the air, the near-surface temperature of the city rises, thus showing a high THI value range [39], [40], [41]. There is a close correlation between LST and near-surface temperatures [42]. Therefore, it is widely used in the inversion of THI. Vegetation cover is an important reason for the spatial difference in evapotranspiration and influences the near-surface temperature and humidity through the evapotranspiration process [43]. Therefore, LST and vegetation cover were used as surface factors to downscale the CLDAS THI. As a topographic factor of the air temperature model, altitude is significantly negatively correlated with air temperature. Shi et al. [33] analyzed the distribution characteristics of THI in Zhejiang with terrain factors, and the results showed that THI changed little with aspect in midsummer. Anhui Province and Zhejiang Province both belong to the Yangtze Valley with similar climate characteristics and topography. It can be considered that the THI of Anhui Province changes little with aspect in midsummer and is mainly affected by topographic factors such as elevation and slope. In summary, LST, vegetation coverage, altitude, and slope were used as downscaling factors to simulate midsummer THI in Anhui.

B. MLR Algorithm

Multiple linear regression (MLR) models make predictions by adding a weighted sum of input features and a biased term. The model can be expressed as the following expression:

$$\hat{y} = \theta_0 \ + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n. \tag{3}$$

In this equation, $\widehat{y}$ is the predicted value, n is the eigenvector, x_i is the eigenvalue, and θ_i is the model parameter (including the bias term θ_0 and the eigenweight $\theta_1 \theta_2 \ldots \theta_n$).

C. RF Algorithm

The random forest (RF) algorithm was proposed by Breiman [44]. The RF algorithm is a type of ensemble learning model that takes a decision tree as a basic classifier. Bootstrap sampling is used in RF to extract multiple samples from the original sample. The decision tree models each sample. Thereafter, the predictions of multiple decision trees are combined, and the final prediction results are achieved by voting [45].

The formula is as follows:

$$Y \ = \frac{1}{S} \sum_{1}^{S} F_S \left(X \right) \tag{4}$$

where Y is the prediction result, X is the input characteristic data vector, S is the number of regression tree models, and $F_S(X)$ is a single CRAT regression tree model, and its formula is

$$F_X = \sum_{t-1}^{t} C_t I \left(X \in R_t \right) \tag{5}$$

where R_t is the unit domain divided by optimal segmentation variables with different characteristics, and I is the logical value, if $X \in R_t$, $I = 1$, otherwise $I = 0$; C_t is the average of all output values contained within the cell domain, and t is the cell domain label. The essence of this formula is to first determine which unit domain the input variable belongs to and then return the predicted value of that unit domain.

D. SVM Algorithm

The basic idea of support vector machine regression (SVR) is to map data into a high-dimensional feature space where linear regression processing is performed on the data [46], [47], [48]. Assume that the training sample set is $\{(x_i, y_i)|i = 1, 2 \cdots p\}$, where p is the number of training samples. The SVR regression function can be transformed into a dual function by introducing Lagrange multipliers, and the dual function and constraints are (6) and (7), respectively:

$$\min_{w} \frac{1}{2}\|w\|^2 + c\sum_{i=1}^{p}(\xi_i + \xi_i^*) \tag{6}$$

$$s.t. \begin{cases} y_i - w \cdot \phi(x_i) - b \le \varepsilon + \xi_i \\ w \cdot \phi(x_i) + b - y_i \le \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \ge 0 \end{cases} \tag{7}$$

where w is the weight;$c$is the penalty factor; ξ and ξ^* are relaxation factors;ϕ is the mapping function; ε is the insensitive loss function; and b is the threshold value. Equation (8) can be obtained by solving (6) and (7).

$$f(x) = \sum_{j=1}^{n_{SV}} (\alpha_j - \alpha_j^*) K(x_j, x) + b \tag{8}$$

where x is the sample value; α and α^* are Lagrange multipliers; n_{SV} is the number of support vectors; $K(x_j, x)$ is the kernel function, and RBF and polynomial involved in this article are (9) and (10), respectively:

$$K(x_j, x) = exp\left(-g\|x - x_j\|_2\right) \tag{9}$$

$$K(x_j, x) = \left[g\left(x_j^T x_j\right)\right]^d \tag{10}$$

where g is the kernel function parameter and d is the polynomial order.

E. GBM Algorithm

Gradient boosting is a machine learning technique for classification and regression problems. The main idea is to generate multiple weak learners serially. The goal of each weak learner is to fit the negative gradient of the loss function of the previously accumulated model so that the cumulative model loss after adding the weak learner is reduced in the direction of the negative gradient [49], [50]. XGBoost is an implementation of the boosting elevator. Newton's method is used to solve the extreme value of the loss function. The loss function is expanded to the second order by Taylor expansion, and the regularization term is added to the loss function. The objective function of training is composed of two parts: the first part is the loss of the gradient boosting algorithm, and the second part is the regularization term. The loss function is defined as follows:

$$\mathrm{L}(\phi) = \sum_{i=1}^{n} l(y'_i, y_i) + \sum_{k} \Omega(f_k) \tag{11}$$

where n is the number of training samples, l is the loss to a single sample, y'_i is the predicted value, y_i is the real label of the sample, and ϕ is the parameter of the model. The regularization term defines the complexity of the model.

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda\|w\|^2 \tag{12}$$

where γ and λ are the coefficients manually set, w is the vector formed by all leaf node values in the decision tree, and T is the number of leaf nodes. The regularization term is composed of the number of leaf nodes and the modular square of the leaf node value vector. The first term reflects the complexity of the decision tree structure, and the second term reflects the complexity of the predicted value of the decision tree.

