applied sciences

Article

Modelling Soil Temperature by Tree-Based Machine Learning Methods in Different Climatic Regions of China

Jianhua Dong [1], Guomin Huang [1], Lifeng Wu [1,2,*], Fa Liu [3], Sien Li [4], Yaokui Cui [5], Yicheng Wang [2], Menghui Leng [6], Jie Wu [7] and Shaofei Wu [1]

[1] School of Hydraulic and Ecological Engineering, Nanchang Institute of Technology, Nanchang 330099, China; djh0530dyz@whu.edu.cn (J.D.); g.huang@nit.edu.cn (G.H.); wsf17@nit.edu.cn (S.W.)
[2] State Key Laboratory of Simulation and Regulation of Water Cycle in River Basin, China Institute of Water Resources and Hydropower Research, Beijing 100038, China; wangych@iwhr.com
[3] Key Laboratory of Water Cycle and Related Land Surface Processes, Institute of Geographic Science and Science and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China; liufa@igsnrr.ac.cn
[4] Center for Agricultural Water Research in China, China Agricultural University, Beijing 100083, China; lisien@cau.edu.cn
[5] Institute of RS and GIS, School of Earth and Space Sciences, Peking University, Beijing 100871, China; yaokuicui@pku.edu.cn
[6] Jiangxi Key Laboratory of Hydrology-Water Resources and Water Environment, Nanchang Institute of Technology, Nanchang 330099, China; 18279130275@163.com
[7] School of Civil Engineering and Architecture, Wuhan Polytechnic University, Wuhan 430023, China; wujiemc@whpu.edu.cn
* Correspondence: wulifeng@nit.edu.cn

Abstract: Accurate estimation of soil temperature (T_s) at a national scale under different climatic conditions is important for soil–plant–atmosphere interactions. This study estimated daily T_s at the 0 cm depth for 689 meteorological stations in seven different climate zones of China for the period 1966–2015 with the M5P model tree (M5P), random forests (RF), and the extreme gradient boosting (XGBoost). The results showed that the XGBoost model (averaged coefficient of determination (R^2) = 0.964 and root mean square error (RMSE) = 2.066 °C) overall performed better than the RF (averaged R^2 = 0.959 and RMSE = 2.130 °C) and M5P (averaged R^2 = 0.954 and RMSE = 2.280 °C) models for estimating T_s with higher computational efficiency. With the combination of mean air temperature (T_{mean}) and global solar radiation (R_s) as inputs, the estimating accuracy of the models was considerably high (averaged R^2 = 0.96–0.97 and RMSE = 1.73–1.99 °C). On the basis of T_{mean}, adding R_s to the model input had a greater degree of influence on model estimating accuracy than adding other climatic factors to the input. Principal component analysis indicated that soil organic matter, soil water content, T_{mean}, relative humidity (RH), R_s, and wind speed (U_2) are the main factors that cause errors in estimating T_s, and the total error interpretation rate was 97.9%. Overall, XGBoost would be a suitable algorithm for estimating T_s in different climate zones of China, and the combination of T_{mean} and R_s as model inputs would be more practical than other input combinations.

Keywords: soil temperature; machine learning models; climatic zones; extreme gradient boosting; principal components analysis

check for updates

Citation: Dong, J.; Huang, G.; Wu, L.; Liu, F.; Li, S.; Cui, Y.; Wang, Y.; Leng, M.; Wu, J.; Wu, S. Modelling Soil Temperature by Tree-Based Machine Learning Methods in Different Climatic Regions of China. *Appl. Sci.* **2022**, *12*, 5088. https://doi.org/10.3390/app12105088

Academic Editor: Gerhard Soja

Received: 21 April 2022
Accepted: 17 May 2022
Published: 18 May 2022

1. Introduction

Soil temperature (T_s), as a consequence of the combined effect of hydrothermal circulation in both the atmosphere and the land surface, is an important factor in atmospheric–ecological–environmental systems [1–3]. T_s has significant impacts on the energy balance of global atmospheric–ecological–environmental systems [4,5], thereby directly affecting the growth and development of plants [6]. Under natural conditions, T_s is influenced by other environmental factors such as atmospheric temperature, vegetation type, soil moisture, and topography, which together cause the spatial and temporal variations in T_s [7]. Given its importance and complexity, accurate prediction of T_s is of substantial value in both scientific research and practical applications [8,9]. However, the current approaches for measuring T_s are generally complicated, costly, and time consuming [1], which make the wide range monitoring of T_s inaccessible in many countries, especially for undeveloped countries. The incomplete T_s data caused by this predicament presents an obstacle to the estimation and analysis of T_s conditions and poses a significant challenge for accurate T_s prediction on a large scale [10].

The prevailing approaches for T_s measurements can be divided into three major categories, each of which has its own advantages and disadvantages. The first category is the in situ monitoring of T_s with equipment installed in meteorological stations, which allows the actual T_s to be recorded in real-time [11,12]. However, considering the geographical limitation and the economic cost of constructing meteorological stations and the potential data deficiencies due to equipment problems, this approach is not suitable for studies covering various regions [13]. The second category is the estimation of T_s, which consists of the interpolation method, the Newhorl model estimation, and direct estimation. The interpolation method estimates T_s at any depth based on known T_s at similar depths through certain formulas [14]. The Newhorl model estimation method uses air temperature plus a fixed value to roughly estimate T_s at a certain depth. For instance, Deboer [15] obtained the T_s values at a depth of 50 cm indirectly by adding 2.5 °C to the mean yearly temperature. The direct estimation method simply uses the T_s from known depths to represent T_s at unknown depths [16]. Although approaches in the second category can solve certain problems in T_s measurements to some extent, differences in several factors such as soil depth, elevation, and environment are not considered during estimation, which certainly have impacts on T_s. Therefore, these methods may be suitable for the local scale, but results obtained from them are not universally applicable and cannot be generalized [17]. Therefore, it is necessary to find methods that are applicable to a large-scale range. The third category is to establish a regression equation for T_s calculation. Research has verified similar patterns in the variations between T_s and air temperature [18,19], and therefore relationships between the two temperatures can be used for T_s calculation at various depths through regression equations [20]. Some studies even consider the impacts of geographical factors (e.g., longitude, latitude, and elevation) on T_s and build a multiple linear regression equation to quantify the impact of each geographical factor on T_s [21]. Although approaches in this category overcome the deficiencies of the methods in the other two categories, there are still disadvantages in these approaches as T_s are not necessarily linearly related to geographical factors [22], which might lead to significant biases during calculation [23]. Therefore, it is necessary to explore a more efficient method to perform the calculation.

In recent years, machine learning algorithms have been widely used in studies for estimating hydrological and meteorological indicators, such as air temperature [24], dew point temperature [25], precipitation [26], solar radiation [27], diffuse solar radiation [28], and evapotranspiration [29,30], among which an artificial neural network (ANN) is probably the most common machine learning algorithm used for modeling. For T_s estimation, ANN has been applied in different areas of the world by scholars. The first study on record for estimating T_s with machine learning algorithms was carried out by Yang et al. [31], in which T_s at three depths (i.e., 10, 50, and 150 cm) were estimated with ANN models. Mihalakakou [32] evaluated the potential of ANN in daily and yearly T_s estimation in Athens and Dublin, and the research concluded that the ANN model had an adequate performance in estimating T_s. Bilgili [33] applied an ANN model to estimate monthly T_s at multiple depths in Adana, Turkey, and the result showed that ANN was a suitable model for estimating T_s. Tabari et al. [34] utilized ANN and multiple linear regression model to estimate daily T_s at six soil depths, in which they found that temperature and relative humidity (RH) were the most influential parameters affecting T_s estimation among meteorological factors. Although the ANN model is widely used by everyone for estimating T_s, as scholars continue their research, it was found that the ANN model is not necessarily better than other machine learning algorithms in estimating T_s. Therefore, it becomes more meaningful to keep investigating new models to improve the estimation performance.

Recent advances of newly developed machine learning algorithms have enabled scholars to evaluate and compare the capabilities of various algorithms in estimating T_s, e.g., M5 model tree (M5 Tree), random forests (RF), M5P model tree (M5P), multiple linear regression (MLR) model, support vector machine (SVM) model, and extreme learning machine (ELM) model, etc. Sanikhani et al. [2] compared the capabilities of ELM, ANN, and M5 Tree in modeling monthly T_s at 5, 50, and 100 cm depths, and they found that the ELM model would be a more desirable tool for estimating T_s at a wide range of depths. Mehdizadeh et al. [35] reported that the adaptive neuro-fuzzy inference system (ANFIS) provided superior results in monthly T_s estimation than ANN and gene expression programming (GEP) with data from 31 stations in Iran. Bilgili et al. [36] estimated monthly T_s at different soil depths (i.e., 5, 10, 20, 50, and 100 cm) using nonlinear regression and MLR models, and obtained the most accurate results at the depth of 5 cm. Feng et al. [37] evaluated the capabilities of ELM and RF models in estimating half-hourly T_s for maize fields, and they concluded that the ELM model had better estimated performance. Sihag et al. [38] utilized M5P, MLP, and RF models to estimate daily T_s in arid regions, the results of which showed that MLP outperformed other models with better performance metrics. Kisi et al. [39] also compared the modeling capabilities of MLP, MLR, and radial basis neural networks (RBNN) for monthly T_s estimation at different soil depths, with inconsistent results obtained depending on depth. At depths of 5 and 10 cm, the RBNN model had the best estimating accuracy among all models used, while at other soil depths (50 and 100 cm), MLR performed better. Mehdizadeh et al. [40] applied SVM and multivariate adaptive regression splines (MARS) to estimate month-by-month T_s for 30 stations in Iran, and it was found that MARS outperformed the SVM model.

Overall, according to the literature, machine learning algorithms have been extensively used in the research of T_s estimation, with the scopes of such studies generally covering various time scales, study regions, and/or soil depths. However, to our best knowledge, the performance of machine learning algorithms on T_s estimation has rarely been reported in the different climatic zones of China, especially for studies with large spatial scales (such as the whole country). Accurate estimation of T_s in different climatic zones would be beneficial to studies of crop modeling, hydrological patterns, and soil properties under large-scale conditions, and may also provide theoretical support for better crop production in different regions [2,41,42]. Recently, extreme gradient boosting (XGBoost) is considered as a promising machine learning algorithm, and has been widely applied in many fields such as meteorology [28,43], hydrology [44,45], and agronomy [46,47]. However, to date no study has been conducted on a national wide estimation of T_s with XGBoost. Therefore, in this study we utilize six different combinations of meteorological variables as inputs to develop XGBoost models for estimating daily T_s at the 0 cm depth, with records from th eperiod 1966–2015 from 689 meteorological stations that covers seven different climatic zones in China. To better evaluate the estimating performance of XGBoost, the M5P and RF algorithms that have applied for T_s estimation in previous studies are added for comparison. The primary purpose of this research is to evaluate the performance of tree-based models to estimate daily T_s in seven different climatic zones (large-scale) of China and its applicability. The effects of different climatic and environmental factors on the performance of each model for estimating T_s are then explored.

2. Materials and Methods

2.1. Study Area

According to previous studies, there are seven different climatic zones in China based on geographical and meteorological data [48,49], which are the arid desert of northwest China (NWC), the semi-arid steppe of Inner Mongolia (IM), the (semi-) humid cold-temperate northeast China (NEC), the semi-humid warm-temperate north China (NC), the humid subtropical central China (CC), the humid tropical south China (SC), and the Qinghai-Tibetan Plateau (QTP) [50] (Figure 1). Among these climatic zones, the QTP has the highest average elevation (up to 4000 m), with considerably high UV intensity and relatively low vegetation coverage. The region is mainly composed of mountains, plateaus, river valleys, and basins. The geology is mainly schist, millstone, sandstone, shale, and volcanic rocks, The QTP region has a variety of ecosystem types and mainly grows barley. The NWC generally has longer daylight hours than other climatic zones. The landscape is dominated by mountains, basins, and deserts. The area is sparsely vegetated, and the desert covers a large area. The crops and fruits are of high yield with excellent quality in places where water is relatively abundant. The IM region receives less rainfalls than other regions. The altitude is generally between 1000 and 1200 m. The terrain is high in the south and low in the north, with 80% of the plateau covered by grassland, which is mainly clumped with grasses. The thickness of the grass layer is between 10 and 60 cm. There are seven main lake areas in the region, including the Hetao and Hulunbuir Lakes. The NEC region has long winters and short summers, with a significant amount of snow and a wet climate. The region is dominated by plains and mountains. There are many marshes and thick mounds of earth. The proximity to the Bohai Sea leads to a long rainy season in the NEC region. The region has large areas of coniferous and mixed coniferous forests and rich black soils. Forestry accounts for 14.7% of the country's land. The main food crop is wheat. The NC is a relatively small region of China with less than 1000 m^3 of water resources per capita. Summers are hot and rainy. Winters are cold and dry. Rainfall varies between 400 and 800 mm. The land type is dominated by the great plains and plateaus, with the terrain decreasing from west to east. The soil type is mainly brown loam. The CC and SC regions generally receive heavy rainfalls in spring, where flooding tend to occur during that time period. The CC region has red loamy soils of low fertility. The hills, plains, basins, rivers and lakes dominate. The main crops grown are rice and oilseed rape. The topography of the SC region is mainly hilly and plain. The average annual temperature is 18–24 °C. Moreover, the area has a large variety of plants, mainly tropical scrub, grassy slopes, and secondary forests. The soil types are mainly brick red loam and russet red loam.

