agriculture

MDPI

Article

Spatial Prediction and Mapping of Soil Water Content by TPE-GBDT Model in Chinese Coastal Delta Farmland with Sentinel-2 Remote Sensing Data

Dexi Zhan, Yongqi Mu, Wenxu Duan, Mingzhu Ye, Yingqiang Song *, Zhenqi Song, Kaizhong Yao, Dengkuo Sun and Ziqi Ding

School of Civil Engineering and Geomatics, Shandong University of Technology, Zibo 255000, China
* Correspondence: songyingqiang@sdut.edu.cn

Abstract: Soil water content is an important indicator used to maintain the ecological balance of farmland. The efficient spatial prediction of soil water content is crucial for ensuring crop growth and food production. To this end, 104 farmland soil samples were collected in the Yellow River Delta (YRD) in China, and the soil water content was determined using the drying method. A gradient boosting decision tree (GBDT) model based on a tree-structured Parzen estimator (TPE) hyperparametric optimization was developed, and then the soil water content was predicted and mapped based on the soil texture and vegetation index from Sentinel-2 remote sensing images. The results of statistical analysis showed that the soil water content had a high coefficient of variation (55.30%), a non-normal distribution, and complex spatial variability. Compared with other models, the TPE-GBDT model had the highest prediction accuracy (RMSE = 6.02% and R^2 = 0.71), and its mapping results showed that the areas with high soil water content were distributed on both sides of the river and near the estuary. Furthermore, the results of Shapley additive explanation (SHAP) analysis showed that the soil texture (PC2 and PC5), modified normalized difference vegetation index (MNDVI), and Sentinel-2 red edge position (S2REP) index provided important contributions to the spatial prediction of soil water content. We found that the hydraulic physical properties of soil texture and the vegetation characteristics (such as vegetation coverage, root action, and transpiration) are the key factors affecting the spatial migration and heterogeneity of the soil water content in the study area. The above results show that the TPE algorithm can quickly capture the hyperparameters that are most suitable for the GBDT model, so that the GBDT model can ensure prediction accuracy, reduce the loss function with less training data, and accurately learn of the nonlinear relationship between soil water content and environmental factors. This paper proposes a machine learning method for hyperparameter optimization that shows considerable potential to predict the spatial heterogeneity of soil water content, which can effectively support regional farmland soil and water conservation and high-quality agricultural development.

Keywords: machine learning; GBDT; hyperparameter; soil water content; Sentinel-2; farmland

Citation: Zhan, D.; Mu, Y.; Duan, W.; Ye, M.; Song, Y.; Song, Z.; Yao, K.; Sun, D.; Ding, Z. Spatial Prediction and Mapping of Soil Water Content by TPE-GBDT Model in Chinese Coastal Delta Farmland with Sentinel-2 Remote Sensing Data. *Agriculture* **2023**, *13*, 1088. https://doi.org/10.3390/agriculture13051088

Academic Editors: Jeffrey Walker, Carsten Montzka and Liujun Zhu

Received: 30 March 2023
Revised: 13 May 2023
Accepted: 17 May 2023
Published: 19 May 2023

1. Introduction

Soil water content is one of the key indicators affecting crop growth and plays an important role in the protection of regional farmland ecosystems [1]. At present, many related studies on the prediction of soil water content have been conducted [2–4], and the accurate prediction of soil water content can provide strong support for smart agriculture, ecological risk prevention and control, etc. [1]. Due to it being environmentally friendly and low-cost, the spatial prediction of soil water content can help with preventing soil erosion and nonpoint pollution on intensive farmland through the use of auxiliary variables. However, complex relationships (i.e., high spatial variability [5–7] and non-linear effects) exist between soil water content and environmental factors such as soil physical and chemical properties, climate,

vegetation, and spectrum [8,9]. Therefore, increasing the spatial prediction accuracy of soil water content at the regional scale is vital to ensure the rational use of water and soil resources and the sustainable development of high-quality agriculture.

Many methods have been applied for the spatial prediction of soil water content [10–12]. Filgueiras et al. [12] used linear models, such as multiple linear regression (MLR) and partial least squares (PLS), to predict the spatial variability of soil water content by combining the vegetation index, daily irrigation amount, crop coefficient, extraterrestrial solar radiation, and other environmental factors. Linear models are easy to use and are particularly useful for the spatial prediction of soil water content using large data volumes. When the linear relationship between environmental factors and soil water content is significant, the prediction accuracy of the linear regression model is higher. However, linear models have poor generalization performance and are generally used for feature variable screening [13]. The relationship between environmental factors and soil water content is complex, and the predictive accuracy of linear models tends to be poor.