TABLE I
AREA PROPORTION OF DIFFERENT THI INTERVAL IN JULY

Data	Area proportion (%)					
	[17.0,19.4]	[19.5,21.4]	[21.5,23.4]	[23.5,25.4]	[25.5,27.5]	[27.6, +∞)
CLDAS	0	0	0.51	7.34	91.76	0.39
MLR	0	0	0.26	6.33	93.26	0.15
RF	0	0	0.27	6.44	93.11	0.18
SVM	0	0	0.24	6.12	93.48	0.16
GBM	0	0	0.18	6.52	93.13	0.17

IV. RESULT ANALYSIS

A. Spatial Distribution of Four Downscaling Results

According to the average THI in July and August from 2002 to 2021, the downscaling effects of four methods, including MLR, RF, support vector machine (SVM), and gradient boosting machine (GBM), were compared. Fig. 2(a) and (b) show the 20-year average THI of July and August calculated from CLDAS data, with a spatial resolution of 0.0625° × 0.0625°. According to Fig. 2(a) and (b), the THI in July and August of midsummer in Anhui were both greater than 21.5 In July, the suitable summer areas where the THI value ranges in [17.0, 25.4] are mainly located in the Dabie Mountains and Wannan Mountains. In August, apart from the above areas, Dangshan and Bozhou in the northernmost regions also appear with low THI. Fig. 2(c) to (j) show the downscaling results of the four methods. The scaling results of the four methods and the CLDAS THI have the same numerical interval, and the spatial distribution has a similar rule. The spatial distribution of THI after downscaling presents more abundant details. The four downscaling methods all reflect the suitable summer areas of the Dabie Mountains and Wannan Mountains well, and the RF algorithm can well simulate the suitable summer areas in August located in northern Anhui.

Tables I and II show the area proportion of different intervals of average THI in July and August from 2002 to 2021. It can be seen that in July and August, in the interval of [17.0, 19.4] and [19.5, 21.4], the area proportion of CLDAS and the four simulation methods is 0. In the intervals of [21.5, 23.4] and [23.5, 25.4], the area ratios of the four simulation methods are all smaller than that of CLDAS. The sum of the area of the two intervals is the interval [21.525.4], and the area of July is as follows: CLDAS (7.85%) >RF (6.71%) >GBM(6.70%) >MLR(6.59%)

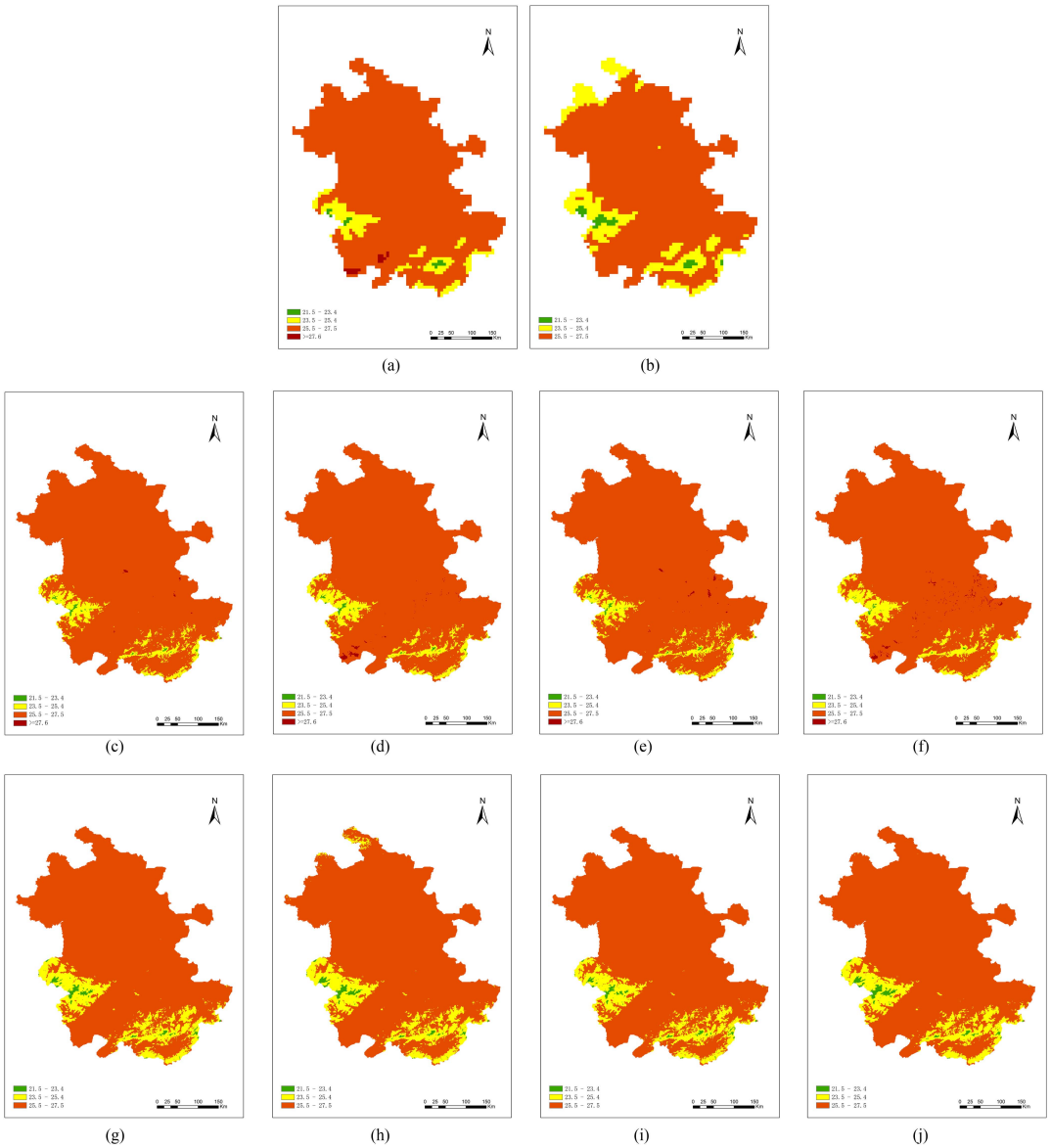

Fig. 2. Spatial distribution of average THI in July and August from 2002 to 2021 by CLDAS data and different downscaling method. (a) July, CLDAS. (b) August, CLDAS. (c) July, MLR. (d) July, RF. (e) July, SVM. (f) July, GBM. (g) August, MLR. (h) August, RF. (i) August, SVM. (j) August, GBM.