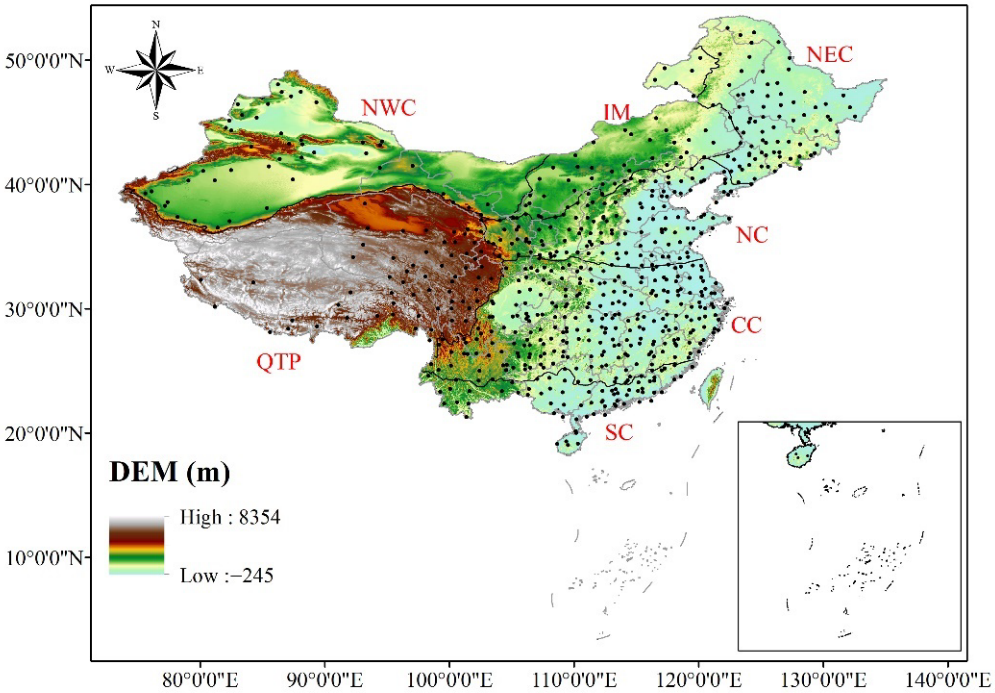

Figure 1. The geographical distribution of meteorological stations used in this study.

2.2. Case Study and Data

In this study, daily meteorological data from 1966 to 2015 from 689 meteorological stations located in the seven climatic zones of China were collected. Meteorological data consisted of mean air temperature (T_{mean}), maximum air temperature (T_{max}), minimum air temperature (T_{min}), relative humidity (RH), global solar radiation (R_s), and wind speed (U_2). The selected data were quality controlled. All weather-related variables were measured at a 2 m height. However, it should be noted that R_s was not recorded in all the 689 stations. To fix this problem, the extraterrestrial solar radiation (R_a) was used for interpolation through modelling, if the R_s data were absent. T_s at a depth of 0 cm were measured through probes on the soil surface. Six input combinations were utilized for training the M5P, RF, and XGBoost models. Table 1 showed the types of input combinations. Raw meteorological data used in the study were provided and quality inspected by the National Meteorological Information Center (NMIC) of the China Meteorological Administration (CMA). A further examination was applied to datasets before computing. A number of data were deleted from the original data because they were missing or the ratio between measured T_s and theoretical T_s was greater than one. T_{mean} was converted from T_{max} and T_{min}. The M5P, RF, and XGBoost programs were written in R software (version 3.2.3; The R Project for Statistical Computing). All the simulations were performed in a computer with a single Intel Core i7-6700 at 3.4–4.0 GHz and 16 GB of random-access memory (RAM).

Table 1. The input combinations of meteorological variables for various machine learning models.

Models			Input Combinations
M5P	**RF**	**XGBoost**	
M5P1	RF1	XGBoost1	T_{mean}, RH, R_s, U_2
M5P2	RF2	XGBoost2	T_{mean}
M5P3	RF3	XGBoost3	T_{mean}, RH
M5P4	RF4	XGBoost4	T_{mean}, R_s
M5P5	RF5	XGBoost5	T_{max}, T_{min}
M5P6	RF6	XGBoost6	T_{mean}, U_2

2.3. Methodology

2.3.1. M5P Model Tree (M5P)

The M5P model tree (M5P) is a regression tree algorithm that develops conventional decision trees with the addition of linear regression functions to the nodes [51]. This technique has been successful in predicting continuous values, which can be achieved by employing the conversion of the classification problem into a functional optimization problem [52]. The M5 Tree denotes a segmented linear function, testing the value of a specific property at each internal node and predicting the class value at each leaf node. To predict the class value of a new sample, the tree is interpreted starting from the root node. At each internal node, the left or the right branch is selected based on the value of a particular attribute of the sample. The advantage of model trees over regression trees is that regression trees' computational load increases rapidly as the dimensionality increases, while model trees are significantly less than conventional models. Therefore, model trees are more efficient when handling with high-dimensional data as leaf nodes use linear functions rather than constants, leading to more accurate predictions. All enumerated attributes in the M5P algorithm are converted into binary variables before the tree is constructed. This algorithm can efficiently handle missing values. There are three main steps in the M5P tree algorithm, which are tree construction, tree pruning, and tree smoothing [53]. The basic tree is constructed with the splitting criterion, which takes the standard deviation of the class values reaching a node as the error of that node. It computes the expected error reduction resulting from testing each property at that node. The property that maximizes the expected error reduction is then selected.

2.3.2. Random Forests (RF)

The random forests (RF) algorithm, proposed by Breiman [54], was developed using classification and regression trees (CART) and the concept of "bagging". As a machine learning algorithm that can effectively solve high-dimensional regression problems, RF has been used extensively in research of regression and estimation, which uses subsets of 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 the samples may be collected repeatedly. A random subset of the training dataset needs to be randomly extracted from the original dataset for the development and training of the model (see the flowchart in Figure 2). Datasets that are not used in the model training are often called out-of-bag (OOB) data. These OOB datasets will not be used for model fitting, but in turn will be used for testing the estimation ability of the model [25].

The RF algorithm is a feature selection based on the Gini coefficient. The criterion for selecting the Gini coefficient is that each child node needs to achieve the highest purity. The smaller the Gini coefficient, the higher the stability of the model and the higher the purity. CART is a binary tree, which means that each non-leaf node can only produce two branches. If multiple (taller 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 produced from all the features, which reduces the risk of overfitting. In contrast to other decision trees, each RF tree is part of the selected feature [25]. Among the selected features in this part, the best feature is picked to partition the left and right subtrees of the decision tree, thereby providing more randomness and further improving the model's inductive power [25]. In short, the final estimation of the RF algorithm is the average of all factors. Breiman's research has more details about the RF algorithm [54].

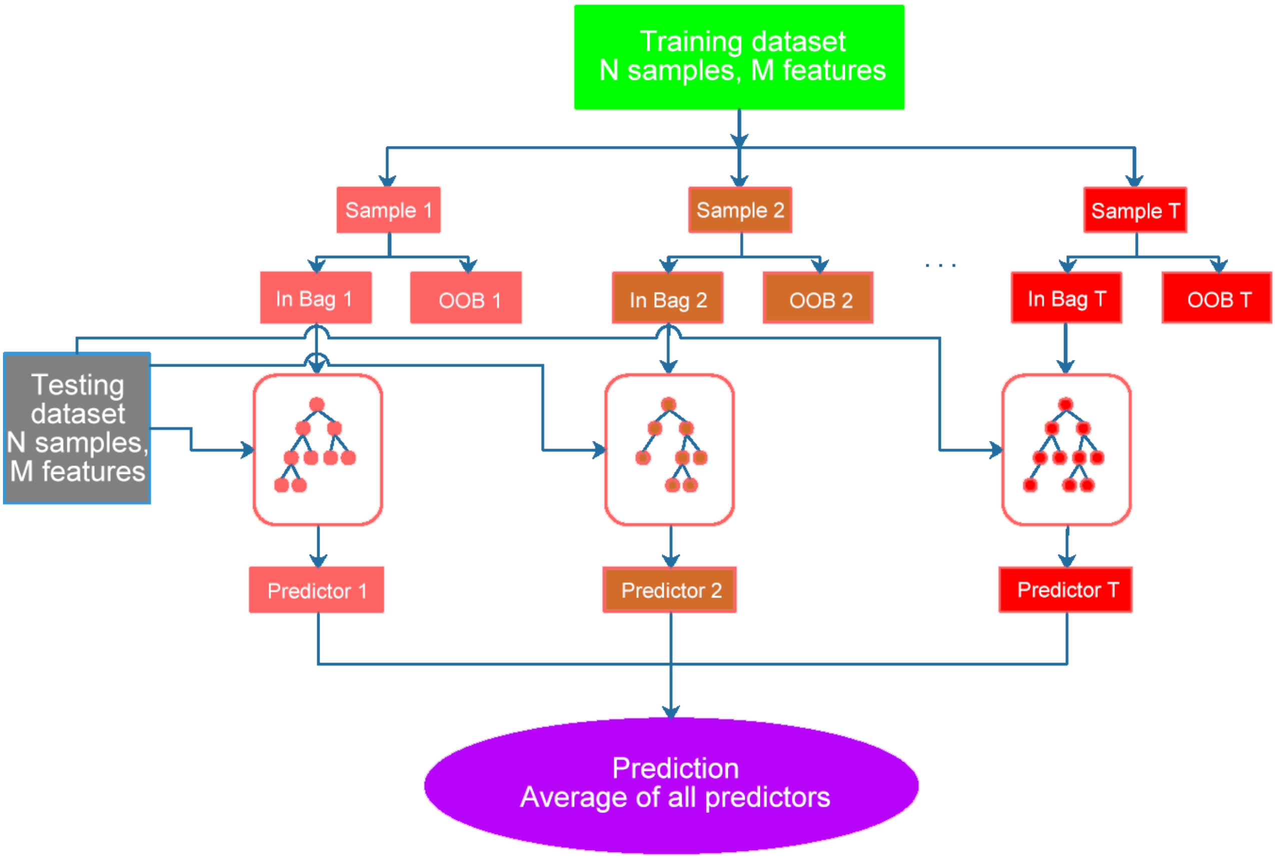

Figure 2. The flow chart of the RF algorithm.

2.3.3. Extreme Gradient Boosting (XGBoost)

Extreme gradient boosting (XGBoost) is a new paradigm for Gradient Boosting Machines (GBMs), which improves the processing of databases through the optimization of decision tree algorithms [55]. The XGBoost algorithm solves the overfitting problem through regularization and built-in cross-validation, which improve the computational accuracy and allow for optimum speed. Furthermore, functions in the XGBoost algorithm are operated and computed automatically, and therefore XGBoost is extensively applied in applications such as classification [56] and estimation [57]. The XGBoost algorithm is derived from the concept of "boosting", which combines the forecasts of all "weak" learners with special training to foster "strong" learners. The expressions are as follows:

$$f_i^{(t)} = \sum_{k=1}^{t} f_k(x_i) = f_i^{t-1} + f_t(x_i) \tag{1}$$

where $f_k(x_i)$ and $f_t(x_i)$ are the predicted values for the k-th and t-th iterations of the XGBoost model, respectively; $f_i^{(t)}$ and f_i^{t-1} are the predicted values for the t and $t-1$ iterations of the i-th sample; x_i is the input variable; k = [1, 2, . . . , t], i = [1, 2, . . . , n].

To prevent overfitting problems without compromising the speed of the algorithm, the XGBoost algorithm can be derived as follows:

$$Obj^{(t)} = \sum_{i=1}^{n} l\left(\overline{f_i^{(t)}}, f_i^{(t)}\right) + \sum_{i=1}^{n} \Omega(f_i) \tag{2}$$

where $Obj^{(t)}$ is the objective function; l is the loss function; $\overline{f_i^{(t)}}$ is the true value of the t-th iteration of the i-th sample; and $\Omega(f_i)$ is the canonical term of the objective function, which is given by:

$$\Omega(f) = \beta T + \frac{1}{2}\lambda\|\omega\|^2 \tag{3}$$

where β and λ are regularization parameters and T is the number of leaf nodes.

2.3.4. Model Evaluation

In this study, three statistical indicators, including coefficient of determination (R^2), root mean square error (RMSE), and mean absolute error (MAE), were chosen to analyze and compare the accuracy and stability of different models for estimating T_s [50,58]. The corresponding formulas are:

$$R^2 = \frac{\left[\sum_{i=1}^{n} (Y_{i,m} - \overline{Y}_{i,m})(Y_{i,e} - \overline{Y}_{i,e})\right]^2}{\sum_{i=1}^{n} (Y_{i,m} - \overline{Y}_{i,m})^2 \sum_{i=1}^{n} (Y_{i,e} - \overline{Y}_{i,e})^2} \tag{4}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (Y_{i,m} - Y_{i,e})^2} \tag{5}$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} |Y_{i,m} - Y_{i,e}| \tag{6}$$

where $Y_{i,m}$, $Y_{i,e}$, $\overline{Y}_{i,m}$, $\overline{Y}_{i,e}$, and n are the measured daily T_s, the estimated daily T_s, the measured average value of daily T_s, the estimated average value of daily T_s, and the number of observations, respectively. Higher R^2 values (closer to 1) and lower RMSE and MAE values indicate better estimation performance of the model.

3. Results and Discussion

3.1. Comparison of the Accuracy of Various Machine Learning Models

The overall and average performance of each data-driven model in the seven different climatic zones is shown in Tables 2 and 3, respectively. As listed in Table 2, averaged across the six input combinations and the seven climatic zones, the XGBoost model (on average R^2 = 0.964, RMSE = 2.066 °C, MAE = 1.597 °C) overall performed better in T_s estimation than the RF (on average R^2 = 0.959, RMSE = 2.130 °C, MAE = 1.647 °C) and M5P models (on average R^2 = 0.954, RMSE = 2.280 °C, MAE = 1.742 °C). Among the six input combinations, models with the complete combination of meteorological variables as inputs (i.e., the input combination 1) had the greatest estimating accuracy in each of the seven climatic zones (Table 3). Under this input combination, averaged across the meteorological stations in each climatic zone, the ranges of the mean statistical indicator values for M5P1, RF1, and XGBoost1 at different climatic zones were 1.722–2.875 °C, 1.354–2.257 °C, and 1.342–2.208 °C for RMSE, 0.928–0.975, 0.956–0.984, and 0.959–0.985 for R^2, and 1.245–2.106 °C, 1.028–1.702 °C, and 0.992–1.651 °C for MAE, respectively. The second best estimating accuracy among input combinations was observed in models with the combination of T_{mean} and R_s (i.e., the input combination 4) as inputs, which were M5P4, RF4, and XGBoost4 (with corresponding ranges of 1.658–2.830 °C, 1.506–2.556 °C, and 1.446–2.435 °C for RMSE, 0.929–0.975, 0.941–0.980, and 0.9494–0.983 for R^2, and 1.214–2.093 °C, 1.119–1.915 °C, and 1.084–1.840 °C for MAE, respectively). Although input combination 4 yielded slightly worse model accuracy than the complete combination, it required less input variables and a much smaller dataset. It was noted that models with T_{mean} as inputs all achieved relatively good performance in estimation, including models with the complete input combination (M5P1, RF1, and XGBoost1) and the combinations of T_{mean} and RH (M5P3, RF3, and XGBoost3), T_{mean} and R_s (M5P4, RF4, and XGBoost4) and T_{mean} and U_2 (M5P6, RF6, and XGBoost6). These results confirm the argument that temperature is the most fundamental meteorological factor for data-driven models in estimating T_s. Among the other three meteorological factors, R_s had the highest degree of influence on estimating accuracy, followed by RH, and then U_2. Therefore, considering the model estimating accuracy and the size of input data jointly, models using the input combination of T_{mean} and R_s would have a higher potential in T_s estimation across different climatic zones in China than models with other input combinations. For the three algorithms in this study, the overall model performance was generally ranked as XGBoost > RF > M5P with any of the input combinations in any of the climatic zones, with the exception that M5P2 was always better than RF2 in all climatic zones and even better than XGBoost2 in some climatic zones (e.g., NWC, IM, and QTP).