With the development of computer science, machine learning models are being widely used in the prediction of soil water content; the commonly used models include support vector machines (SVM), random forest (RF), and eXtreme gradient boosting (XGBoost). The advantage of machine learning models is that they have a higher prediction accuracy than linear methods. For example, Fuentes et al. [14] used MLP models to predict surface soil water content using multiple-source environmental factors such as the soil moisture active and passive (SMAP) remote sensing data, moderate resolution imaging spectroradiometer (MODIS) surface reflectivity, surface temperature and land cover, and gridded soil properties. They found that the NASA-USDA SMAP is important for land surface prediction, and the importance of soil properties and LULC increased with increasing depth. Chaudhary et al. [15] collected a large amount of soil water content data from the Varanasi and Guntur districts of India, which they combined with Sentinel-1 ground distance detection (GRD) data for preprocessing. They then generated the backscatter coefficient of radiation topography correction and used the backscatter coefficients as predictors. They proved that prediction accuracy of the RF model is high for soil water content.

Recently, advanced ensemble learning boosting algorithms such as gradient boosting machine (GBM) and XGBoost have improved performance in the spatial prediction of soil water content. For example, Babaeian et al. [16] selected environmental factors such as near-infrared transformed reflectance (NTR), normalized difference vegetation index (NDVI), texture (Text), organic matter content (OC), soil porosity, etc. GBM and other models were used to predict soil moisture space in the root zone, and they found that when NDVI and NTR were used as model inputs together with soil physical and hydraulic information, the machine learning model accurately captured the changes in field-scale soil water content. This type of model explains the nonlinear relationship between environmental factors and soil water content, which effectively increases the prediction accuracy of spatial soil water content. In addition, it effectively controls the overfitting problem of the predicted model. For example, the XGBoost model adds regularization terms to the definition of its loss function and prevents overfitting by column sampling [17]. However, boosting algorithms also suffer from one critical shortcoming: due to the complexity of the algorithm, many hyperparameters need to be adjusted, and the hyperparameters also affect each other. Achieving high-quality prediction performance using default parameters and manual parameter adjustment is difficult because the optimal hyperparameters cannot be found.

The hyperparameter settings strongly impact the prediction accuracy of the machine learning model [18], and the hyperparameters cannot be obtained during training. When the learning rate and the number of base estimators of the machine learning model are low, the model does not learn enough about the features correlated to soil water content, and the high learning rate can cause overfitting and increase the loss value. Therefore, automatic parameter tuning is one of the most important machine learning model processes [19]. The Sentinel-2 satellite captures high-resolution remote sensing images, is equipped with multispectral instruments (MSI), and has a shorter revisit cycle, which is beneficial for soil

detection [20]. Sentinel-2 remote sensing images have been widely used for the spatial prediction of soil properties [21,22]. However, whether the optimal hyperparameters obtained by the optimization algorithm can reduce the loss value of the prediction model and control overfitting, as well as its response mechanism in the spatial prediction of soil water content, is unclear. Therefore, the objectives of this study were to (1) build TPE machine learning models (i.e., TPE-AdaBoost, TPE-RF, and TPE-GBDT models) with soil texture and vegetation index factors from Sentinel-2 remote sensing images for soil water content; (2) spatially map the soil water content in the study area with TPE machine learning models; (3) explore the capability of environmental factors to drive soil water content using the Shapley additive explanation (SHAP) method.

2. Materials and Methods

2.1. Study Area

The study area is located in the Yellow River Delta of Shandong Province, China (Figure 1) (118°13′–118°58′ E, 37°20′–38°02′ N), which is influenced by Eurasia and the western Pacific Ocean and has a continental monsoon climate. According to the 2022 Dongying Statistical Yearbook, the annual average temperature, annual average precipitation, and annual sunshine hours in the study area are 14.4 °C, 902.8 mm, and 2289.8 h, respectively. The geological background is alluvial loess parent material, and the soils are alluvial. The main crop types in the study area are wheat, maize, and rice. In addition, soil salinization on farmland is a serious problem in the study area, and freshwater resources are not abundant, which seriously restricts the development of high-quality agriculture. As such, the spatial prediction of soil water content is crucial for cultivated land conservation and crop production.

2.2. Soil Sampling and Laboratory Analysis

In September 2022, 104 samples from surface farmland soils (0–10 cm) were collected. The terrain of the study area is gentle, and the influence of terrain on soil water content is relatively small. We tried to avoid the impact of climate conditions on soil water content. We observed that during the entire sampling period, there were no rainfall or extreme weather conditions. The temperature, wind speed, and other conditions in the study area were in a very stable state. In addition, during this period, the farmland is in the crop (corn) harvest season, and this process will not carry out artificial irrigation activities. Moreover, due to the coverage of corn plants, the water evapotranspiration caused by solar radiation and wind speed in the soil surface is better avoided.