TABLE II
AREA PROPORTION OF DIFFERENT THI INTERVAL IN AUGUST

Data	Area proportion (%)					
	[17.0,19.4]	[19.5,21.4]	[21.5,23.4]	[23.5,25.4]	[25.5,27.5]	[27.6, +∞)
CLDAS	0	0	1.42	15.85	82.73	0
MLR	0	0	0.81	11.56	87.63	0
RF	0	0	0.87	12.02	87.11	0
SVM	0	0	0.67	10.86	88.47	0
GBM	0	0	0.96	11.28	87.76	0

>SVM(6.36%).The area proportion in August was as follows: CLDAS(17.27%) >RF(12.89%) >MLR(12.37%) > GBM(12.24%) >SVM(11.53%). In comparison, the simulation of the suitable summer area shows that the RF is closer to the CLDAS value.

B. *Precision Analysis of Four Downscaling Methods*

To check the accuracy of different machine learning downscaling methods, the 1 km × 1 km THI grid value was extracted according to the locations of 81 national stations. The correlation between the average monthly THI in July and August of the national station from 2002 to 2021 and the downscaled THI was analyzed. The correlation coefficient (R), bias (B), and root mean square error (RMSE) were used to quantitatively analyze the correlation between the THI of observation stations and the THI of grid points after downscaling. The results are shown in Table III.

There was a significant correlation between the four machine learning downscaling results and the THI of the site, R ⩾ 0.975,

TABLE III
CORRELATION BETWEEN DOWN-SCALING RESULT AND SITE THI

Methods	July			August		
	R	RMSE	B	R	RMSE	B
MLR	0.979	0.2383	0.131	0.976	0.2656	0.147
RF	0.979	0.2421	0.126	0.981	0.2390	0.098
SVM	0.975	0.2561	0.098	0.979	0.2426	0.078
GBM	0.976	0.2504	0.124	0.979	0.2429	0.094

TABLE IV
DEVIATION OF THI IN DIFFERENT INTERVALS IN JULY

Methods	Bias					
	[17.0,19.4]	[19.5,21.4]	[21.5,23.4]	[23.5,25.4]	[25.5,27.5]	[27.6,29.0]
MLR	2.238	1.231	0.886	0.315	0.087	-0.100
RF	2.373	1.399	0.833	0.220	0.081	-0.026
SVM	2.214	1.403	0.899	0.269	0.055	-0.139
GBM	2.649	1.505	0.947	0.224	0.077	-0.045

TABLE V
DEVIATION OF THI IN DIFFERENT INTERVALS IN AUGUST

Methods	Bias					
	[17.0,19.4]	[19.5,21.4]	[21.5,23.4]	[23.5,25.4]	[25.5,27.5]	[27.6,29.0]
MLR	1.705	1.128	0.765	0.256	0.092	-0.209
RF	2.078	0.988	0.538	0.162	0.048	-0.198
SVM	1.786	1.077	0.801	0.214	0.005	-0.261
GBM	2.370	0.962	0.605	0.165	0.037	-0.216

RMSE $\leqslant$ 0.2656, B $\leqslant$ 0.147 (all passed the 99% confidence test). The results indicate that downscaling factors such as LST, vegetation coverage, altitude, and slope are highly correlated with the THI, and it is appropriate to select these factors for downscaling. Fig. 3 shows the fitting between the downscaled THI (labeled I_d) of each algorithm and the THI of the site (labeled I_z). Compared with Table I, the difference between the simulated value and the observation value can be analyzed more intuitively.

In Fig. 3, the downscaling results of the four methods, I_d are generally higher than I_z in the interval [17.0, 25.4], and the lower the value of the THI is, the higher the value. This is also the reason why the simulated area of the suitable summer area is smaller than that of CLDAS. I_d are close to I_z in the interval [25.5, 29.0].

The average deviation is calculated according to different intervals, and the results are shown in Tables IV and V. Since the THI interval of [17.0, 25.4] is the suitable area for summer where we are concerned, the simulation accuracy of the THI of this interval largely determines the downscaling effect. In this interval, the accuracy of the same method in August is better than that in July. The reason may be that there are more samples in this interval in August, and the more samples there are, the better the effect of machine learning. Compared with the other algorithms, the RF algorithm in the interval [21.525.4] has a smaller deviation than the other algorithms. Previous analysis showed that the THI values in the summer resort area of Anhui are mainly concentrated in this interval, and the correlation coefficient R of the RF algorithm is the highest among the four algorithms, so it can be considered that the RF algorithm has the best downscaling effect.