Table 2. Summary of the performance of data-driven models for soil temperature in different climate zones of China.

Different Zone	M5P			RF			XGBoost		
	RMSE	R^2	MAE	RMSE	R^2	MAE	RMSE	R^2	MAE
NWC	2.358	0.971	1.800	2.213	0.974	1.709	2.175	0.976	1.694
IM	2.679	0.967	2.071	2.486	0.971	1.944	2.417	0.974	1.907
NEC	2.912	0.957	2.211	2.700	0.963	2.075	2.584	0.967	1.988
NC	2.197	0.966	1.688	2.042	0.970	1.588	1.984	0.973	1.539
CC	1.748	0.961	1.302	1.643	0.965	1.239	1.608	0.970	1.202
SC	1.735	0.925	1.330	1.645	0.931	1.274	1.604	0.941	1.226
QTP	2.330	0.930	1.794	2.181	0.938	1.697	2.090	0.945	1.625
Mean	2.280	0.954	1.742	2.130	0.959	1.647	2.066	0.964	1.597

Note: RMSE and MAE are in °C.

Table 3. Performance of data-driven models for the estimation of soil temperature in the different climatic zones of China.

Model	NWC			IM			NEC			NC			CC			SC			QTP		
	RMSE	**R^2**	**MAE**	**RMSE**	**R^2**	**MAE**	**RMSE**	**R^2**	**MAE**	**RMSE**	**R^2**	**MAE**	**RMSE**	**R^2**	**MAE**	**RMSE**	**R^2**	**MAE**	**RMSE**	**R^2**	**MAE**
M5P1	2.155	0.975	1.601	2.515	0.971	1.878	2.875	0.958	2.106	2.077	0.969	1.552	1.722	0.962	1.245	1.728	0.928	1.283	2.239	0.934	1.688
RF1	1.738	0.984	1.319	1.996	0.981	1.533	2.257	0.974	1.702	1.642	0.981	1.257	1.374	0.975	1.028	**1.354**	0.956	**1.035**	1.760	0.959	1.352
XGBoost1	**1.656**	**0.985**	**1.258**	**1.904**	**0.985**	**1.453**	**2.208**	**0.976**	**1.651**	**1.580**	**0.983**	**1.200**	**1.342**	**0.979**	**0.992**	1.367	**0.959**	1.036	**1.671**	**0.963**	**1.274**
M5P2	2.212	0.974	1.723	2.506	0.972	1.994	2.737	0.962	2.146	2.103	0.969	1.673	1.709	0.963	1.321	1.713	0.925	1.365	2.219	0.936	1.748
RF2	2.301	0.972	1.795	2.607	0.969	2.070	2.851	0.959	2.228	2.162	0.967	1.719	1.745	0.962	1.349	1.736	0.923	1.384	2.271	0.933	1.788
XGBoost2	2.474	0.970	1.937	2.710	0.968	2.154	2.719	0.964	2.114	2.166	0.969	1.684	1.673	0.968	1.267	1.626	0.940	1.250	2.304	0.935	1.806
M5P3	2.364	0.971	1.810	2.656	0.968	2.054	2.877	0.958	2.191	2.183	0.966	1.668	1.698	0.963	1.254	1.695	0.927	1.290	2.261	0.935	1.735
RF3	2.273	0.973	1.756	2.515	0.971	1.962	2.729	0.962	2.099	2.073	0.969	1.601	1.655	0.965	1.233	1.694	0.927	1.296	2.187	0.938	1.697
XGBoost3	2.212	0.975	1.735	2.455	0.974	1.952	2.651	0.966	2.057	2.039	0.972	1.597	1.639	0.969	1.223	1.664	0.937	1.274	2.111	0.945	1.647
M5P4	2.153	0.975	1.623	2.438	0.973	1.847	2.830	0.960	2.093	2.006	0.972	1.520	1.658	0.965	1.214	1.678	0.929	1.262	2.234	0.935	1.707
RF4	1.951	0.980	1.489	2.203	0.978	1.684	2.556	0.967	1.915	1.811	0.977	1.389	1.506	0.971	1.119	1.532	0.941	1.168	2.035	0.946	1.571
XGBoost4	1.870	0.983	1.440	2.118	0.980	1.640	2.435	0.971	1.840	1.759	0.978	1.356	1.448	0.974	1.084	1.446	0.949	1.105	1.896	0.954	1.465
M5P5	2.849	0.958	2.185	3.176	0.955	2.474	3.154	0.950	2.414	2.490	0.957	1.906	1.858	0.957	1.375	1.774	0.921	1.351	2.632	0.913	2.025
RF5	2.658	0.963	2.057	2.926	0.961	2.304	2.921	0.957	2.258	2.303	0.963	1.780	1.746	0.962	1.302	1.716	0.925	1.312	2.458	0.923	1.913
XGBoost5	2.520	0.968	1.973	2.737	0.967	2.176	2.730	0.963	2.124	2.166	0.968	1.681	1.697	0.967	1.277	1.645	0.938	1.259	2.302	0.934	1.797
M5P6	2.416	0.969	1.855	2.781	0.965	2.182	2.999	0.955	2.316	2.322	0.962	1.812	1.844	0.957	1.401	1.821	0.916	1.430	2.397	0.927	1.860
RF6	2.360	0.971	1.834	2.671	0.967	2.114	2.884	0.958	2.249	2.261	0.964	1.780	1.834	0.957	1.405	1.838	0.914	1.448	2.377	0.927	1.864
XGBoost6	2.317	0.974	1.822	2.580	0.972	2.067	2.758	0.963	2.145	2.196	0.969	1.715	1.849	0.963	1.369	1.875	0.927	1.432	2.255	0.938	1.758

Note: RMSE and MAE are in °C. The bolded numbers are the best values for model estimation performance.

The box plots of the RMSE values of the three data-driven models (i.e., M5P, RF, and XGBoost models) for estimating T_s with the six input combinations at the 689 stations of China are shown in Figure 3. The RMSE values of the XGBoost model under the six input combinations were 1.001–2.823 °C, 1.272–3.480 °C, 1.121–3.259 °C, 1.050–2.981 °C, 1.173–3.540 °C, and 1.231–3.422 °C, respectively, which overall were lower than the corresponding values of the M5P model (1.241–3.618 °C, 1.056–3.447 °C, 1.133–3.638 °C, 1.174–3.478 °C, 1.250–4.059 °C, and 1.107–3.758 °C, respectively) or the RF model (1.051–2.873 °C, 1.082–3.611 °C, 1.111–3.431 °C, 1.090–3.123 °C, 1.170–3.761 °C, and 1.153–3.619 °C, respectively). Averaged across all stations, the mean RMSE values for the XGBoost model under the six input combinations decreased by 23.1%, −1.4%, 5.5%, 13.2%, 11.2%, and 3.4% when compared with the M5P model, and by 3.0%, 1.4%, 1.9%, 4.3%, 4.5%, and 1.7% when compared with the RF model. The median RMSE values of the three models showed a similar pattern to the mean RMSE values. These results together indicated that, among the three models in this study, the XGBoost model had the highest overall performance, followed by the RF model and then the M5P model.

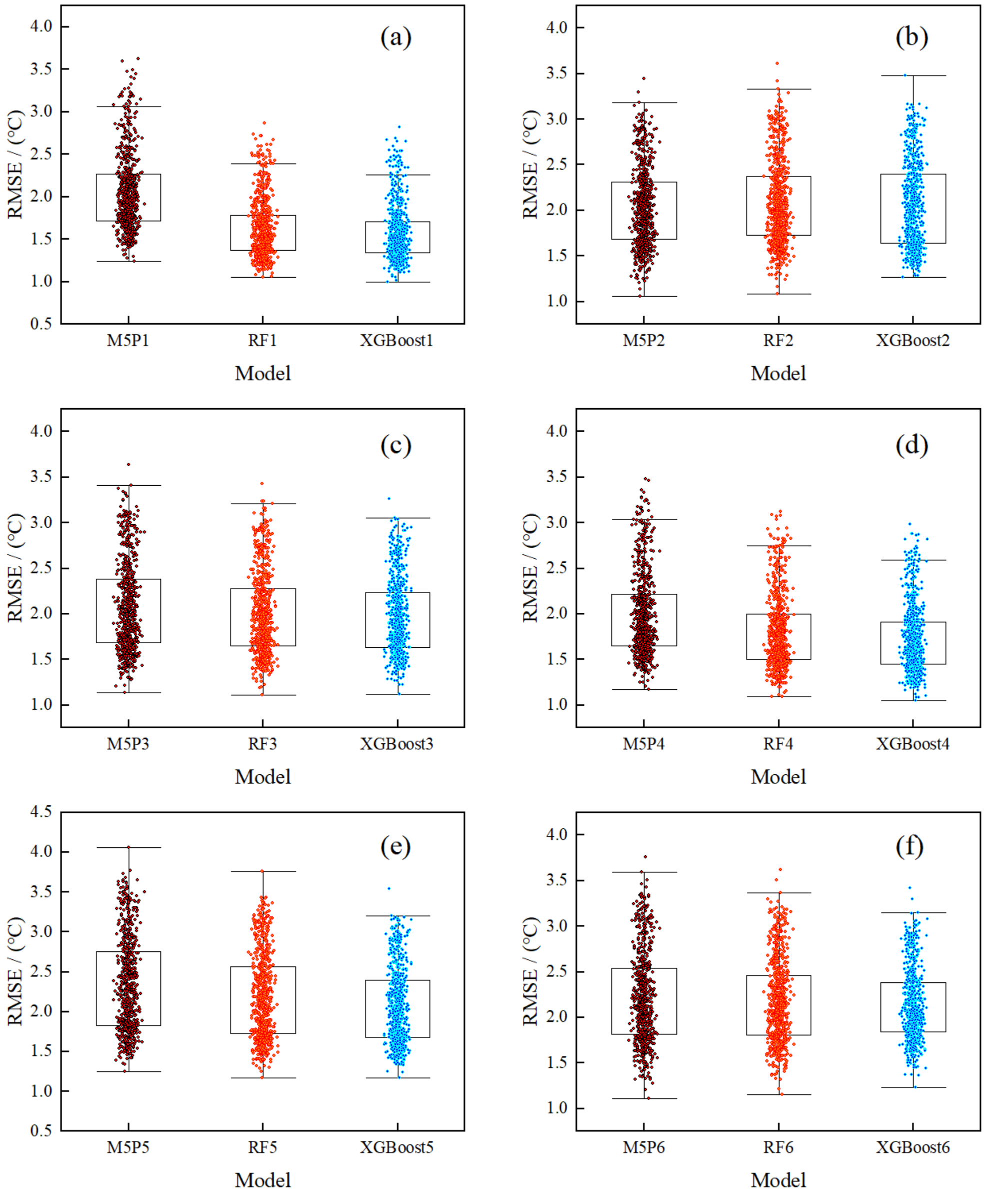

Figure 3. Box plots of the RMSEs of the three models for estimating T_s under six input combinations (**a**–**f**).

3.2. Comparison of the Spatial Distribution of Errors in Estimating Soil Temperature

Spatial distributions of the performance in estimating T_s for the data-driven models with all climatic variables as inputs (M5P1, RF1, and XGBoost1 models) are shown in Figure 4. Generally, the models all performed well in estimating T_s at the majority of stations, with the values of RMSE less than 2.0 °C and the values of MAE less than 1.50 °C at 76.8% of all stations studied. Overall, there was some spatial variation in model performance between climatic zones. The accuracy of the model was relatively better in the CC and SC regions than in other regions, while the model performance was relatively poor in the NEC region, especially in the northeast provinces of China. Among the three models, M5P1 had the worst performance in estimation, with RMSE values greater than 2.01 °C and MAE values greater than 1.50 °C in up to 47.7% of the stations in all regions, except for the CC region where only a small fraction of stations had similar poor performance. On the contrary, the XGBoost1 model had the most satisfactory in estimating T_s, with RMSE values less than 1.51 °C and MAE values less than 1.25 °C in most stations of all regions, except for the NEC region.

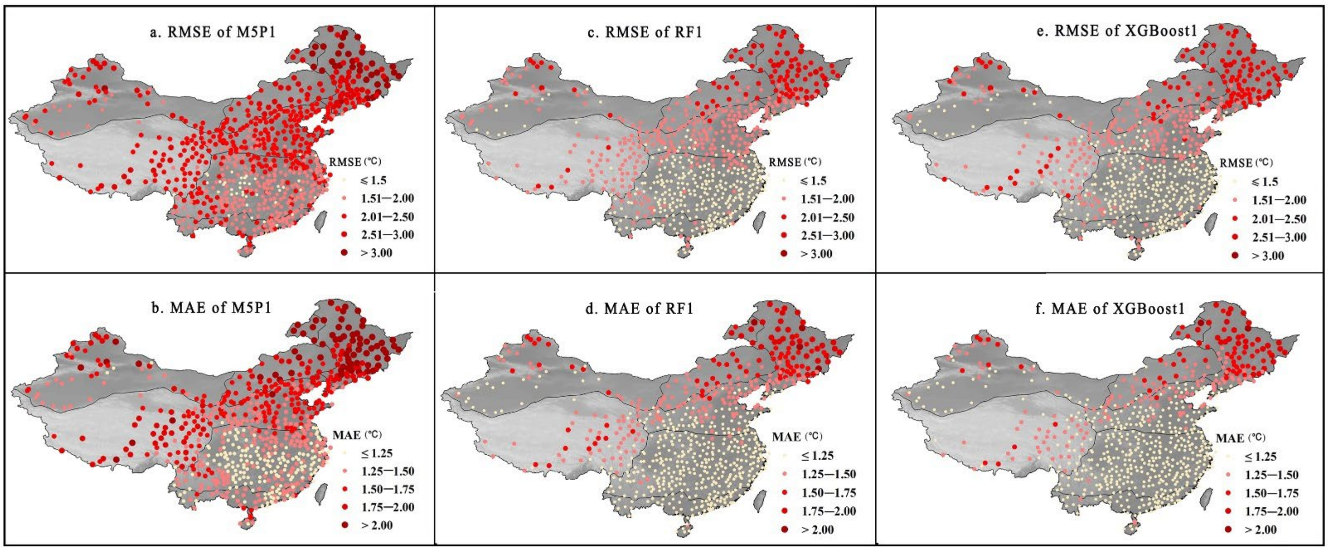

Figure 4. Performance of data-driven models (M5P1: (**a**,**b**); RF1: (**c**,**d**); XGBoost1: (**e**,**f**)) for the estimation of T_s with T_{mean}, RH, R_s, and U_2 as inputs.