For soil sampling, we used the cutting ring method, in which a ring knife is used to collect soil. First, we used a shovel to smooth the soil surface and pressed the ring blade vertically downward into the soil. Second, after removing the ring knife, the soil adhered to the outer wall of the ring knife was quickly scraped, and then the soil surface was flattened from the edge to the middle with a soil cutter. Finally, the soil in the ring knife was quickly loaded into a covered aluminum box to measure the initial weight. The collected soil samples were dried in an oven and weighed every 24 h until the aluminum box was a constant weight. The detailed sampling criteria is based on Soil testing Part III: Determination of soil mechanical composition (NY/T 1121.3-2006) (https://www.moa.gov.cn, accessed on 18 January 2022), which was published by the Ministry of Agriculture and Rural Affairs of the People's Republic of China. According to formula (1), soil water content ($Soil_W$) was calculated based on the initial aluminum box with wet soil weight (m_1), the aluminum box with dry soil weight (m_2), and the aluminum box weight (m_3).

$$Soil_W = (m_1 - m_2)/(m_2 - m_3) \times 100 \quad (1)$$

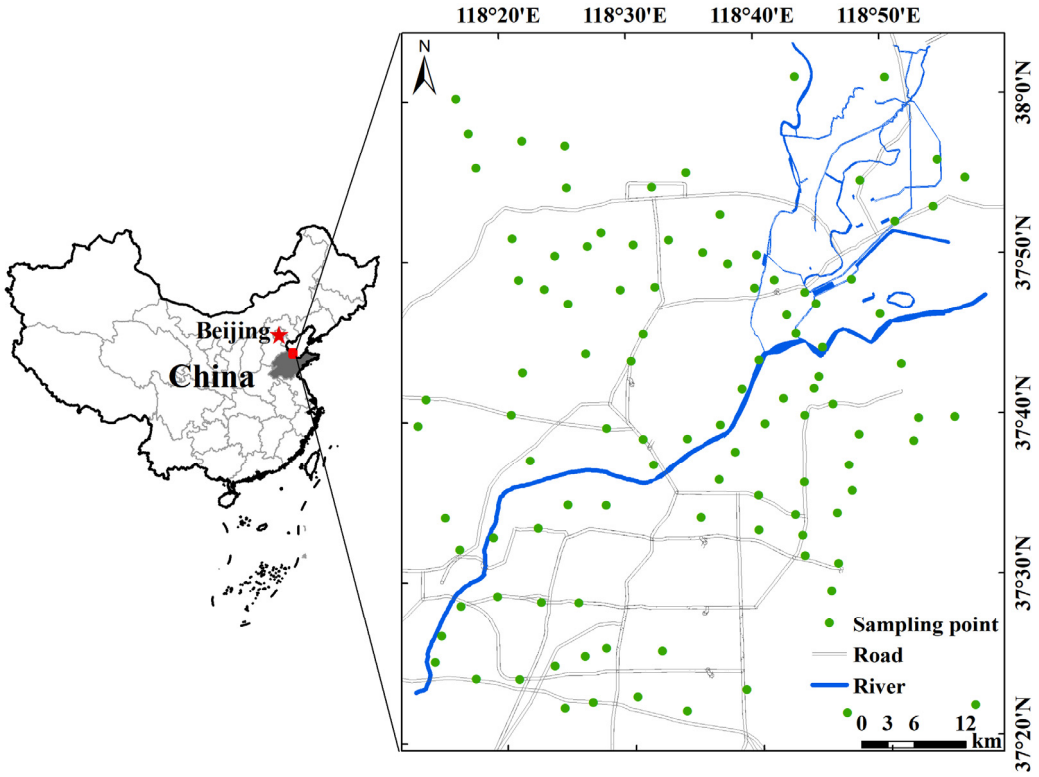

Figure 1. Information on the geographic location and distribution of samples in the study area. (Projection coordinate system: WGS_1984_UTM_Zone_50N, Geographical coordinate system: GCS_ WGS_1984).

2.3. Auxiliary Data

The auxiliary data included soil texture and the vegetation index, which were calculated based on a Sentinel-2 remote sensing image (23 September 2022), which is from the European Space Agency (https://scihub.copernicus.eu/, accessed on 12 November 2022). Sentinel-2's MSI has 13 spectral bands covering the visible, near-infrared, and short-wave near-infrared spectral bands, with wavelengths ranging from 443 to 2190 nm. The bands (bands 1–11 (B1–B11)) of Sentinel 2 image data were preprocessed (radiometric correction and geometric correction) in ENVI 5.3 and obtained ten vegetation indices, including the MERIS terrestrial chlorophyll index (MTCI), modified normalized difference vegetation index (MNDVI), modified chlorophyll absorption in reflectance index (MCARI), plant senescence reflectance index (PSRI), red edge normalized difference vegetation index (RENDVI), sentinel 2 red edge position index (S2REP), inverted red-edge chlorophyll index (IRECI), optimized soil regulated vegetation index (OSAVI), SAVIred, and modified simple ratio index (MSR). The formulas for calculating the vegetation index factors are shown in Table 1.