C. Revision of RF Downscaling Results

As seen from the above, when comparing the four machine learning algorithms, the downscaling effect of RF is better, but there are still some systematic errors between the simulation value I_d and the site value I_z. According to the error analysis results in Tables IV and V, the average deviation B was used as the revisal value to make piecewise correction to the monthly average THI in July and August. The revised THI is denoted as I_d^c, $I_d^c = I_d - B$. according to the linear fitting relationship between the simulated value and the site value in Fig. 3(b) and (f), when $I_z \in [17.0, 19.4]$,[19.52, 1.4], [21.5, 23.4], [23.5, 25.4], [25.5, 27.5], [27.6, 29.0], in July $I_d \in [18.8, 20.8]$, [20.9, 22.4], [22.5, 24.0], [24.1, 25.7], [25.8, 27.4], [27.5, 28.6], the corresponding value of B is 2.373,1.399,0.833,0.220,0.081, -0.026; In August $I_d \in [18.6, 20.6]$, [20.7, 22.2], [22.3, 23.9], [24.0, 25.6], [25.7, 27.3], [27.4, 28.6], the corresponding value of B is 2.078, 0.988, 0.538, 0.162, 0.048, 0.198. Fig. 4 shows the comparative analysis between the revised results and the site values. After the revision, the correlation between the RF downscaling results and the site value is improved, and the correlation coefficient is 0.986 in July and 0.985 in August. The root-mean-square error was 0.2377 in July and 0.2478 in August. The deviation was 0.0028 in July and 0.0033 in August.

For the temporal trends, from Fig. 5, it can be figured out that the values and temporal trends for the site THI (red line) are similar to the revised THI (blue line), whether in July or August. In comparison, the THI fluctuation from 2012 to 2021 is larger than that from 2002 to 2011.

Fig. 6 shows the spatial distribution of THI drawn with the RF correction result. It can be seen that the area of summer resort where the THI is in [17.0, 25.4] has increased compared with that before the correction, both in July and August. especially in August, the summer resort area in the northernmost Xiaoxian County and Dangshan County has increased obviously. After the revision, the proportions of summer resort areas in July and August were 9.58% and 19.29%, respectively, which increased by 2.87% and 6.4% compared with before the revision and increased by 1.73% and 2.02% compared with CLDAS.

V. CONCLUSION AND DISCUSSION

Data from Anhui Province were analyzed in this study, and the average THI values in July and August from 2002 to 2021 were calculated by using the live grid data of CLDAS. Four machine learning algorithms, MLR, RF, SVM, and GBM, were adopted. Using LST, vegetation coverage, altitude, and slope as downscaling factors, the CLDAS THI was downscaled. The grid distance of monthly THI in July and August of Anhui Province from 2002 to 2021 was downscaled from 6 to 1 km, and the accuracy of the downscaling results was analyzed using site data

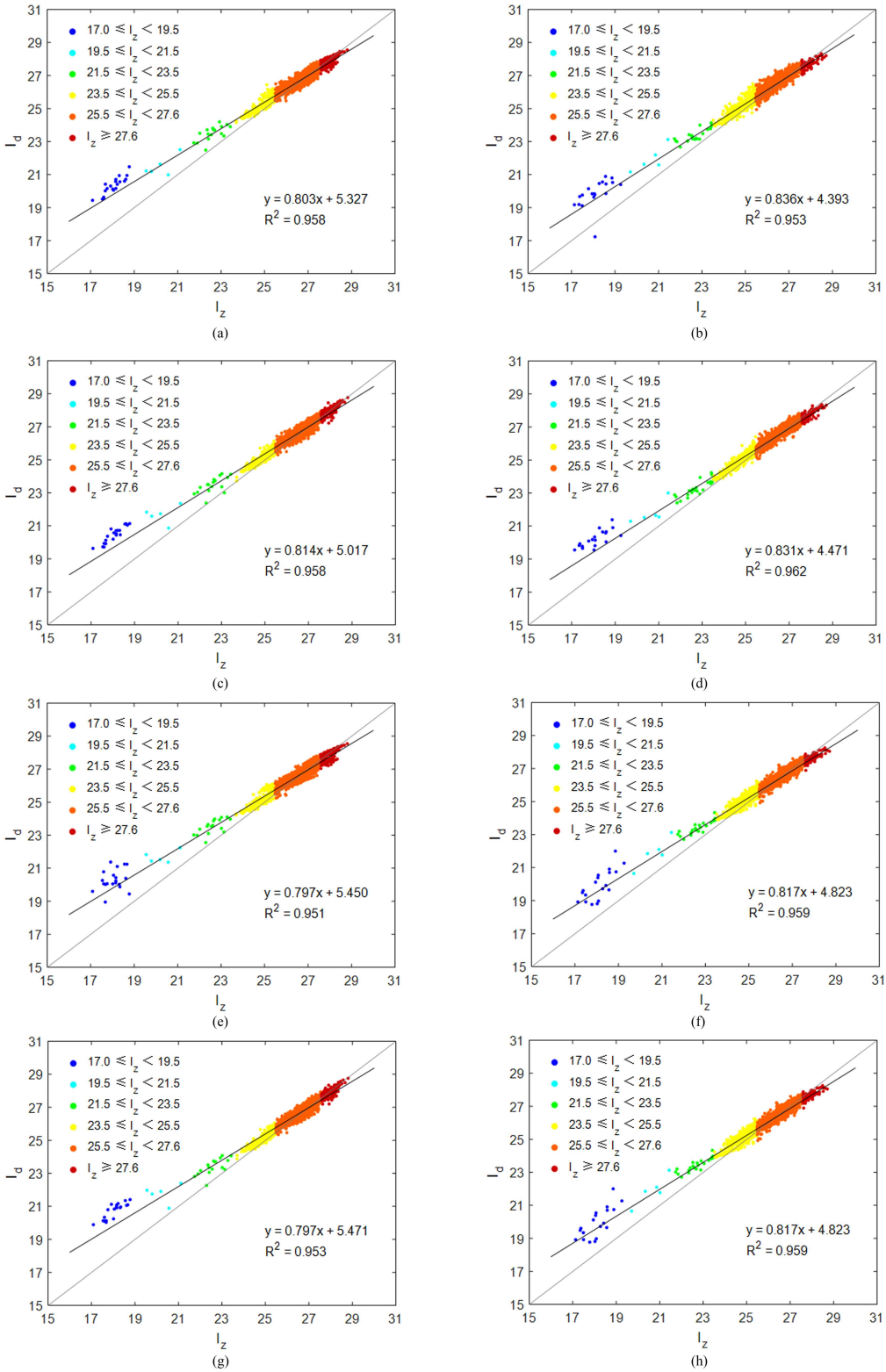

Fig. 3. Scatter plots of the site THI and down-scaling result by different downscaling method. (a) July, site, and MLR. (b) July, site, and RF. (c) July, site, and SVM. (d) July, site, and GBM. (e) August, site, and MLR. (f) August, site, and RF. (g) August, site, and SVM. (h) August, site, and GBM.