The performance of the M5P2, RF2, and XGBoost2 models for estimating T_s are illustrated in Figure S1. The three models were trained using T_{mean} as the only input available. The three models generally had a positive performance in estimating T_s at the majority of stations, with RMSE values greater than 2.01 °C and MAE values greater than 1.50 °C at 48.8% of the stations. Although there was still relatively good estimating accuracy in some stations of the CC and SC regions, the overall performance of the three models was poor in other regions of this study. In particular, the IM and NEC regions had up to 71.9% of the stations with model RMSE values greater than 2.51 °C and MAE values greater than 2.00 °C. Overall, the estimation accuracy of the XGBoost2 model was slightly greater than both other models. These results suggest that there is space for improvement in the estimation accuracy of the three models when using T_{mean} as the only input. Our findings also confirm the conclusion of Talaee's study that the appropriate addition of other meteorological factors to T_{mean} could effectively improve model performance in estimating T_s [59].

The spatial distribution of the performance for the data-driven models (M5P3, RF3, and XGBoost3 models) on T_s estimation with T_{mean} and RH as inputs is illustrated in Figure S2. The models generally had positive performance in estimating T_s at the majority of stations, with RMSE values less than 2.0 °C and MAE values less than 1.50 °C at

56.2% of the stations. Among the climatic zones, the three models performed relatively better in the CC region than other regions, while the poorest model performance was observed in the NEC region, where the RMSE and MAE values in 78.5% of the stations were greater than 2.51 °C and 2.00 °C, respectively. Models showed very similar estimating performance between the NWC and the QTP regions. Compared with the M5P1, RF1, and XGBoost1 models, the accuracy of models with only T_{mean} and RH as inputs was slightly worse. Therefore, reducing the number of meteorological factors as inputs would lower the estimation performance of each model, which is consistent with findings obtained by Feng et al. [60]. This result suggests that using a complete set of relevant meteorological factors as inputs in studies of T_s estimation could effectively improve the estimating performance of models, which confirms the findings from previous studies [61].

The spatial distribution of estimating performance for M5P4, RF4, and XGBoost4 models that were trained with T_{mean} and R_s as inputs are illustrated in Figure S3. The models all showed positive performance in modeling T_s for a large proportion of stations, with RMSE values less than 2.0 °C and MAE value less than 1.50 °C at up to 71.6% of the stations. Among the climatic zones, the estimating accuracy of the three models was relatively excellent in the CC and SC regions, with RMSE values less than 1.50 °C and MAE values less than 1.25 °C at 45.5% of the stations, while in contact the model performance in the NEC region was relatively poor, with up to 64.9% of the stations showing values of RMSE and MAE greater than 2.51 °C and 1.75 °C, respectively. Among the three models, the estimating performance of XGBoost4 was slighter better than RF4, but both of them obviously outperformed M5P4. Overall, patterns of the spatial variability in model performance for this input combination (i.e., T_{mean} and R_s) were very similar to patterns shown in models with the complete combination of meteorological variables as inputs, indicating that the removal of RH and U_2 from the inputs did not noticeably decrease the estimation accuracy of models in this study. Moreover, the input combination of T_{mean} and R_s required far less meteorological data than the complete input combination. Therefore, it can be concluded that T_{mean} and R_s are key input variables for estimating T_s in different climate zones, consistent with findings obtained by Talaee [59] and Bilgili [21].

For the data-driven models with T_{max} and T_{min} as inputs (i.e., M5P5, RF5, and XGBoost5), the spatial distribution of their estimating performance is demonstrated in Figure S4. Overall, the three models performed poorly in estimating T_s at most stations, with 55.8% of the stations showing the values of RMSE and MAE above 2.01 °C and 1.50 °C, respectively. Models with T_{max} and T_{min} as input variables performed worse than the M5P1, RF1, and XGBoost1 models, indicating that using temperature as the only input climatic variable would not be sufficient enough to obtain satisfactory model performance.

The spatial distributions of the estimating performance for the M5P6, RF6, and XGBoost6 models that were trained using T_{mean} and U_2 as inputs are shown in Figure S5. The three models showed poor performance in modeling T_s at most stations, with RMSE values greater than 2.01 °C and MAE values greater than 1.50 °C at 58.7% of the stations. Although the estimating accuracy of the three models was relatively good in the CC region, their estimating performance was poor in other regions. For instance, in the IM and NEC regions, up to 78.4% of the stations showed values of RMSE and MAE greater than 2.51 °C and 2.00 °C, respectively.

A combined comparison of Figures S2–S5 indicates that the models showed an overall trend of decreasing accuracy from south to north. This would lead us to think that it is related to latitude, whereas latitude could affect T_s due to the angle of incidence by solar radiation [62]. Regarding the effect of latitude on T_s, scholars have reached similar conclusions in studies from other countries. In a study conducted by Fitton and Brooks [63] in the USA, it was concluded that T_s decreased with increasing latitude. The model with inputs T_{mean} and R_s would be slightly more accurate in estimation than under several other combinations of inputs, which was very close to the performance of each model at full input of the parameter variables. Based on T_{mean}, the factor R_s had a greater influence on the accuracy of each model than several other meteorological factors, followed by RH and U_2. Because this research estimated T_s at 0 cm depth, the air temperature was most closely related to T_s [64]. However, the greater degree of influence of R_s might be due to the source of soil heat. T_s is one of the components of soil heat. However, soil heat is mainly formed by solar radiation, heat emitted from the Earth's interior, and heat generated by the decomposition of microorganisms in the soil [65]. Of these, solar radiation dominates. Therefore, R_s had a greater degree of influence in estimating T_s. Nahvi et al. [66] and Huang et al. [67] also concluded that T_{mean} was significantly relevant in estimating T_s. It suggested that the input combination of T_{mean} and R_s was the most generalizable for studies estimating T_s in the different climatic zones of China.

3.3. Analysis of Factors Affecting the Model's Estimation of Soil Temperature

To compare the level of linearity between the estimated and the measured values of T_s for each model, we randomly selected one meteorological station per region (including stations 51,477, 50,618, 50,353, 52,986, 56,188, 59,287, and 52,984, respectively) out of the 689 stations in the seven different climatic zones across the country as examples for description. Outputs from models with input combinations 1, 5, and 6 were plotted for comparison (Figure 5, Figures S6 and S7). It is clear from Figure 5, Figures S6 and S7 that the three types of models (i.e., the M5P, RF, and XGBoost models) all performed very well in estimating T_s. For any of the models constructed with the selected input combinations in the randomly selected stations, there was a significant positive correlation between the estimated and the measured T_s values, with R^2 values at least higher than 0.92.

For models with the complete combination of climatic variables (i.e., combination 1) as inputs, XGBoost outperformed both M5P and RF at each of the randomly selected stations (Figure 5). Across the stations, the average R^2 values of XGBoost models were 0.981, higher than RF (averaged R^2 = 0.978) and M5P (averaged R^2 = 0.966) models. For each type of model, the estimating accuracy differed among stations, suggesting that variations in geographical and environmental factors among different climatic zones could affect the model performance in T_s estimation to some extent, as suggested by Kassaye et al. [68]. Taking XGBoost as an example, except for station 59,287 (R^2 = 0.966), the model accuracy was considerably high at all stations (all R^2 values no less than 0.980; averaged R^2 = 0.983), with the highest accuracy observed at station 51,477 (R^2 = 0.989). The significant difference in model estimating accuracy between station 51,477 and station 59,287 could be attributed to the differences in climatic factors (e.g., temperature) between climatic zones. Station 51,477 is located in the NWC region, with sufficient light, low soil water contents, and low variation of T_s. In contrast, station 59,287 is within the SC region, where the soil moisture is sufficient and the variation in T_s is high. First of all, we selected stations that are all weather stations. The terrain of the station measurement points is short grass and flat, so the slope and some human factors are not considered to influence T_s. However, soil characteristics differ between climatic zones. For example, the soil in the CC area is mainly red loam. The main characteristic of red loam soil is strong clay and little organic matter. Therefore, the soil heat transfer is faster, which means that T_s does not change much. Thus, the model estimated T_s with high accuracy. However, the soil in the NEC area is mainly black soil. It is characterized by fertile land and high organic matter content. However, as we know, high organic matter content means higher heat requirements for microbial decomposition. Therefore, the variation in T_s varied widely, leading to a decrease in the accuracy of model estimation. Zhang et al. [69] concluded that the variation of T_s in different climatic zones was influenced by a combination of air temperature and precipitation. More precipitation occurs in humid areas and less in arid regions. These findings are consistent with the conclusion made by Knight et al. [70], where they suggested that T_s could be directly affected by air temperature and indirectly impacted by soil moisture. In addition, other factors such as soil depth, soil characteristics, irrigation levels, and precipitation could also affect the estimation of T_s to some extent, as described in the literature [9,37,71–73]. For example, the variability of T_s at different soil depths generally vary, with greater variability between measured and estimated values observed in deeper soils. The literature indicates that the thermal conductivity of soil decreases with increasing soil porosity but increases with increasing soil water content [74]. Therefore, both enhanced rainfall and reduced porosity could improve the thermal conductivity of the soil, which could make T_s less variable [75]. In addition, the increase in the organic matter content of the soil also reduces the thermal conductivity [76]. The degree of irrigation, similar to precipitation, can indirectly affect the soil water content. The increase in soil water content increases the heat capacity in the soil, which results in a smaller range of variation in T_s [74].

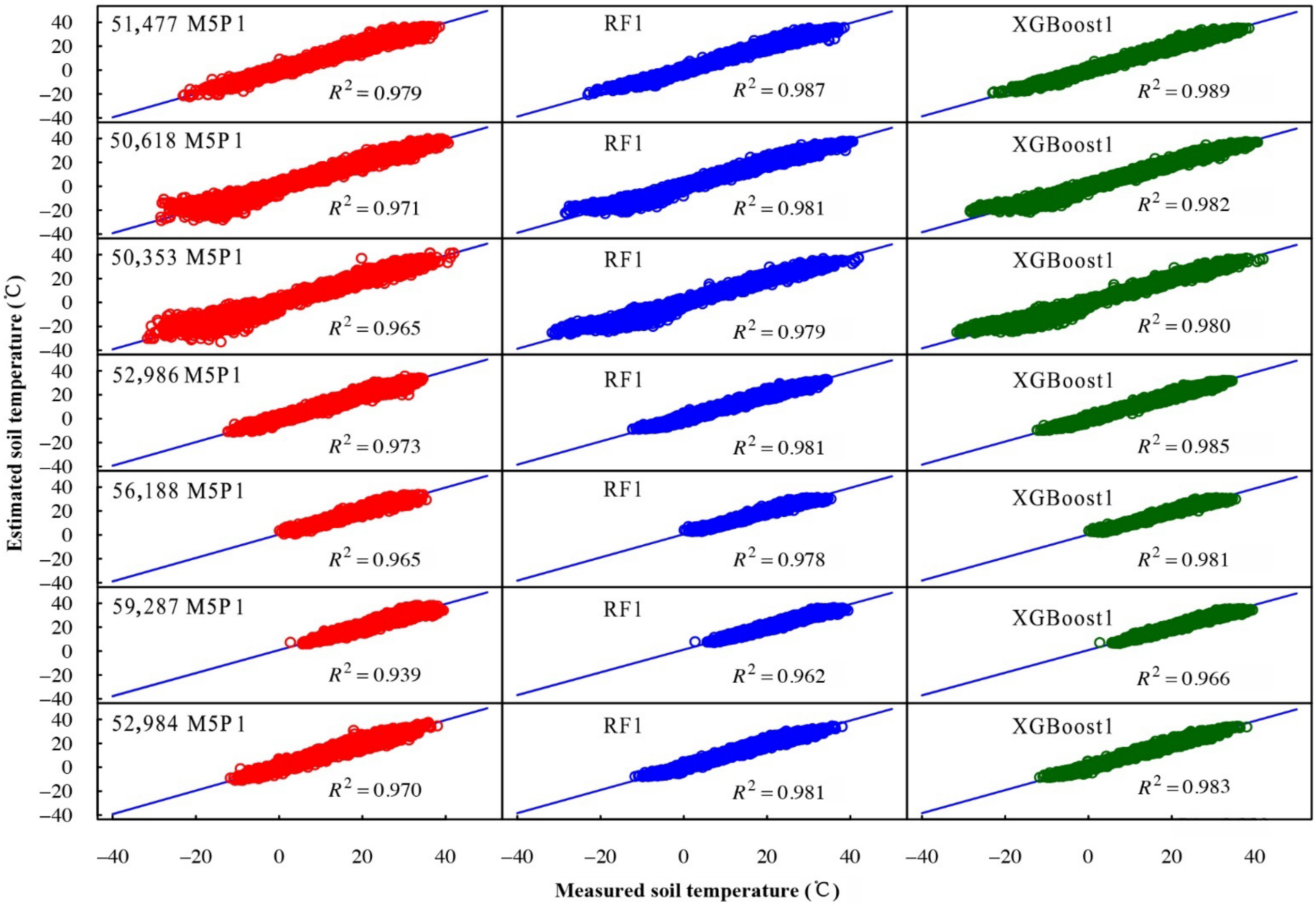

Figure 5. Scatter plot of estimated T_s versus measured T_s from the three machine learning models with T_{mean}, RH, R_s, and U_2 as inputs.