Table 1. The formula for vegetation index based on Sentinel-2 remote sensing image.

Index	Formula	Reference
MTCI	$(B6 - B5)/(B5 - B4)$	[23]
MNDVI	$(B7 - B4)/(B7 + B4)$	[23]
MCARI	$((B6 - B5) - 0.2 \times (B6 - B3)) \times (B6/B5)$	[24]
PSRI	$(B4 - B3)/B6$	[25]
RENDVI	$(B8 - B5)/(B8 + B5)$	[26]
S2REP	$705 + 35 \times ((B7 + B4)/2 - B5)/(B6 - B5)$	[23]
IRECI	$(B7 - B4)/(B5/B6)$	[23]
OSAVI	$(1 + 0.16) \times (B6 - B5)/(B6 + B5 + 0.16)$	[27]
SAVIred	$(B8 - B5) \times (1 + 0.5)/(B8 + B5 - 0.5)$	[28]
MSR	$(B6/B5 - 1)/\sqrt{B6/B5 + 1}$	[29]

Based on the 11 original bands of the Sentinel-2 remote sensing image, for each band, we calculated eight texture feature statistics in the Filter/ Co-current Measure analysis of ENVI 5.3 software, including correlation (CORR) (B1–B11), contrast (CTRA) (B12–B22), dissimilarity (DIS) (B23–B33), entropy (ENT) (B34–B44), homogeneity (HOM) (B45–B55), mean (B56–B66), second moment (SECM) (B67–B77), and variance (VAR) (B78–B88), and 88 soil texture feature bands were obtained. To reduce the dimensions of the soil texture feature variables, the forward PCA rotation new statistics and rotation tool was used to perform principal component analysis (PCA) on the 88 soil texture feature bands. The cumulative variance contribution rate of the first five principal components (PCs) to the soil texture features was 75.52%, and PC1 contained the most soil texture feature information (variance = 38.84%) (Figure 2a). Among the five principal components, the bands with high loading values in PC1-PC3 were mainly concentrated in those experiencing CTRA (B12–B21), DIS (B23–B33), ENT (B34–B42), and VAR (B78–B86) transformation. In addition, PC2 has high band load values after mean transformation (B56–B60). The bands with high load values in PC4 and PC5 were the bands after CORR(B3–B11) and MEAN (B60–B66) transformation, respectively (Figure 2b).

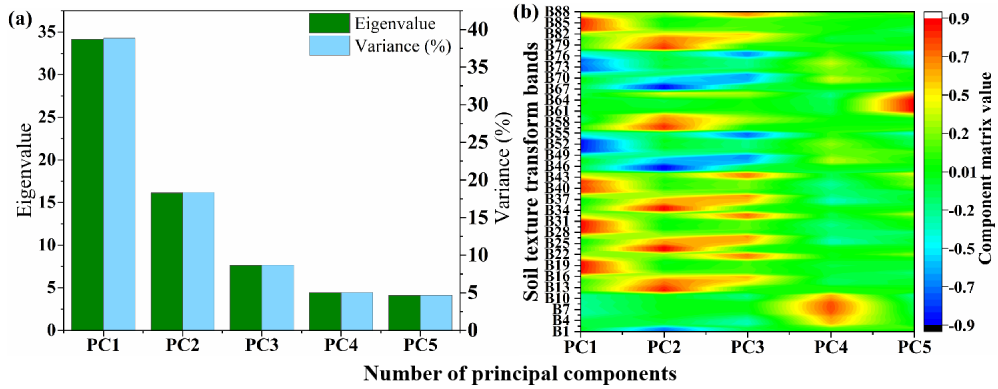

Figure 2. Principal component analysis results (**a**) eigenvalue and variance and (**b**) factor loading matrix based on Sentinel-2 remote sensing image.

2.4. Estimation Methods

2.4.1. MLR

The multiple linear regression (MLR) model uses linear functions to fit the relationship between one or more predictors and dependent variables, using the least squares method to reduce the residuals [30]. In this study, the MLR model was used to fit the linear relationship due to the need to explore the relationship between multiple features and soil water content, and its mathematical model is shown in formula (2) [31].

$$y_{sm} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \varepsilon \quad (2)$$

y_{sm} is a predicted soil water content for a sample, β_0 is the intercept, n is the number of features in the sample; so, x_1,x_2, ... , x_n are features in the sample, $\beta_1, \beta_2, \ldots, \beta_n$ are the coefficients calculated in the model, ε is a distractor.

The linear model is very simple, easy to understand, and the learning process is very fast. It can complete the modeling of large data sets in a very short time [32].