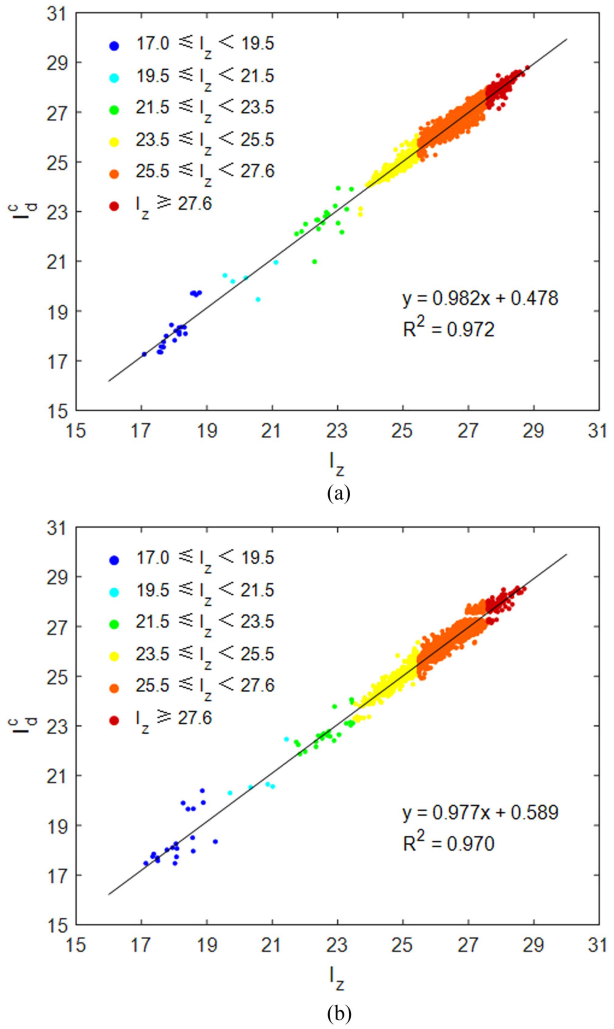

Fig. 4. Scatter plot of the site THI and the RF correction THI. I_z is the site THI; I_d^c is the RF correction THI. (a) July. (b) August.

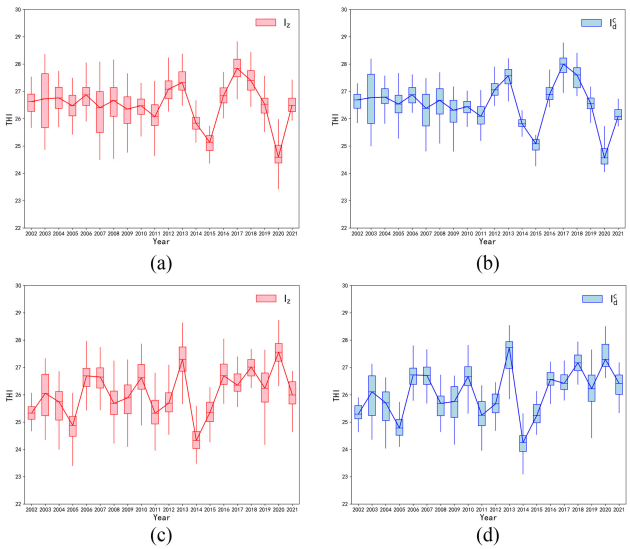

Fig. 5. Temporal trends of THI. I_z is the site THI; I_d^c is the RF correction THI. (a) July, I_z. (b) July, I_d^c. (c) August. I_z. (d) August I_d^c.

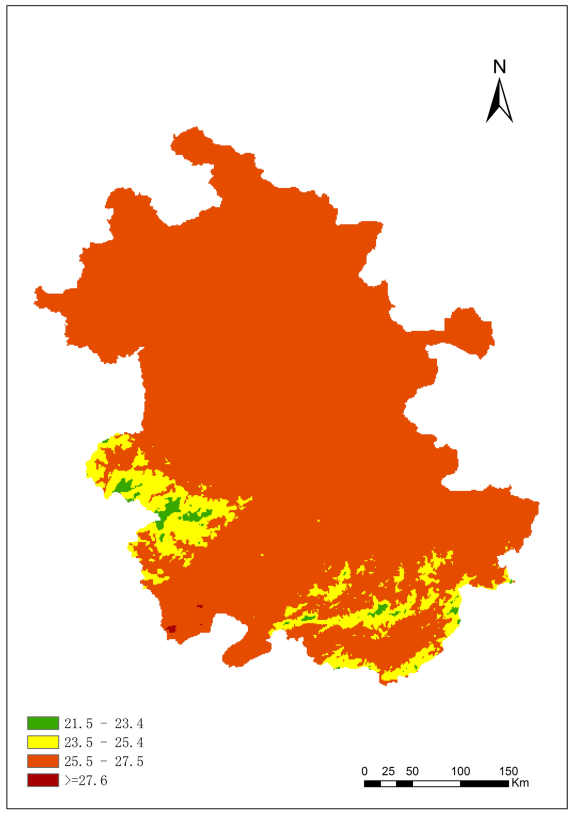

Fig. 6. Spatial distribution of average THI in July and August from 2002 to 2021 with RF method correction result. (a) July. (b) August.

to select the optimal model and propose the correction method. The results show the following:

First, the downscaling results of the four methods and the CLDAS THI had a consistent value range, and the 20-year average THI values in July and August were both above 21.5. The spatial distribution of [21.5, 25.4] is similar. The suitable areas for summer are mainly located in the Dabie Mountains and Wannan Mountains. After downscaling, the spatial distribution of THI showed richer details, which effectively improved the spatial resolution.