With combination 5 (i.e., including T_{max} and T_{min}) as inputs (Figure S6), the accuracy of each model at each station was lower than the accuracy of its corresponding model with the complete input combination, indirectly suggesting that adding more relevant meteorological factors to the inputs would improve the model performance in estimating T_s. This result is consistent with the conclusion made by Sanikhani et al. [2] and Kim and Singh [6]. Among the selected stations, the best-performing models were observed at station 56,188, where the R^2 values of the M5P, RF, and XGBoost models were 0.969, 0.971, and 0.976, respectively (Figure S6). For station 56,188 (within the CC region) and station 59,287 (within the SC region), the range of variations in T_s was smaller than other stations within other climatic zones, resulting in the scatters of the regression at the two stations being denser than at other stations. This outcome might be due to the fact that these two stations are located at regions where the climatic difference between them is similar, but they are distinct from other stations in terms of climate. Similar patterns were found by other studies, in which the estimation of reference crop evapotranspiration (ET_0) in different climatic zones was conducted [50].

For combination 6 (i.e., including T_{mean} and U_2 as inputs), XGBoost models outperformed both RF and M5P models in terms of estimating accuracy, similar to other input combinations (Figure S7). Across the selected stations, the R^2 values of XGBoost models averaged 0.966, slightly higher than the RF and M5P models, which had the same estimating accuracy on average (both averaged R^2 = 0.960). Under this input combination, the accuracies of the M5P and RF models were very close at each of the stations. At some of the stations (e.g., stations 56,188 and 59,287), the M5P model even slightly outperformed the RF model. Among these stations, the best model performance in estimating T_s was observed in station 51,744, where the XGBoost model had the highest accuracy (R^2 = 0.976). Compared to models with input combination 5, there were improvements to some extent in the estimating accuracy of models with input combination 6 at all selected stations, except for station 59,287. The estimating accuracies of the M5P, RF, and XGBoost models at station 59,287 decreased by 2.0%, 2.5%, and 2.1%, respectively, when comparing the input combinations. This result might be due to the high variability of wind speeds in the region where station 59,287 was located, which could impair the model accuracy in estimation. A previous study by Kim and Singh [6] had similar findings, which confirmed that U_2 could significantly affect the performance of estimated T_s.

To quantify the effects of soil water content, soil organic matter content, and other meteorological factors on T_s estimation by models in different climatic zones, the relationships between these factors and the RMSE values of XGBoost1 were analyzed (Figure 6). The spatial distribution of soil organic matter content and soil water content for the selected stations are plotted in Figure 7. Data from different climatic zones were fitted separately. For each of the soil texture factors or the meteorological factors, its correlation with the RMSE of T_s differed between climatic zones, suggesting that the model accuracy in estimating T_s was related to the soil texture and the climate of the region where the meteorological stations were located. Different soil textures have different permeability and nutrient contents, which partially leads to differences of T_s variability and consequently affects the model accuracy for T_s estimation. This pattern is consistent with results described in the previous section.

As shown in Figure 6a, the most significant correlation between soil organic matter content and model estimating accuracy was found in the NC region (R^2 = 0.255), followed in order by the IM, the NWC, and the CC regions. In other regions, including NEC, SC, and QTP, no significant correlation was found. In regions where the correlation was significant, the model accuracy for estimating T_s decreased with the increasing soil organic matter content. This is due to the fact that the specific heat capacity of soil organic matter is less than the soil specific heat capacity. Under certain conditions, the higher the organic matter content is, the more heat absorption should be required for decomposition, leading to more variable and unstable T_s. Wang et al. [77] also concluded that soil organic matter could significantly affect the variation of T_s. According to the results, the NEC, SC, and QTP regions had relatively high soil organic matter contents, indicating that too much soil organic matter may lead to other types of reactions within the soil and make the model estimation of T_s less stable [78,79].

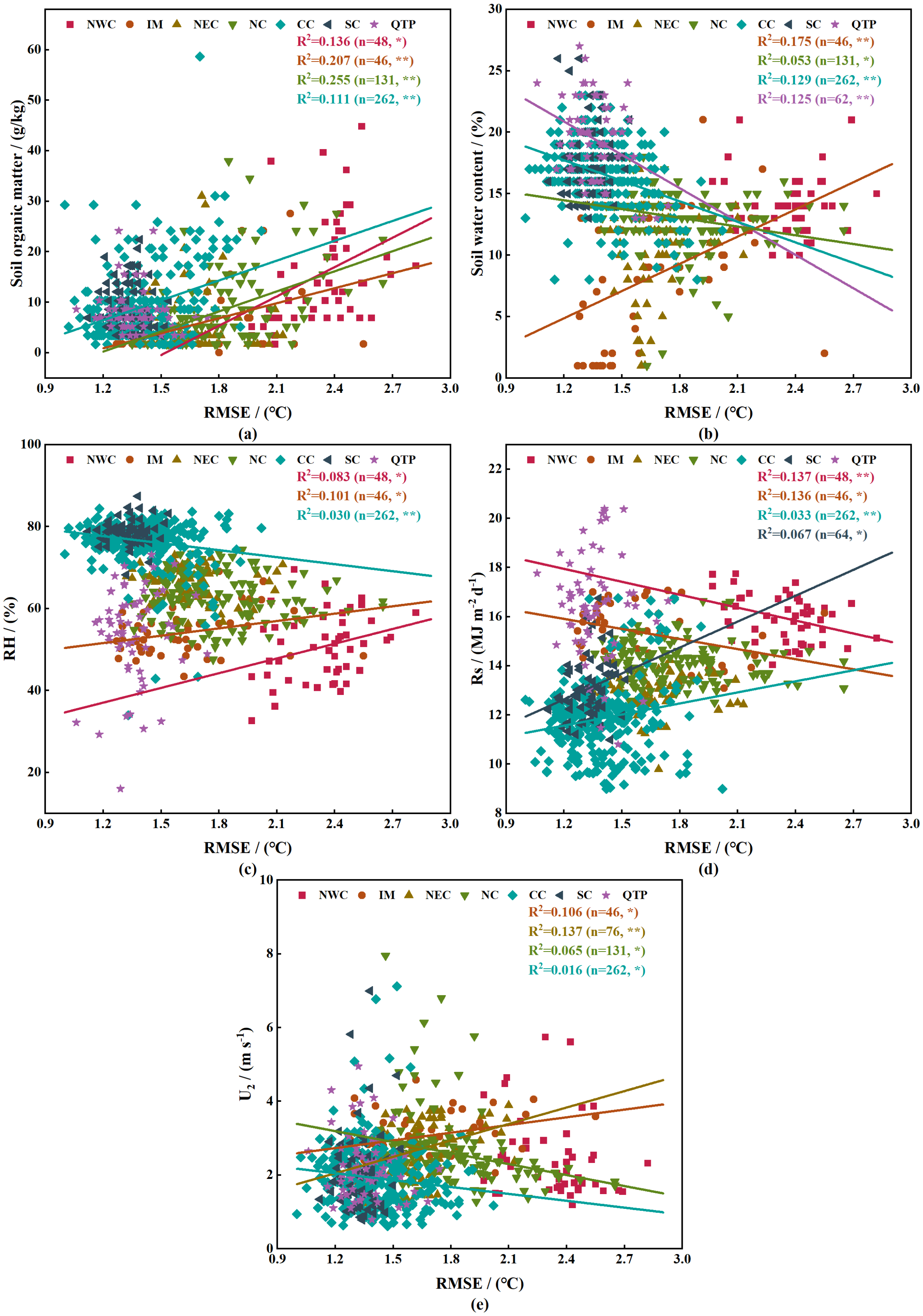

Figure 6. Fitting relationships between several factors ((**a**) soil organic matter; (**b**) soil water content; (**c**) RH; (**d**) R_s; (**e**) U_2) and model performance in estimating T_s (* indicates significant correlation. ** indicates highly significant correlation.).

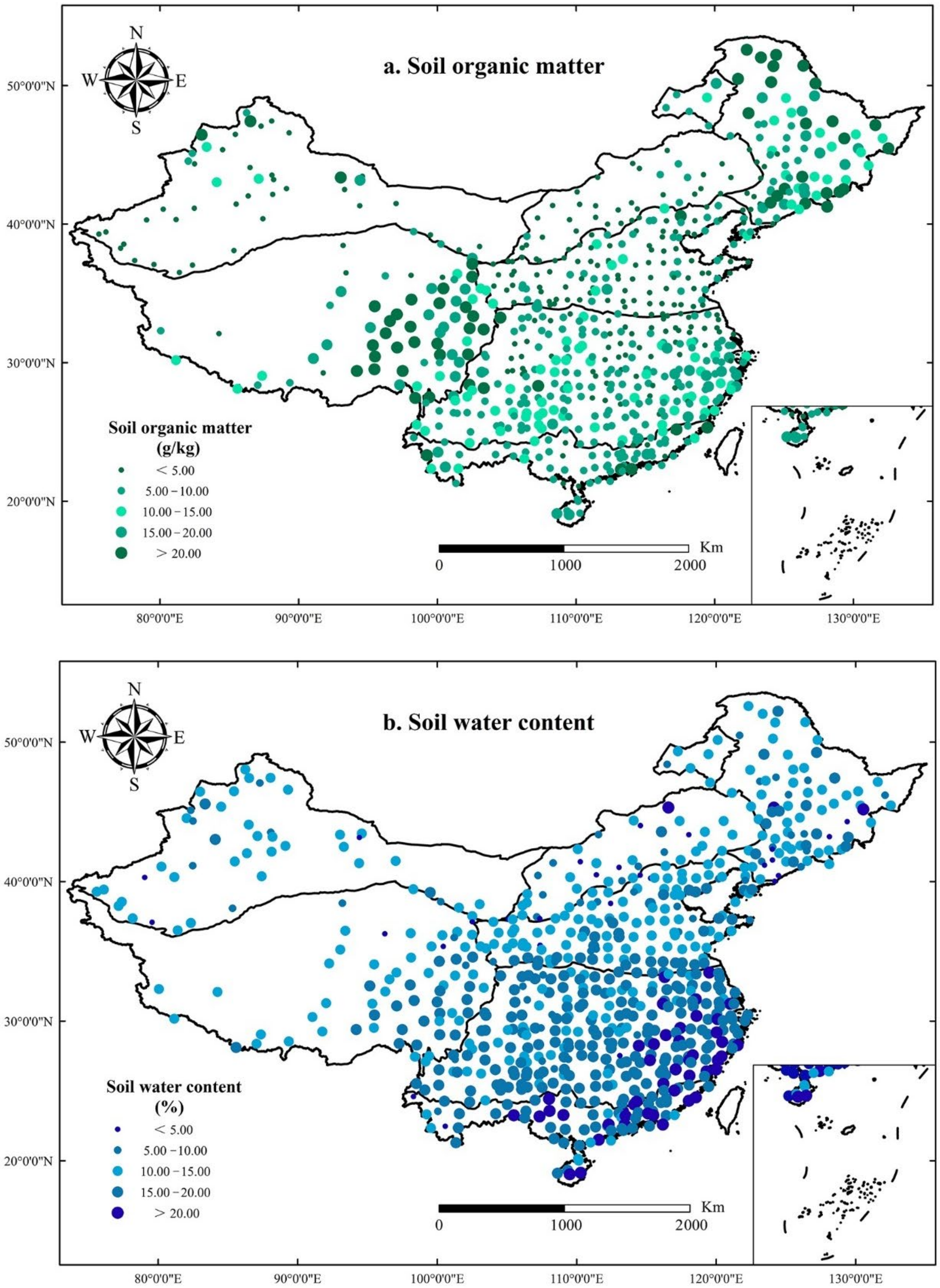

Figure 7. Spatial distribution of soil organic matter content (**a**) and soil water content (**b**).

The correlation between soil water content and model estimating accuracy was most significant in the IM region ($R^2 = 0.175$; Figure 6b), followed by the CC region, and then the QTP and the NC regions. Soil water content was not significantly correlated with the accuracy of estimating T_s in the NWC, NEC, and SC regions. In most regions where the correlation was significant, the model estimating accuracy increased with the increasing soil water content. This might be attributed to the fact that the specific heat capacity of water is about four times higher than that of soil. The temperature change of the soil water is much smaller than the soil itself when equally heated or cooled, suggesting that soil water can indirectly regulate T_s and reduce the range of T_s variation. Yuan [80] also concluded that soil water content had an important influence on soil thermal properties and could effectively regulate T_s. Therefore, under certain conditions, the higher the soil water content is, the smaller is the expected range of T_s variation, leading to higher model accuracy in T_s estimation. This finding is consistent with conclusions from previous studies [81,82]. It should be noted that a contrasting pattern was observed in the IM region, where the model estimating accuracy decreased with the increasing soil water content. This phenomenon might be attributed to the relatively low soil water content in this region. For most stations in the IM region, their soil water contents were lower than 10%, which was distinct from the other six regions in the study. Therefore, soil water content might not play an important role in determining the pattern of model estimating accuracy in the IM region, in which other factors such as soil organic matter might be more influential.

RH was highly significantly correlated with the accuracy of model-estimated T_s in the CC region, which might be due to the fact that the region is in a humid area with high perennial air humidity (Figure 6c). When the relative humidity is high, there is considerable moisture in the air. Given the fact that the specific heat capacity of water is larger than air, the heat transferred by the air temperature and solar radiation to soils is partially absorbed by the moisture in the air, thus reaching the amount of heat to the surface of the soil. Sawan [83] also suggested that RH had a strong influence on air temperature variation and indirectly affected the heat absorbed by the soil.

Solar radiation is one of the main sources of soil heat. In our study, R_s showed significant correlation with model accuracy for estimating T_s in the NWC and CC regions, which might be caused by the longer insolation time in the NWC region and the higher insolation intensity in the low-latitude CC region (Figure 6d). Solar radiation can be directed to the surface, allowing the surface to absorb heat and thus increase the T_s. High latitudes have a high tilt of solar irradiation, so that less solar radiation energy is absorbed on the soil per unit area. This results in lower T_s. In contrast, the T_s is higher at low latitudes because the solar radiation absorbed per unit area of soil is higher due to direct sunlight to the ground [62]. Olchev et al. [84] also confirmed the important role of solar radiation in the conversion of T_s.