2.4.2. AdaBoost

Adaptive boosting (AdaBoost) was first proposed by Yoav Freund and Robert Schapire in 1996 [33]. This study mainly uses this model to complete the regression prediction task. The core of AdaBoost regression is to find the weight of weak learner and update the weight of training samples. In the iterative process, the prediction error training data is given more weight, and the weight of the training data is updated whenever a new weak learner is

created [34]. When calculating the weight of the learner, the maximum error on the training set and the relative error of each sample are first calculated [33],

$$E_k = \max|y_i - G_k(x_i)|i = 1, 2, \ldots, m, \tag{3}$$

where m is the number of samples and it is the i-th sample, x_i is the feature of the $i-$th sample, y_i is the true value of soil water content in the i-th sample, G_k is the k-th weak learner, $G_k(x_i)$ is the predicted value of the weak learner based on the sample, E_k is the maximum error of the k-th weak learner found in the formula. The relative error is then calculated,

$$e_{ki} = \frac{(y_i - G_k(x_i))^2}{E_k^2}, \tag{4}$$

the error rate of the weak evaluator is calculated by relative error,

$$e_k = \sum_{i=1}^{m} w_{ki} e_{ki}, \tag{5}$$

where w_{ki} is the weight distribution of the current training data.

According to the error rate to calculate its weight, the small error rate of the learner has a better prediction effect, so it has a higher weight,

$$a_k = \frac{e_k}{1 - e_k}, \tag{6}$$

the final strong learner is the sum of all weak learners multiplied by the corresponding weights:

$$f_{Ada}(x) = \sum_{k=1}^{K} (\ln \frac{1}{a_k}) G(x) \tag{7}$$

In the formula (7), K is the total number of weak learners, that is, the hyperparameter n_estimators, $G(x)$ is the median of all $a_k G_k(x), k = 1, 2, \ldots, K$, and $f_{Ada}(x)$ is the final strong learner. It can be seen that it is a kind of boosting algorithm, which makes good use of all weak learners. However, AdaBoost model is sensitive to outlier data, and outlier data may obtain higher weights in the modeling process, which affects the final prediction performance [35–37]. The hyperparameter information of AdaBoost model and the final hyperparameters we use are shown in Table 2.

Table 2. Information about main parameters of AdaBoost model.

Hyperparameters	Type	Default Value	Range	Explanation	Final Hyperparameters
n_estimators	int	50	[10, 500]	number of trees	58
learning_rate	float	1.0	[0.01, 3]	learning speed in the iterative process	0.973
loss	string	"linear"	["linear", "square", "exponential"]	type of loss function	"linear"

2.4.3. RF

The RF model is a machine learning algorithm first proposed by Leo Breiman and Adele Cutler [38]. It uses the decision tree as a weak learner, randomly samples the training data, establishes multiple different decision trees. In order to improve the generalization performance of the model, each decision tree is as different as possible. Therefore, the RF model trains different training sets by random sampling with playback. About two-thirds of the samples are randomly selected (in-bag data) for training models, and the rest is used for testing [39,40]. Then, the model randomly selects a part of the sample that features as the node splitting attribute of the decision tree, repeats the above steps, establishes many decision trees, and forms a forest. According to the bagging idea, we take the classification results with the most decision trees to complete the classification task and calculate the average of all learner results to complete the regression task [38,41].

Compared with other machine learning models, RF can calculate the importance of features by the error contribution of features to the model [42], which makes it easier to explain the model. Because of the characteristics of the RF bagging method, it can be used conveniently without dividing training and test data, so it has been widely used in various fields. The detailed hyperparameters of the RF model are shown in Table 3.

Table 3. Information about main parameters of RF model.

Hyperparameters	Type	Default Value	Range	Explanation	Final Hyperparameters
n_estimators	int	100	[10, 500]	number of trees	94
max_depth	int	1.0	[1, 10]	the maximum depth of each tree	6
min_impurity_decrease	float	0	[0, 5]	minimum impurity of node partition	0.8
min_samples_split	int	2	[1, 15]	the minimum number of samples required for internal node re-division	2
max_features	int, float, string	1.0	["log2", "sqrt", 2, 4, 16, 32, None]	the number of features to consider when finding the best segmentation	2
criterion	string	"squared_error"	["mae", "mse", "squared_error"]	node division standard	"squared_error"

2.4.4. GBDT

The gradient boosting decision tree (GBDT) is an ensemble learning model based on the boosting idea; it applies the CART regression tree as a weak learner, which uses the gradient descent algorithm to continuously fit the loss function by constructing multiple weak learners to obtain the optimal model. Different from RF, a strong dependence exists between the individual learners in GBDT. GBDT first initializes the weak learner [43],

$$f_0(x) = \arg\min_c \sum_{i=1}^{m} L(y_i, c), \tag{8}$$

where $L(y_i, c)$ is the loss function, c is a constant, and the sum of the difference between the constant and the true value of the data set is the smallest.