Second, there was a significant correlation between the downscaling estimates of the four methods and the site values ($R \geqslant 0.975$). The downscaling estimates were all higher than the site values in the interval [17.0, 25.4], and the lower the THI values were, the higher they were. In the interval [25.5, 29.0], the values are close to the site values. The area of summer resorts after downscaling is underestimated.

Third, among the four methods, the RF method has the best downscaling effect, the correlation coefficient between the simulated THI and site THI is the largest, the deviation is the smallest in the interval [21.5, 25.4], and the fitting effect of the summer resort is the best.

Fourth, taking the average deviation as the revisal value, the RF simulation results were revised by sections. The correlation coefficients in July and August increased to 0.986 and 0.985, the root-mean-square error was reduced to 0.2377 and 0.2478, and the deviation was reduced to 0.0028 and 0.0033, respectively. According to the correction result with a spatial grid distance of 1 km, the proportion of summer resort areas in Anhui Province was 9.58% in July and 19.29% in August. Compared with CLDAS, they increased by 1.73% and 2.02%, respectively.

Four machine learning methods were used to downscale the THI calculated by CLDAS live grid data, and the THI with a spatial grid distance of 1 km was obtained, providing refined data support for the assessment of summer resort resources. According to these refined data, an advisory report was provided to the relevant management department for finding summer tourist destination.

However, the study had a research period of 20 years, so the subsequent time series should be further extended with more data. In addition, the proposed method is too simple to correct the downscaling results with the average deviation as the revisal value, and various correction methods can be tried to improve the data accuracy in further work.

ACKNOWLEDGMENT

The authors would like to thank the National Aeronautics and Space Administration (NASA) for providing the EOS/MODIS data, and also the National Meteorological Information Center of China for providing the CLDAS data and observational data; the Anhui Institute of Meteorological Sciences for providing the Geographic information data.

REFERENCES

[1] Y. F. Wang and Y. Shen, "The temperature-humidity effect and human comfort in Shanghai summer," *J. East China Normal Univ. (Natural Sci.)*, vol. 3, pp. 60–66, 1998.

[2] R. Emmanuel, "Thermal comfort implication of urbanization in a warm-humidity: The Colombo Metro Politian Region (CMR), Sri Lanka," *Building Environ.*, vol. 40, pp. 1591–1601, 2005, doi: 10.1016/j.buildenv.2004.12.004.

[3] S. Wang, H. Tian, W. S. Xie, W. A. Tang, and X. Ding, "Change characteristics and regionalization of climate comfort level in Anhui Province in recent 50 years," *Prog. Geography*, vol. 31, no. 1, pp. 40–45, 2012.

[4] A. Q. Jin, A. Zhang, and X. Y. Zhao, "Prediction of future climate comfort level in Eastern China under climate change scenario," *Acta Sci. Natural, Univ. Peking*, vol. 55, no. 5, pp. 887–898, 2019, doi: 10.13209/j.0479-8023.2019.057.

[5] E. C. Thom, "The discomfort index," *Weathervise*, vol. 12, pp. 57–60, 1959.

[6] R. G. Steadman, "The assessment of sultriness. Part 1: A temperature-humidity index based on human physiology and clothing science," *J. Appl. Meteorol. Climatol.*, vol. 18, pp. 861–873, 1979.

[7] R. G. Steadman, "The assessment of sultriness. Part 2: Effect of wind extra radiation and barometric pressure on apparent temperature," *J. Appl. Meteorol.*, vol. 18, pp. 874–884, 1979.

[8] R. G. Steadman, "A universal scale of apparent temperature," *J. Appl. Meteorol. Climatol.*, vol. 23, pp. 1674–1167, 1984.

[9] K. Blazejczyk, Y. Epstein, G. Jendritzky, H. Staiger, and B. Tinz, "Comparison of UTCI to selected thermal indices," *Int. J. Biometeorology*, vol. 56, pp. 34–39, 2012, doi: 10.1007/s00484-011-0453-2.

[10] G. Jendritzky, R. Dear, and G. Havenith, "UTCI-why another thermal index?," *Int. J. Biometeorology*, vol. 56, pp. 421–428, 2012, doi: 10.1007/s00484-011-0513-7.

[11] M. Liu, B. Yu, and K. M. Yao, "Research status and application prospect of human comfort," *Meteorol. Sci. Technol.*, vol. 30, no. 1, pp. 11–14, 2002, doi: 10.19517/j.1671-6345.2002.01.003.

[12] Y. C. Yan, S. P. Yue, X. H. Liu, D. D. Wang, and H. Chen, "Advances in assessment of bioclimatic comfort conditions at home and abroad," *Adv. Earth Sci.*, vol. 28, no. 10, pp. 1119–1125, 2013.

[13] Climatic Suitability Evaluation on Human Settlements, Chinese National Standard 27963- 2011, Mar. 2012. [Online]. Available: http://www.cmastd.cn/standardView.jspx?id=473

[14] W. Xie, L. X. Ren, and L. P. Jiang, "A study on spatial and temporal distribution of temperature-humidity index in China based on MODIS data," *Geographic Inf. Sci.*, vol. 22, no. 5, pp. 31–35, 2006.