Wind speed affects the flow of air. As shown Figure 6e, U_2 was highly significantly correlated with the accuracy of model-estimated T_s in the NEC region, which might be due to the fact that the NEC is in a cold-temperate region with frequent cold air and high wind speed all year round. When the U_2 is low, heat exchange between soil and air is reduced, thus reducing the variation in T_s. Conversely, the heat loss from air is generally fast when U_2 is high, resulting in significant variation in T_s. Kong et al. [85] studied the effect of seasonal variation in U_2 on T_s and suggested that U_2 affects the heat uptake by soils to some extent, thus causing changes in T_s.

We also explored the extent of the contribution of these factors to the estimating error of the model and the relationship between them through principal component analysis (PCA) and correlation coefficients (Figure 8). PCA was used to explain the degree of influence of the factors on the model error. As shown in Figure 8a, the total model error explained by each factor was 97.9%. Among them, PC1 accounted for 96.2% and PC2 accounted for 1.7%. Soil water content, T_{mean}, R_s, and U_2 were all located in the same quadrant. These factors, illustrated in Figure 8a, are shown to be the main factors contributing to the error in estimating T_s. Zhang [86], when he studied the effect of snow cover on T_s, also proposed that air temperature alone does not fully explain the variation in T_s. The R_s, soil water content, and RH can also combine to influence the variation of T_s, whereas soil organic matter content and U_2 were additionally proposed to explain the T_s variation in this study. For each of the factors, there was a significant relationship between it and the estimated RMSE values of T_s (Figure 8b). Among these factors, U_2 had the lowest correlation (R = 0.14), whereas RH had the highest correlation (R = −0.48) and soil water content had the second highest correlation (R = 0.38), indicating that the heat absorption capacity of water had a more obvious effect on T_s than other factors such as T_{mean}, R_s, and soil organic matter content. To summarize, soil organic matter, soil water content, T_{mean}, RH, R_s, and U_2 are the main factors that cause errors in estimating T_s, all of which would have, to some extent, a critical impact on the model estimation of T_s.

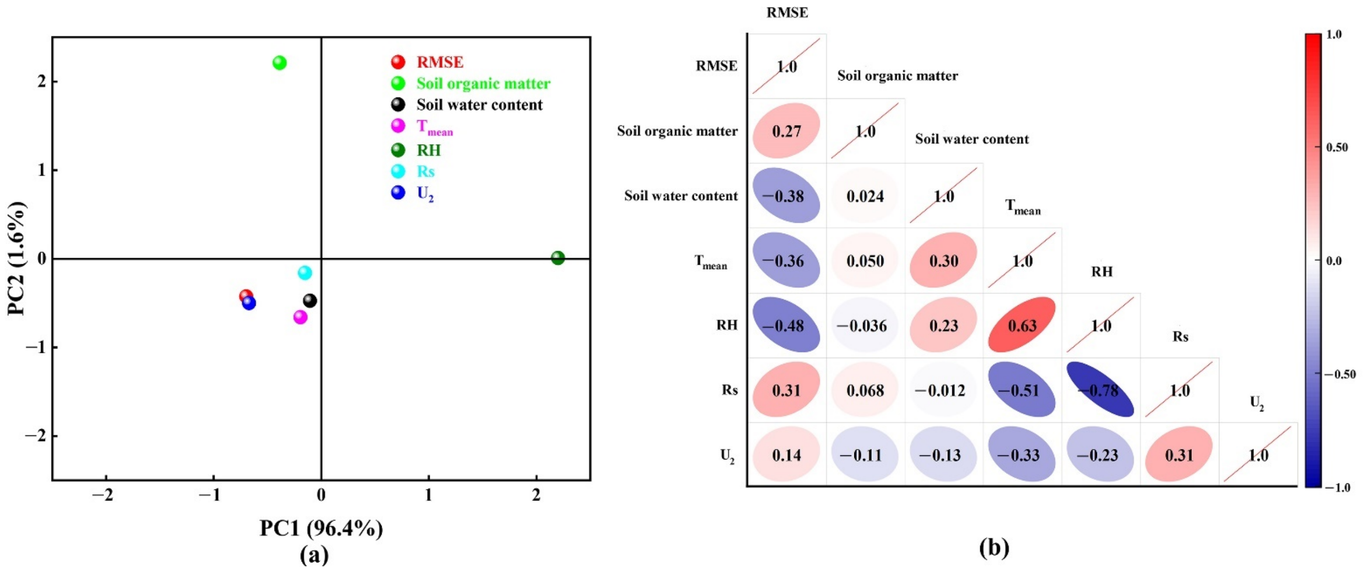

Figure 8. Principal component analysis (**a**) and correlation analysis (**b**) between several factors and model estimation T_s errors.

3.4. Comparison of Estimation Performance in Different Climate Zones

To compare the estimating performance for each of the 18 models (3 machine learning algorithms × 6 input combinations) at each of the stations within each of the climatic zones in this study, a combined metric based on the RMSE and MAE values of each model was calculated. The value of the combined metric ranged from 0 to 1, in which the closer the value to 0, the better the model performance was, while on the contrary the closer the value to 1, the worse the model performance was. Values of the combined metric at different stations for each climate zone are illustrated with heatmaps (see Figure 9 and Figures S8–S13). Overall, under the same input combination, the estimation performance of the XGBoost model was generally better than that of the RF and M5P models at most stations (Figure 9 and Figures S8–S13). Among all the 18 models with different input combinations, the XGBoost1 and RF1 models generally performed best at most stations in the seven climate zones, of which XGBoost1 had better performance than RF1. Among the incomplete input combinations, models with combination 4 (i.e., including T_{mean} and R_s) generally performed better than models with other combinations, of which the models could be ranked as XGBoost4 > RF4 > M5P4 in terms of estimation performance. It is noted that the performance of the XGBoost4 model was relatively close to that of the XGBoost1 model, while the former required fewer input datasets than the latter. Therefore, the XGBoost4 model would be more generalizable than the XGBoost1 model for estimating T_s across different climatic zones. Overall, the model performance differed between different climatic zones. For example, the M5P5 and RF5 models had relatively poor estimating performance in the NWC, IM, NEC, NC, and QTP regions, while the estimation performance of the M5P6, RF6, and XGBoost6 models was poor in the SC and CC regions. These results confirmed that the model estimating performance would be impacted by variations among different climatic zones.

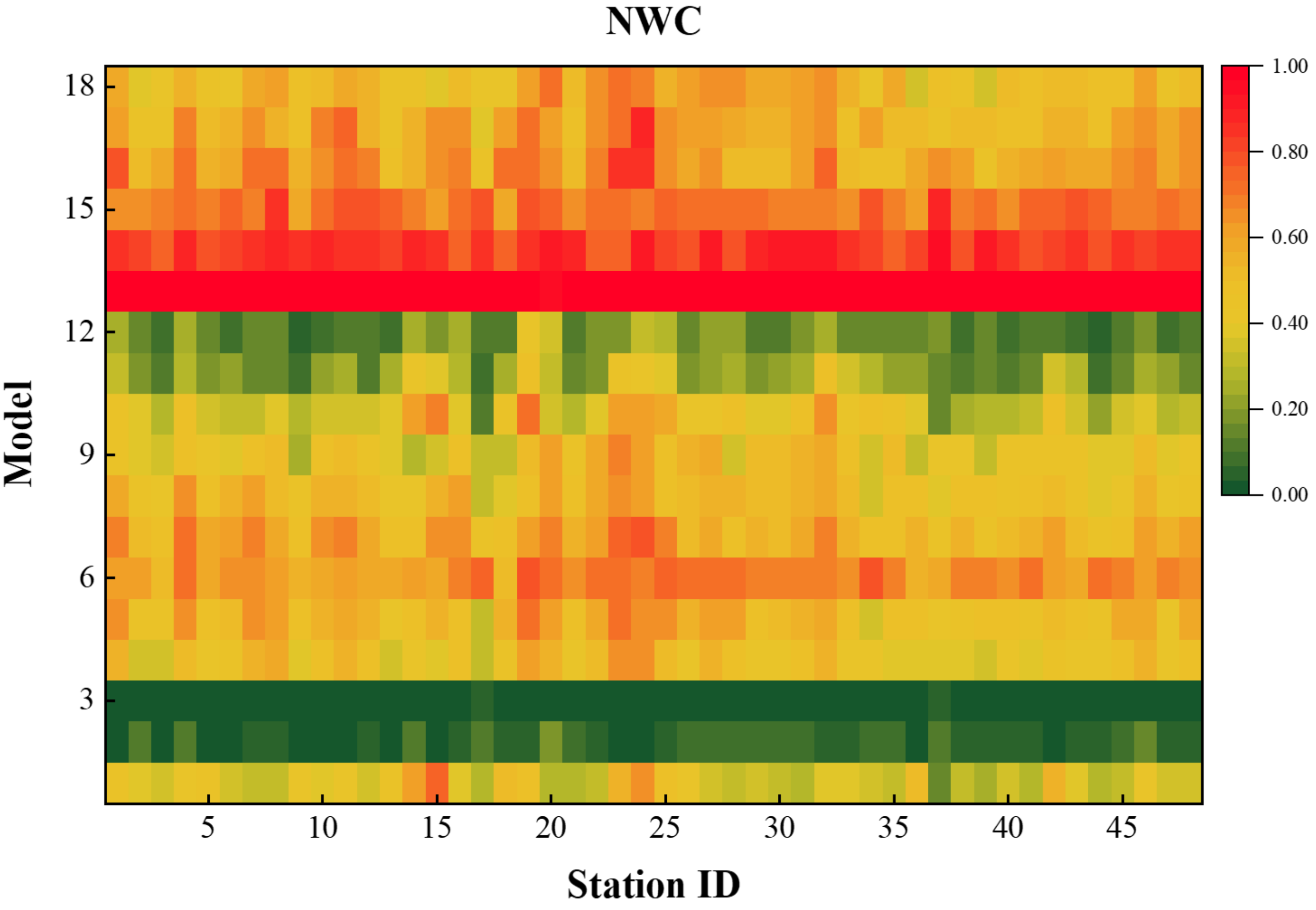

Figure 9. Combined indicator values for each model at all stations in the NEC region. (Model 1–18 in the *Y* axis represents M5P1, RF1, XGBoost1, M5P2, RF2, XGBoost2, M5P3, RF3, XGBoost3, M5P4, RF4, XGBoost4, M5P5, RF5, XGBoost5, M5P6, RF6, and XGBoost6, respectively, which also applies for Figures S8–S13).

3.5. Comparison of Stability of Various Machine Learning Models

To further compare the performance of different machine learning models, the absolute errors (AE) for each model was also adopted as a statistical metric. The frequency distribution of AE between measured and estimated daily T_s derived from the three machine learning models under various input combinations during the testing stage was plotted (Figure 10). As shown in the histograms, the AE of the three models mainly distributed at the ranks of <0.3 °C, while the proportions of AE greater than 0.5 °C for the models only ranged from 1.0% to 3.9%, which together suggested that the three types of models all performed well in T_s estimation. The proportions of AE less than 0.3 °C differed among model types and input combinations. For the input combination of T_{mean} and R_s (i.e., combination 4), the proportions of M5P and XGBoost models with AE less than 0.3 °C ranged from 87.4% to 90.9%, higher than the proportions of their corresponding counterparts with the input combination of T_{max} and T_{min} (ranging from 81.6% to 85.1%). Kisi et al. [39] also concluded that the input combination of T_{mean} and R_s provided a more effective estimation of T_s in the Turkish region. For the complete input combination, the proportions of M5P and XGBoost models with AE less than 0.3 °C ranged from 87.0% to 93.1%. Moreover, the XGBoost model had the highest proportion of AE less than 0.1 (up to 55.2%) with this input combination among all models, indicating that the XGBoost model had stronger overall estimating performance than both other models. In addition, the XGBoost model outperformed the RF and M5P models in data processing speed, with less computing time required, consistent with results obtained by former research [29,87,88].

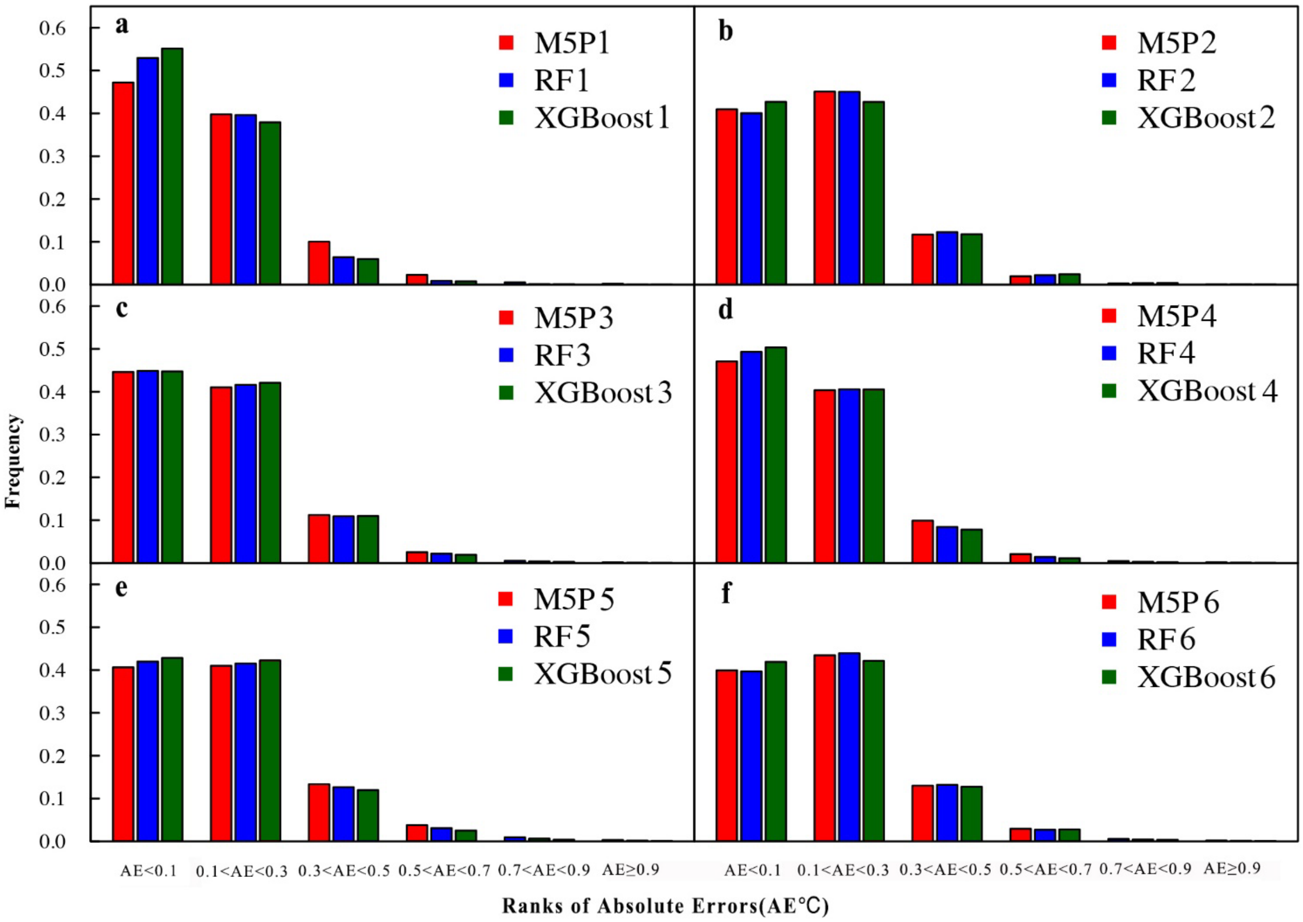

Figure 10. Frequency distribution of absolute errors (AE) between measured and estimated daily T_s obtained by three machine learning models under various input combinations (**a**–**f**) during testing.