Then, each weak learner adjusts the training samples according to the performance of the previous weak learner; that is, the residual caused by training the previous weak learner [44–46],

$$r_{ki} = y_i - f_{k-1}(x_i), i = 1, 2 \ldots m, \tag{9}$$

where f_{k-1} is the k-th weak learner. Then, fitting residual r_{ki}, create a new weak learner $T_k(x)$, and update the learner,

$$f_k(x) = f_{k-1}(x) + T_k(x) \tag{10}$$

Repeat until the number of weak learners reaches the previously set hyperparameter n_estimators. Finally, the prediction results of all weak learners are weighted and summed to form a strong learner and output the prediction results,

$$f_{GBDT}(x) = f_0(x) + \sum_{k=1}^{K} T_k(x), \tag{11}$$

where $f_{GBDT}(x)$ is the strongest learner that is finally obtained. The GBDT model involves many hyperparameters. The hyperparameter details and the final hyperparameters are shown in Table 4.

Table 4. Information about main parameters of GBDT model.

Hyperparameters	Type	Default Value	Range	Explanation	Final Hyperparameters
criterion	string	friedman_mse	[“friedman_mse”, “squared_error”]	specify the evaluation criteria for partitioning subtrees.	“squared_error”
loss	string	squared_error	[“squared_error”, ”absolute_error”, “huber”, “quantile”]	specify the loss function type	“squared_error”
n_estimators	int	100	[10, 500]	specifies the number of iterations, the number of trees	97
learning_rate	float	0.1	[0.01, 3]	specify the learning speed of the model	0.5
subsample	float	1.0	[0.3, 1]	specify the proportion of subsampling in the modeling process	0.66
max_depth	int	3	[1, 20]	specify the maximum depth of each tree	2
max_features	int, float, string	None	[“log2”, “sqrt”, 2, 4, 16, 32, None]	limit the number of features considered when branching the tree	4
min_impurity_decrease	float	0	[0, 5]	limit the amount of information gain when splitting nodes	0.15

2.4.5. TPE Optimization Algorithm

At present, the common parameter adjustment methods include Bayesian optimization, grid search, random search, artificial search, etc. However, some optimization algorithms are time-consuming and do not obtain ideal results. One of the best current parameter optimization methods, the tree structure Parzen estimator method (TPE) algorithm, was used in this study. The TPE optimization algorithm is an algorithm that uses the Gaussian mixture model to learn the hyperparameters model. Compared with many other optimization algorithms, the TPE optimization algorithm can often more efficiently achieve better results. Therefore, it is widely used in the field of machine learning parameter tuning and is one of the best hyperparameter optimization algorithms. Compared with the grid and random searches, the TPE algorithm has strong global exploration ability, does not easily fall into the local optimum, and considerably increases the optimization efficiency [47].

TPE algorithm defines two probability density functions [48]:

$$p(x|y) = \begin{cases} l(x) \, if \, y < z \\ g(x) \, if \, y \geq z \end{cases}, \tag{12}$$

where z is the value at quantile r in the y-set, and the value of r was 0.15 in the original TPE study [49].

$$r = p(y < z) \tag{13}$$

The evaluation criterion selected by the TPE optimization algorithm is expected improvement (EI), which is the expectation that $f(x)$ is less than a certain threshold z, and its calculation formula is:

$$EI_{y^*}(x) = \frac{ry^* l(x) - l(x)\int_{-\infty}^{y^*} p(y)dy}{rl(x) - (1-r)g(x)} \propto \left(r + \frac{g(x)}{l(x)}(1-r)\right)^{-1}, \tag{14}$$

Formula (14) shows that EI has an inverse relationship with $g(x)$ and a positive relationship with $l(x)$. To maximize the EI in the iterative process, the probability of $l(x)$ is increased and the probability of $g(x)$ is decreased, so $\frac{g(x)}{l(x)}$ is used to select the most appropriate x value. In each iteration, the algorithm returns x^* with the maximum EI value, and the returned x^* participates in the next iteration, which is repeated to find the best hyperparameter combination for the model [19].

2.5. Validation Methods

In this study, the data set was randomly divided: 75% into the training set and 25% into test set. The input features of the model in this study were 10 vegetation indices and 5 soil textures, and the prediction target was soil water content. Then, the model was optimized with the TPE optimization algorithm, and soil water content was predicted according to the test set after training. To evaluate the soil water content predictive performance of the model, the coefficient of determination (R^2) and the root mean square error (RMSE) were used as the criteria, which are calculated as follows [50,51]:

$$R^2 = \frac{\left[\sum_{i=1}^{m}(y_i - y_i)(\hat{y}_i - \hat{y}_i)\right]^2}{\sum_{i=1}^{m}(y_i - y_i)^2 \sum_{i=1}^{m}(\hat{y}_i - \hat{y}_i)^2}, \quad (15)$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(\hat{y}_i - y_i)^2}, \quad (16)$$

where $\hat{y}_i$ is the predicted value of the model; y_i and $\hat{y}_i$ are the average of the true value and predicted value, respectively.