[15] Y. Shen, H. Shen, Q. Cheng, and L. Zhang, "Generating comparable and fine-scale time series of summer land surface temperature for thermal environment monitoring," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 2136–2147, Dec. 23 2020, doi: 10.1109/JSTARS.2020.3046755.

[16] Z. Wan, Y. Zhang, Q. Zhang, and Z. L. Li, "Quality assessment and primary productivity in the complicated mountainous area: A case study of Yunnan, China," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 12, pp. 4637–4648, Dec. 2018.

[17] Z. M. Wan, "New refinement and validation of the MODIS land-surface temperature/emissivity products," *Remote Sens. Environ.*, vol. 112, no. 1, pp. 59–74, Jan. 2008.

[18] S.-B. Duan and Z.-L. Li, "Inter comparison of operational land surface temperature products derived from MSG-SEVIRI and Terra/Aqua-MODIS data," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 8, pp. 4163–4170, Aug. 2015.

[19] Y. J. Kaufman and B.-C. Gao, "Remote sensing of water vapor in the near IR from EOS/MODIS," *IEEE Trans. Geosci. Remote Sens.*, vol. 30, no. 5, pp. 871–884, Sep. 1992.

[20] Y. Sun and Y. Ma, "Research on atmospheric water vapor inversion by combining ground GPS data with MODIS data," *Geo. Spat. Info. Tech.*, vol. 42, no. 12, pp. 95–98, Dec. 2019.

[21] B. C. Gao, "Water vapor retrievals using moderate resolution imaging spectroradiometer (MODIS) near-infrared channels," *J. Geophys. Res.*, vol. 108, no. 13, pp. 1–10, 2003.

[22] P. Q. Qu, R. H. Shi, C. S. Liu, and H. L. Zhong, "The evaluation of MODIS data and geographic data for estimating near surface air temperature," *Remote. Sens. Land Resour.*, vol. 91, no. 4, pp. 78–82, Dec. 2011.

[23] W. Zhu, A. Lű, S. Jia, J. Yan, and R. Mahmood, "Retrievals of all-weather daytime air temperature from MODIS products," *Remote Sens. Environ.*, vol. 189, pp. 152–163, Feb. 2017, doi: 10.1016/j.rse.2016.11.011.

[24] W. Wang et al., "All-weather near-surface air temperature estimation based on satellite data over the Tibetan Plateau," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 3340–3350, Mar. 23 2022, doi: 10.1109/JSTARS.2022.3161800.