Taking all the previous results together, it can be concluded that, among the different climate zones of China in this study, the best-performed T_s estimations were observed in the CC region, followed by the SC region. In contrast, models constructed in the study generally showed poor performance in the NEC and IM regions. Models with the combination of T_{mean} and R_s as inputs could archive relatively high performance in estimation, approaching the performance of models in the fully parametric input pattern. Moreover, the former required fewer datasets and was more generalizable. In terms of different types of models, the XGBoost model would outperform the other two models. Moreover, soil organic matter content, soil water content, T_{mean}, RH, R_s, and U_2 were significantly related to the accuracy of the model estimation of T_s and were the main factors that caused errors in the model estimation of T_s.

4. Conclusions

Soil temperature (T_s) has an important influence on the energy conversion between the atmosphere and the soil. In this study, three data-driven algorithms (i.e., M5P, RF, and XGBoost) were selected to model T_s at a 0 cm depth using different combinations of meteorological factors (including T_{max}, T_{min}, T_{mean}, RH, R_s, and U_2) as inputs. The three types of tree-based models were evaluated and compared at 689 meteorological stations in seven different climate zones of China. The results showed that the XGBoost model performed better and was more computationally efficient for estimating T_s compared to the M5P and RF models in different climate zones. With the inputs of T_{mean} and R_s, the models exhibited similar performance to that of models with the complete combination of meteorological factors as inputs, both of which had good estimation performance, but the former required fewer meteorological factors and datasets. On the basis of T_{mean}, adding R_s to the model input had more significant influence and importance on model estimating accuracy than adding other climatic factors such as RH and U_2 to the input. There was diversity in the estimating performance of the models across climate zones for the same combination of inputs. Comparing different climatic zones, the lower variability of temperature and radiation in the CC and SC regions resulted in a better model estimating accuracy, while the larger variability of radiation in the IM and NEC regions led to poorer model performance for estimating T_s. Soil organic matter content, soil water content, T_{mean}, RH, R_s, and U_2 were significantly related to model accuracy for estimating T_s, suggesting that these factors all contributed to model errors during the estimation of T_s. Overall, for estimating T_s at the 0 cm depth, the XGBoost model is highly recommended, and if the study area covers different climatic zones of China, the XGBoost model with the combination of T_{mean} and R_s as inputs would obtain satisfactory results for T_s estimation across different regions. However, it is to be noted that there is variability in T_s at different soil depths, and other meteorological factors (e.g., precipitation) would also impact T_s. Therefore, for further studies of T_s estimation, the use of different soil depths as well as additional meteorological and environmental factors should be considered.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/app12105088/s1. Figure S1. Performance of data-driven models (M5P2: a, b; RF2: c, d; XGBoost2: e, f) for the estimation of T_s with T_{mean} as inputs. Figure S2. Performance of data-driven models (M5P3: a, b; RF3: c, d; XGBoost3: e, f) for the estimation of T_s with T_{mean} and RH as inputs. Figure S3. Performance of data-driven models (M5P4: a, b; RF4: c, d; XGBoost4: e, f) for the estimation of T_s with T_{mean} and R_s as inputs. Figure S4. Performance of data-driven models (M5P5: a, b; RF5: c, d; XGBoost5: e, f) for the estimation of T_s with T_{max} and T_{min} as inputs. Figure S5. Performance of data-driven models (M5P6: a, b; RF6: c, d; XGBoost6: e, f) for the estimation of T_s with T_{mean} and U_2 as inputs. Figure S6. Scatter plot of estimated T_s versus measured T_s from the three machine learning models with T_{max} and T_{min} as inputs. Figure S7. Scatter plot of estimated T_s versus measured T_s from the three machine learning models with T_{mean} and U_2 as inputs. Figure S8. Combined indicator values for each model at all stations in the IM region. Figure S9. Combined indicator values for each model at all stations in the NEC region. Figure S10. Combined indicator values for each model at all stations in the NC region. Figure S11. Combined indicator values for each model at all stations in the CC region. Figure S12. Combined indicator values for each model at all stations in the SC region. Figure S13. Combined indicator values for each model at all stations in the QTP region.

Author Contributions: Conceptualization, J.D. and L.W.; methodology, J.D. and L.W.; software, L.W.; validation, G.H., J.D. and L.W.; formal analysis, J.D. and L.W.; investigation, F.L., S.L., and Y.C.; data curation, J.D. and L.W.; writing—original draft preparation, J.D.; writing—review and editing, L.W. and G.H.; visualization, Y.W., M.L., J.W. and S.W.; funding acquisition, L.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China, grant number 51709143, and Jiangxi Natural Science Foundation of China (No. 20181BBG78078 and 20212BDH80016).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The meteorological station data for this study were obtained from the China Meteorological Data Sharing Network (http://www.cma.cn, accessed on 1 December 2020).

Acknowledgments: Thanks to the National Meteorological Information Center of China Meteorological Administration for offering the meteorological data.

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

References

1. Hu, G.; Zhao, L.; Wu, X.; Li, R.; Wu, T.; Xie, C.; Qiao, Y.; Shi, J.; Cheng, G. An analytical model for estimating soil temperature profiles on the Qinghai-Tibet Plateau of China. *J. Arid Land* **2016**, *8*, 232–240. [CrossRef]
2. Sanikhani, H.; Deo, R.C.; Yaseen, Z.M.; Eray, O.; Kisi, O. Non-tuned data intelligent model for soil temperature estimation: A new approach. *Geoderma* **2018**, *330*, 52–64. [CrossRef]
3. Lizcano-Toledo, R.; Reyes-Martín, M.P.; Celi, L.; Fernández-Ondoño, E. Phosphorus dynamics in the Soil–Plant–Environment relationship in cropping systems: A review. *Appl. Sci.* **2021**, *11*, 11133. [CrossRef]
4. Liu, Z.; Li, D.; Zhang, J.; Saleem, M.; Zhang, Y.; Ma, R.; He, Y.; Yang, J.; Xiang, H.; Wei, H. Effect of simulated acid rain on soil CO_2, CH_4 and N_2O emissions and microbial communities in an agricultural soil. *Geoderma* **2020**, *366*, 114222. [CrossRef]
5. Barros, N. Thermodynamics of soil microbial metabolism: Applications and functions. *Appl. Sci.* **2021**, *11*, 4962. [CrossRef]
6. Kim, S.; Singh, V.P. Modeling daily soil temperature using data-driven models and spatial distribution. *Theor. Appl. Climatol.* **2014**, *118*, 465–479. [CrossRef]
7. Karnieli, A.; Agam, N.; Pinker, R.T.; Anderson, M.; Imhoff, M.L.; Gutman, G.G.; Panov, N.; Goldberg, A. Use of NDVI and land surface temperature for drought assessment: Merits and limitations. *J. Clim.* **2010**, *23*, 618–633. [CrossRef]
8. Jackson, T.; Mansfield, K.; Saafi, M.; Colman, T.; Romine, P. Measuring soil temperature and moisture using wireless MEMS sensors. *Measurement* **2008**, *41*, 381–390. [CrossRef]
9. Samadianfard, S.; Asadi, E.; Jarhan, S.; Kazemi, H.; Kheshtgar, S.; Kisi, O.; Sajjadi, S.; Manaf, A.A. Wavelet neural networks and gene expression programming models to predict short-term soil temperature at different depths. *Soil Tillage Res.* **2018**, *175*, 37–50. [CrossRef]
10. Porada, P.; Ekici, A.; Beer, C. Effects of bryophyte and lichen cover on permafrost soil temperature at large scale. *Cryosphere* **2016**, *10*, 2291–2315. [CrossRef]
11. Hu, Q.; Feng, S. A daily soil temperature dataset and soil temperature climatology of the contiguous United States. *J. Appl. Meteorol.* **2003**, *42*, 1139–1156. [CrossRef]
12. Ozturk, M.; Salman, O.; Koc, M. Artificial neural network model for estimating the soil temperature. *Can. J. Soil Sci.* **2011**, *91*, 551–562. [CrossRef]
13. Yue, J.; Chen, X.; Zhang, W. Research on an evaluation method for the Job-Housing Spaces of megacities using different scales based on multisource data integration: A case study from shenzhen. *IOP Conf. Ser. Earth Environ. Sci. IOP Publ.* **2019**, *264*, 012016. [CrossRef]
14. Changnon, S.A. A rare long record of deep soil temperatures defines temporal temperature changes and an urban heat island. *Clim. Chang.* **1999**, *42*, 531–538. [CrossRef]
15. DeBoer, T.A. Relationships between the Newhall Simulation Model and Dryland Corn Yield in the Major Land Resource Areas of Nebraska. Master's Thesis, University of Nebraska at Omaha, Omaha, NE, USA, 2007.
16. Watson, C.L. Seasonal soil temperature regimes in south-eastern Australia. *Soil Res.* **1980**, *18*, 325–331. [CrossRef]
17. Wu, W.; Tang, X.; Ma, X.; Liu, H. A comparison of spatial interpolation methods for soil temperature over a complex topographical region. *Theor. Appl. Climatol.* **2016**, *125*, 657–667. [CrossRef]
18. Yang, Y.; Wu, Z.; He, H.; Du, H.; Wang, L.; Guo, X.; Zhao, W. Differences of the changes in soil temperature of cold and mid-temperate zones, Northeast China. *Theor. Appl. Climatol.* **2018**, *134*, 633–643. [CrossRef]
19. Cheon, J.; Ham, B.; Lee, J.; Park, Y.; Lee, K. Soil temperatures in four metropolitan cities of Korea from 1960 to 2010: Implications for climate change and urban heat. *Environ. Earth Sci.* **2014**, *71*, 5215–5230. [CrossRef]
20. Abdul-Wahab, S.A.; Bakheit, C.S.; Al-Alawi, S.M. Principal component and multiple regression analysis in modelling of ground-level ozone and factors affecting its concentrations. *Environ. Modell. Softw.* **2005**, *20*, 1263–1271. [CrossRef]
21. Bilgili, M. Prediction of soil temperature using regression and artificial neural network models. *Meteorol. Atmos. Phys.* **2010**, *110*, 59–70. [CrossRef]
22. An, K.; Wang, W.; Wang, Z.; Zhao, Y.; Yang, Z.; Chen, L.; Zhang, Z.; Duan, L. Estimation of ground heat flux from soil temperature over a bare soil. *Theor. Appl. Climatol.* **2017**, *129*, 913–922. [CrossRef]
23. Fan, J.; Wang, X.; Wu, L.; Zhou, H.; Zhang, F.; Yu, X.; Lu, X.; Xiang, Y. 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. *Energ. Convers. Manag.* **2018**, *164*, 102–111. [CrossRef]
24. Zhou, B.; Erell, E.; Hough, I.; Rosenblatt, J.; Just, A.C.; Novack, V.; Kloog, I. Estimating near-surface air temperature across Israel using a machine learning based hybrid approach. *Int. J. Climatol.* **2020**, *40*, 6106–6121. [CrossRef]
25. Dong, J.; Wu, L.; Liu, X.; Li, Z.; Gao, Y.; Zhang, Y.; Yang, Q. Estimation of daily dew point temperature by using bat algorithm optimization based extreme learning machine. *Appl. Therm. Eng.* **2020**, *165*, 114569. [CrossRef]
26. Adnan, R.M.; Petroselli, A.; Heddam, S.; Santos, C.A.G.; Kisi, O. Short term rainfall-runoff modelling using several machine learning methods and a conceptual event-based model. *Stoch. Env. Res. Risk A* **2021**, *35*, 597–616. [CrossRef]
27. Fan, J.; Wu, L.; Zhang, F.; Cai, H.; Zeng, W.; Wang, X.; Zou, H. Empirical and machine learning models for predicting daily global solar radiation from sunshine duration: A review and case study in China. *Renew. Sust. Energ. Rev.* **2019**, *100*, 186–212. [CrossRef]
28. Dong, J.; Wu, L.; Liu, X.; Fan, C.; Leng, M.; Yang, Q. Simulation of daily diffuse solar radiation based on three machine learning models. *Comput. Modeling Eng. Sci.* **2020**, *123*, 49–73. [CrossRef]