To make the model used in our study more intuitive, the Shapley additive explanation (SHAP) package was imported in the Python development environment to explain the contribution of features to the fitting of farmland soil water content during the model prediction process and to visualize the influence of different features. SHAP was proposed by Lundberg and Lee [52]; it uses the Shapley value in game theory to explain the fitting process of the machine learning model. By calculating the Shapley value of each feature, the impact of the feature on the predicted value is measured [47] and visualized so that the impact of each feature on the target can be intuitively observed. SHAP can be used for global or local interpretation. We mainly used it for global interpretation to study the dominant factors affecting the soil water content in the Yellow River Delta.

3. Results

3.1. Statistical Analysis

Figure 3a shows the coefficient of variation (CV) of the soil water content and environmental factors, which were high, with moderate variability (MV) (15% < CV < 35%) and high variability (HV) (CV > 35%) [53]. The CV of the soil water content exceeded 50% and showed high spatial variability. Factors similarly affecting soil water content variability included IRECI, PC1, PC4, and PC5, all of which had a CV of around 50%. Among the 15 environmental factors used in the study, PC3 was the only low variability (LV) factor (CV = 14.9%). MNDVI, MACRI, PSRI, RENDVI, S2REP, OSAVI, MSR, and PC2 had similar spatial variability, with moderate variability. The CV of MTCI, IRECI, SAVIred, PC1, PC4, and PC5 exceeded 35%, showing high spatial variability, of which SAVIred variability was particularly prominent (CV > 160%).

The numerical distribution of the soil water content and environmental factors after normalization is shown in Figure 3b, including the distribution trend and normal distribution curve of each feature factor. The values of MTCI, SAVIred, and PC3 were concentrated below 0.2, which may have been influenced by some extreme values, resulting in a lower overall normal distribution curve. In particular, SAVIred is the surface soil information calculated by remote sensing image bands. Because the sampling time was autumn, the crops on the surface had not been harvested, so little bare soil information was extracted, which may also have led to a higher CV. The other environmental factors showed a normal distribution that was close to the standard. The environmental factors that had curves similar in shape to the normal distribution curve of the soil water content values were MSR, PC2, and PC5. The overall data in PC3 were large, concentrated in the value range above 0.6.

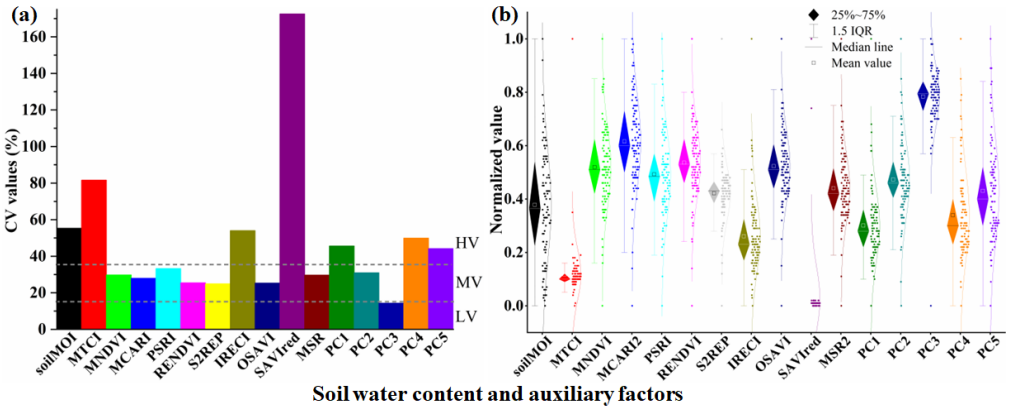

Figure 3. Soil water content and environmental factors (**a**) coefficient of variation and (**b**) statistical analysis.

3.2. Prediction of Soil Water Content using TPE Machine Learning Models

According to the performance of the four models on the test set, the results in Table 5 were calculated. In Table 5, the explanatory ability (R^2) of the four models for the spatial variability of soil water content exceed 0.4. Compared with the other three models, the TPE-GBDT model was the most accurate (RMSE = 6.02% and R^2 = 0.71) for the prediction of soil water content. This showed that the TPE-GBDT model has a strong ability to fit the nonlinear relationship between soil water content and environmental factors. The prediction accuracy of TPE-RF was second only to that of TPE-GBDT (RMSE = 6.94% and R^2 = 0.50); in addition, the predictive accuracy of the MLR was the lowest (RMSE = 7.92% and R^2 = 0.44). Figure 4 shows a scatterplot of the four prediction models regarding the predictions and real soil water content. Figure 4 shows that the distribution of the MLR model is relatively scattered, and the model prediction ability is weaker than that of the other models. The distribution of some test points in the TPE-AdaBoost model is more discrete than that of the training set, and outliers are present.