[25] F. Flores and M. Lillo, "Simple air temperature estimation method from MODIS satellite images on a regional scale," *Chilean J. Agricultural Res.*, vol. 70, no. 3, pp. 436–445, Sep. 2010.
[26] H. Sun, Y. H. Chen, A. D. Gong, X. Zhao, W. F. Zhan, and M. J. Wang, "Estimating mean air temperature using MODIS day and night land surface temperatures," *Theor. Appl. Climatol.*, vol. 118, no. 28, pp. 81–92, Dec. 2014.
[27] C. Vancutsem, T. Dinku, and S. J. Connor, "Evaluation of MODIS land surface temperature data to estimate air temperature in different ecosystems over Africa," *Remote Sens. Environ.*, vol. 114, no. 2, pp. 449–465, 2010.
[28] J. M. Yang and J. H. Qiu, "A method for estimate precipitable water and effective water vapor content from ground humidity parameters," *Chin. J. Atmos. Sci.*, vol. 26, no. 1, pp. 9–22, 2002.
[29] N. Li, Y. M. Xu, M. He, and X. H. Wu, "Retrieval of apparent temperature in Beijing based on remote sensing," *Ecol. Environ. Sci.*, vol. 27, no. 6, pp. 1113–1121, 2018.
[30] G. Peng, J. Li, Y. Chen, A. P. Norizan, and L. Tay, "High-resolution surface relative humidity computation using MODIS image in Peninsular Malaysia," *Chin. Geographical Sci.*, vol. 16, no. 3, pp. 260–264, 2006.
[31] Y. H. Huang, D. Jiang, D. F. Zhuang, and J. Y. Fu, "Estimation of the surface vapor pressure based on the MODIS images," *Prog. Geography*, vol. 29, no. 9, pp. 1137–1142, Sep. 2010.
[32] S. F. Li, L. X. Qian, and J. Wang, "Improved temperature-humidity index based on Landsat TM/ETM+ and its response to impervious surface," *Geography Geo-Inf. Sci.*, vol. 29, no. 2, pp. 112–115, 2002, doi: 10.7702/dlydlxxkx20130222.
[33] G. P. Shi, Y. J. He, M. Zhang, X. F. Qiu, and Y. Zen, "Spatial distribution and topographic influence analysis of temperature and humidity index in Zhejiang province based on GridMet model," *J. Geo-Inf. Sci.*, vol. 21, no. 12, pp. 1923–1933, 2019, doi: 10.12082/dqxxkx.2019.180635.
[34] M. Azarderakhsh, S. Prakash, Y. X. Zhao, and A. Aghakouchak, "Satellite-based analysis of extreme land surface temperatures and diurnal variability across the hottest place on Earth," *IEEE Geosci. Remote Sens.*, vol. 17, no. 12, pp. 2025–2029, Dec. 2020, doi: 10.1109/LGRS.2019.2962055.
[35] X. Liu, Z.-L. Li, J.-H. Li, P. Leng, M. Liu, and M. Gao, "Temporal upscaling of MODIS 1-km instantaneous land surface temperature to monthly mean value: Method evaluation and product generation," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, Feb. 22 2023, Art. no. 5001214, doi: 10.1109/TGRS.2023.3247428.
[36] C. X. Shi et al., "Progress in the development of products based on melting point of multi-source meteorological data," *Acta Meteorologica Sinica*, vol. 77, no. 4, pp. 774–783, 2019, doi: 10.11676/qxxb2019.04.
[37] Y. Liu, C. X. Shi, H. J. Wang, and S. Han, "The applicability of CLDAS temperature data in China," *Trans. Atmos. Sci.*, vol. 44, no. 4, pp. 540–548, 2021, doi: 10.13878 /j.cnki.dqkxxb.20200819001.
[38] Z. H. Yao, D. W. Wu, R. H. Chu, Y. Q. Yao, B. F. He, and Y. Huang, "Dynamic change of vegetation coverage and its response to landform in Anhui Province," *Bull. Soil Water Conservation*, vol. 41, no. 3, pp. 283–289, 2021, doi: 10.13961/j.cnki.stbctb.20210430.001.
[39] T. Shi, Y. J. Yang, J. Ma, L. Zhang, and S. F. Luo, "Spatial-temporal characteristics of urban heat island in typical cities of Anhui province based on MODIS," *J. Appl. Meteorol. Sci.*, vol. 24, no. 4, pp. 484–493, 2013.
[40] F. B. Wu and J. P. Tang, "The impact of urbanization on summer precipitation and temperature in the Yangtze River Delta," *J. Trop. Meteorol.*, vol. 31, no. 2, pp. 255–263, 2015.
[41] Z. H. Yao, Y. Q. Yao, C. H. Wang, F. Fan, and G. P. Shi, "Temporal and spatial characteristics of somatosensory temperature in summer holiday in Anhui province during 1987–2016," *J. Arid Meteorol.*, vol. 37, no. 3, pp. 454–459, 2019, doi: 10.11755 /j.issn.1006-7639(2019)-03-0454.
[42] Y. Y. Hou, J. J. Zhang, H. Yan, and J. L. Wang, "Estimation of regional air temperature using satellite remote sensing data," *Meteorol. Monit.*, vol. 36, no. 4, pp. 75–79, 2010.
[43] H. Q. Zhang, Y. J. Yang, S. P. Xun, B. F. He, A. M. Zhang, and W. Y. Wu, "Seasonal variation and spatial distribution of vegetation and land surface temperature in Anhui Province," *J. Appl. Meteorol. Sci.*, vol. 22, no. 2, pp. 232–240, 2011.
[44] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/a:1010933404324.
[45] A. Liaw and M. Wiener, "Classification and regression by random forest," *R News*, vol. 2, no. 3, pp. 18–22, 2002.
[46] J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining Knowl. Discov.*, vol. 2, pp. 127–167, 1998, doi: 10.1023/A:1009715923555.
[47] V. K. Chauhan, K. Dahiya, and A. Sharma, "Problem formulations and solvers in linear SVM: A review," *Artif. Intell. Rev.*, vol. 52, no. 2, pp. 803–855, 2019, doi: 10.1007/s10462-018-9614-6.
[48] Y. Y. Chen, X. D. Yu, X. H. Gao, and H. Z. Fen, "A new method for non-linear classify and non-linear regression (I): Introduction to support vector machine," *J. Appl. Meteorol. Sci.*, vol. 15, no. 3, pp. 345–354, 2004.
[49] Y. Freund and R. E. Shapire, "Experiments with a new boosting algorithm," in *Proc. 13th Int. Conf. Mach. Learn.*, 1996, vol. 96, pp. 148–156.
[50] G. Ridgeway, "The state of boosting," *Statist. Comput.*, vol. 31, pp. 172–181, 1999.

Danwa Wu received the B.S. degree in weather dynamics from the Nanjing University of Information Science and Technology, Nanjing, China, in 1993, and the M.S. degree in meteorology from the Nanjing University, Nanjing, China, in 2013.

She is currently a Senior Engineer with the Anhui Public Meteorological Service Center, Hefei, China. Her research interests include the professional meteorological services, satellite remote sensing applications in meteorology.

Zhenhai Yao received the M.S. degree in 3S integration and application in meteorology from the Nanjing University of Information Science and Technology, Nanjing, China, in 2015.

He is currently working with the Anhui Public Meteorological Service Center, Hefei, China. His research interests include GIS mapping technique, remote sensing application to land surface, and specialized meteorological service.

Linlin Wu received the M.S. and D.S. degrees in atmospheric physics and atmospheric environment from the Nanjing University of Information Science and Technology, Nanjing, China, in 2006 and 2015, respectively.

He is currently a Professor with the CMA Cloud-Precipitation Physics and Weather Modification Key Laboratory, Beijing, China. His research interests include atmospheric observation and radar meteorology.

Xichang Luo received the M.S. degree in electronic and communication engineering from the Nanjing University of Information Science and Technology, Nanjing, China, in 2014.

He is currently working with the Anhui Public Meteorological Service Center, Hefei, China. His research interests include disastrous weather image recognition and approaching forecast.

Shuai Sun received the B.S. degree in remote sensing science and technology and the M.S. degree in 3S integration and meteorological application from Nanjing University of Information Science and Technology, Nanjing, China, in 2015 and 2018, respectively.

He is currently working with the National Meteorological Information Center, Beijing, China. His research interests include multisource data fusion and land surface data assimilation.

Binfang He received the M.S. degree in engineering information and communication engineering from the University of Science and Technology of China, Hefei, China, in 2009.

He is currently working with the Anhui Institute of Meteorological Sciences, Hefei, China. His research interests include monitoring and evaluating ecological environment based on remote sensing data.

Yali Zhang received the M.S. degree in computer application technology from Central China Teachers University, Wuhan, China, in 2017.

She is currently working with the Anhui Public Meteorological Service Center, Hefei, China. Her research interests include application of image recognition in the field of meteorology and application of deep learning in meteorology.