29. Wu, L.; Peng, Y.; Fan, J.; Wang, Y.; Huang, G. A novel kernel extreme learning machine model coupled with K-means clustering and firefly algorithm for estimating monthly reference evapotranspiration in parallel computation. *Agr. Water Manag.* **2021**, *245*, 106624. [CrossRef]
30. Yan, S.; Wu, L.; Fan, J.; Zhang, F.; Zou, Y.; Wu, Y. A novel hybrid WOA-XGB model for estimating daily reference evapotranspiration using local and external meteorological data: Applications in arid and humid regions of China. *Agr. Water Manag.* **2021**, *244*, 106594. [CrossRef]
31. Yang, C.; Prasher, S.O.; Mehuys, G.R. An artificial neural network to estimate soil temperature. *Can. J. Soil Sci.* **1997**, *77*, 421–429. [CrossRef]
32. Mihalakakou, G. On estimating soil surface temperature profiles. *Energ. Build.* **2002**, *34*, 251–259. [CrossRef]
33. Bilgili, M. The use of artificial neural networks for forecasting the monthly mean soil temperatures in Adana, Turkey. *Turk. J. Agric. For.* **2011**, *35*, 83–93. [CrossRef]
34. Tabari, H.; Sabziparvar, A.; Ahmadi, M. Comparison of artificial neural network and multivariate linear regression methods for estimation of daily soil temperature in an arid region. *Meteorol. Atmos. Phys.* **2011**, *110*, 135–142. [CrossRef]
35. Mehdizadeh, S.; Behmanesh, J.; Khalili, K. Evaluating the performance of artificial intelligence methods for estimation of monthly mean soil temperature without using meteorological data. *Environ. Earth Sci.* **2017**, *76*, 325. [CrossRef]
36. Bilgili, M.; Sahin, B.; Sangun, L. Estimating soil temperature using neighboring station data via multi-nonlinear regression and artificial neural network models. *Environ. Monit. Assess.* **2012**, *185*, 347–358. [CrossRef]
37. Feng, Y.; Cui, N.; Hao, W.; Gao, L.; Gong, D. Estimation of soil temperature from meteorological data using different machine learning models. *Geoderma* **2019**, *338*, 67–77. [CrossRef]
38. Sihag, P.; Esmaeilbeiki, F.; Singh, B.; Pandhiani, S.M. Model-based soil temperature estimation using climatic parameters: The case of Azerbaijan Province, Iran. *Geol. Ecol. Landsc.* **2020**, *4*, 203–215. [CrossRef]
39. Kisi, O.; Tombul, M.; Kermani, M.Z. Modeling soil temperatures at different depths by using three different neural computing techniques. *Theor. Appl. Climatol.* **2015**, *121*, 377–387. [CrossRef]
40. Mehdizadeh, S.; Behmanesh, J.; Khalili, K. Comprehensive modeling of monthly mean soil temperature using multivariate adaptive regression splines and support vector machine. *Theor. Appl. Climatol.* **2018**, *133*, 911–924. [CrossRef]
41. Qi, J.; Li, S.; Li, Q.; Xing, Z.; Bourque, C.P.; Meng, F. A new soil-temperature module for SWAT application in regions with seasonal snow cover. *J. Hydrol.* **2016**, *538*, 863–877. [CrossRef]
42. Li, Z.; Yang, J.Y.; Drury, C.F.; Yang, X.M.; Reynolds, W.D.; Li, X.; Hu, C. Evaluation of the DNDC model for simulating soil temperature, moisture and respiration from monoculture and rotational corn, soybean and winter wheat in Canada. *Ecol. Model.* **2017**, *360*, 230–243. [CrossRef]
43. Lu, X.; Fan, J.; Wu, L.; Dong, J. Forecasting multi-step ahead monthly reference evapotranspiration using hybrid extreme gradient boosting with grey wolf optimization algorithm. *Comput. Model. Eng. Sci.* **2020**, *125*, 699–723. [CrossRef]
44. Jin, Q.; Fan, X.; Liu, J.; Xue, Z.; Jian, H. Estimating tropical cyclone intensity in the South China Sea using the XGBoost Model and FengYun Satellite images. *Atmosphere* **2020**, *11*, 423. [CrossRef]
45. Dong, J.; Zeng, W.; Lei, G.; Wu, L.; Chen, H.; Wu, J.; Huang, J.; Gaiser, T.; Srivastava, A.K. Simulation of dew point temperature in different time scales based on grasshopper algorithm optimized extreme gradient boosting. *J. Hydrol.* **2022**, 127452. [CrossRef]
46. Goydaragh, M.G.; Taghizadeh-Mehrjardi, R.; Jafarzadeh, A.A.; Triantafilis, J.; Lado, M. Using environmental variables and Fourier Transform Infrared Spectroscopy to predict soil organic carbon. *Catena* **2021**, *202*, 105280. [CrossRef]
47. Fan, J.; Zheng, J.; Wu, L.; Zhang, F. Estimation of daily maize transpiration using support vector machines, extreme gradient boosting, artificial and deep neural networks models. *Agr. Water Manag.* **2021**, *245*, 106547. [CrossRef]
48. Zhao, S. A new scheme for comprehensive physicl regionalization in China. *Acta Geographica Sinica.* **1983**, *38*, 1–10. (In Chinese)
49. Yao, N.; Li, Y.; Lei, T.; Peng, L. Drought evolution, severity and trends in mainland China over 1961–2013. *Sci. Total Environ.* **2018**, *616*, 73–89. [CrossRef]
50. Dong, J.; Liu, X.; Huang, G.; Fan, J.; Wu, L.; Wu, J. Comparison of four bio-inspired algorithms to optimize KNEA for predicting monthly reference evapotranspiration in different climate zones of China. *Comput. Electron. Agr.* **2021**, *186*, 106211. [CrossRef]
51. Quinlan, J.R. Learning with continuous classes. In *5th Australian Joint Conference on Artificial Intelligence*; World Scientific: Singapore, 1992; pp. 343–348. [CrossRef]
52. Wang, Y.; Witten, I.H. *Induction of Model Trees for Predicting Continuous Classes*; University of Waikato: Hamilton, New Zealand, 1996; Available online: https://hdl.handle.net/10289/1183 (accessed on 20 March 2022).
53. Yi, H.; Lee, B.; Park, S.; Kwak, K.; An, K. Prediction of short-term algal bloom using the M5P model-tree and extreme learning machine. *Environ. Eng. Res.* **2019**, *24*, 404–411. [CrossRef]
54. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
55. Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794. [CrossRef]
56. Dong, H.; Xu, X.; Wang, L.; Pu, F. Gaofen-3 PolSAR image classification via XGBoost and polarimetric spatial information. *Sensors* **2018**, *18*, 611. [CrossRef] [PubMed]
57. Tziachris, P.; Aschonitis, V.; Chatzistathis, T.; Papadopoulou, M. Assessment of spatial hybrid methods for predicting soil organic matter using DEM derivatives and soil parameters. *Catena* **2019**, *174*, 206–216. [CrossRef]

58. Malone, B.P.; Styc, Q.; Minasny, B.; McBratney, A.B. Digital soil mapping of soil carbon at the farm scale: A spatial downscaling approach in consideration of measured and uncertain data. *Geoderma* **2017**, *290*, 91–99. [CrossRef]
59. Talaee, P.H. Daily soil temperature modeling using neuro-fuzzy approach. *Theor. Appl. Climatol.* **2014**, *118*, 481–489. [CrossRef]
60. Feng, Y.; Jia, Y.; Zhang, Q.; Gong, D.; Cui, N. National-scale assessment of pan evaporation models across different climatic zones of China. *J. Hydrol.* **2018**, *564*, 314–328. [CrossRef]
61. Wang, L.; Kisi, O.; Zounemat-Kermani, M.; Li, H. Pan evaporation modeling using six different heuristic computing methods in different climates of China. *J. Hydrol.* **2017**, *544*, 407–427. [CrossRef]
62. Hara, M. New estimation trial for the soil temperature of Morioka from the meteorological data of Morioka. In *New Trends on System Sciences and Engineering*; IOS Press: Amsterdam, The Netherlands, 2015; pp. 595–608.
63. Fitton, E.M.; Brooks, C.F. Soil temperatures in the United States. *Mon. Weather Rev.* **1931**, *59*, 6–16. [CrossRef]
64. Bai, Y.; Scott, T.A.; Min, Q. Climate change implications of soil temperature in the Mojave Desert, USA. *Front. Earth Sci.-PRC.* **2014**, *8*, 302–308. [CrossRef]
65. Davies, A.; Thomas, H. Rates of leaf and tiller production in young spaced perennial ryegrass plants in relation to soil temperature and solar radiation. *Ann. Bot.* **1983**, *51*, 591–597. [CrossRef]
66. Nahvi, B.; Habibi, J.; Mohammadi, K.; Shamshirband, S.; Al Razgan, O.S. Using self-adaptive evolutionary algorithm to improve the performance of an extreme learning machine for estimating soil temperature. *Comput. Electron. Agr.* **2016**, *124*, 150–160. [CrossRef]
67. Huang, R.; Huang, J.; Zhang, C.; Wen, Z.; Chen, Y.; Zhu, D.; Qingling, W.U.; Mansaray, L.R. Soil temperature estimation at different depths, using remotely-sensed data. *J. Integr. Agr.* **2020**, *19*, 277–290. [CrossRef]
68. Kassaye, K.T.; Boulange, J.; Saito, H.; Watanabe, H. Soil water content and soil temperature modeling in a vadose zone of Andosol under temperate monsoon climate. *Geoderma* **2021**, *384*, 114797. [CrossRef]
69. Zhang, H.; Shi, X.; Yu, D.; Wang, H.; Zhao, Y.; Sun, W.; Huang, B. Spatial prediction of soil temperature in China. *Acta Pedol. Sin.* **2009**, *46*, 3718–3723. (In Chinese) [CrossRef]
70. Knight, J.H.; Minasny, B.; McBratney, A.B.; Koen, T.B.; Murphy, B.W. Soil temperature increase in eastern Australia for the past 50 years. *Geoderma* **2018**, *313*, 241–249. [CrossRef]
71. Hu, G.; Zhao, L.; Li, R.; Wu, X.; Wu, T.; Xie, C.; Zhu, X.; Su, Y. Variations in soil temperature from 1980 to 2015 in permafrost regions on the Qinghai-Tibetan Plateau based on observed and reanalysis products. *Geoderma* **2019**, *337*, 893–905. [CrossRef]
72. Kunkel, V.; Wells, T.; Hancock, G.R. Soil temperature dynamics at the catchment scale. *Geoderma* **2016**, *273*, 32–44. [CrossRef]
73. Zhang, Y.; Hou, W.; Chi, M.; Sun, Y.; An, J.; Yu, N.; Zou, H. Simulating the effects of soil temperature and soil moisture on CO_2 and CH_4 emissions in rice straw-enriched paddy soil. *Catena* **2020**, *194*, 104677. [CrossRef]
74. Zhang, L.L.; Zhao, L.; Li, R.; Gao, L.M.; Xiao, Y.; Qiao, Y.P.; Shi, J.Z. Investigating the influence of soil moisture on albedo and soil ther modynamic parameters during the warm season in Tanggula Range, Tibetan Plateau. *J. Glaciol. Geocryol.* **2016**, *38*, 351–358. (In Chinese)
75. Nikolaev, I.V.; Leong, W.H.; Rosen, M.A. Experimental investigation of soil thermal conductivity over a wide temperature range. *Int. J. Thermophys.* **2013**, *34*, 1110–1129. [CrossRef]
76. Hurrass, J.; Schaumann, G.E. Influence of the sample history and the moisture status on the thermal behavior of soil organic matter. *Geochim. Cosmochim. Ac.* **2007**, *71*, 691–702. [CrossRef]
77. Wang, Y.; Lu, Y.; Horton, R.; Ren, T. Specific heat capacity of soil solids: Influences of clay content, organic matter, and tightly bound water. *Soil Sci. Soc. Am. J.* **2019**, *83*, 1062–1066. [CrossRef]
78. Curtin, D.; Beare, M.H.; Hernandez-Ramirez, G. Temperature and moisture effects on microbial biomass and soil organic matter mineralization. *Soil Sci. Soc. Am. J.* **2012**, *76*, 2055–2067. [CrossRef]
79. Grunwald, D.; Kaiser, M.; Junker, S.; Marhan, S.; Piepho, H.; Poll, C.; Bamminger, C.; Ludwig, B. Influence of elevated soil temperature and biochar application on organic matter associated with aggregate-size and density fractions in an arable soil. *Agric. Ecosyst. Environ.* **2017**, *241*, 79–87. [CrossRef]
80. Yuan, Q. Prediction for the effect of temperature and water content on the soil specific heat by BP neural network. *Trans. Chin. Soc. Agric. Mach.* **2008**, *5*, 108–111. (In Chinese) [CrossRef]
81. Epron, D.; Farque, L.; Lucot, É.; Badot, P. Soil CO_2 efflux in a beech forest: Dependence on soil temperature and soil water content. *Ann. Forest Sci.* **1999**, *56*, 221–226. [CrossRef]
82. Gaumont-Guay, D.; Black, T.A.; Griffis, T.J.; Barr, A.G.; Jassal, R.S.; Nesic, Z. Interpreting the dependence of soil respiration on soil temperature and water content in a boreal aspen stand. *Agr. For. Meteorol.* **2006**, *140*, 220–235. [CrossRef]
83. Sawan, Z.M. Climatic variables: Evaporation, sunshine, relative humidity, soil and air temperature and its adverse effects on cotton production. *Inf. Process Agric.* **2018**, *5*, 134–148. [CrossRef]
84. Olchev, A.; Radler, K.; Sogachev, A.; Panferov, O.; Gravenhorst, G. Application of a three-dimensional model for assessing effects of small clear-cuttings on radiation and soil temperature. *Ecol. Model.* **2009**, *220*, 3046–3056. [CrossRef]
85. Kong, K.; Nandintsetseg, B.; Shinoda, M.; Ishizuka, M.; Kurosaki, Y.; Bat-Oyun, T.; Gantsetseg, B. Seasonal variations in threshold wind speed for saltation depending on soil temperature and vegetation: A case study in the Gobi Desert. *Aeolian Res.* **2021**, *52*, 100716. [CrossRef]
86. Zhang, T. Influence of the seasonal snow cover on the ground thermal regime: An overview. *Rev. Geophys.* **2005**, *43*, 1–23. [CrossRef]

87. Huang, G.; Wu, L.; Ma, X.; Zhang, W.; Fan, J.; Yu, X.; Zeng, W.; Zhou, H. Evaluation of CatBoost method for prediction of reference evapotranspiration in humid regions. *J. Hydrol.* **2019**, *574*, 1029–1041. [CrossRef]
88. Nanda, A.; Sen, S.; Sharma, A.N.; Sudheer, K.P. Soil temperature dynamics at hillslope scale—Field observation and machine learning-based approach. *Water* **2020**, *12*, 713. [CrossRef]