Table 5. The prediction results (validation set) of the machine learning model for the soil water content.

Model	RMSE (%)	R^2
MLR	7.92	0.44
GBDT	6.02	0.71
RF	6.94	0.50
AdaBoost	7.08	0.45

The scatter distribution of the TPE-GBDT model is more compact and evenly distributed near the 1:1 diagonal (Figure 4). By analyzing the scatterplots of the four models, we found that the relationship between soil water content and environmental factors, as captured by the GBDT model, was more comprehensive and accurate than that of other models. To explore the hyperparameter optimization process of the TPE algorithm, each iteration result in TPE optimization was saved to reproduce the algorithm flow. In Figure 5, the hyperparameter samples selected by TPE are basically distributed in the most densely sampled areas (i.e., subsample (0.66–0.72), n_estimators (95–100), etc.), and the possibility of a low RMSE is high in the densely sampled areas, which verifies the principle of the TPE algorithm. TPE always selected the next sampling point based on the previously sampled sample to approximate the optimal parameter. Therefore, in Figure 5, some regional sampling points are sparse (i.e., subsample (0.75–0.84), and n_estimators (80–85), learning_rate (0.4–0.45), etc.), and some regional sampling points are dense. After multiple iterations, the optimal hyperparameters searched by the TPE must exist in the region with the densest sampling points or near the region.

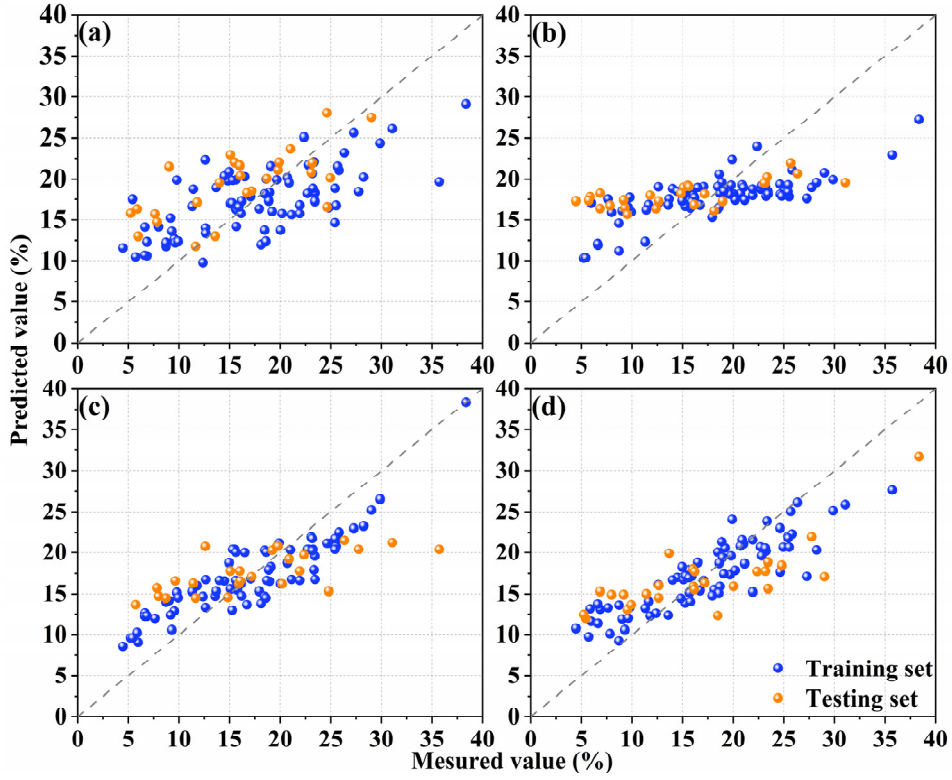

Figure 4. Scatterplot of soil water content measured value vs predicted value using (**a**) MLR, (**b**) TPE-RF, (**c**) TPE-AdaBoost, and (**d**) TPE-GBDT models.

3.3. Spatial Mapping of Soil Water Content Using TPE Machine Learning Models

The spatial distribution of the soil water content predicted by the models is shown in Figure 6. The ranges of the soil water content of MLR, TPE-RF, TPE-AdaBoost, and TPE-GBDT were 0.25–126.15%, 14–28.46%, 9.07–26.99%, and 3.20–39.83%, respectively. Compared with the true range of soil water content (4.47–38.37%), the mapping range of TPE-GBDT was closest. Although the prediction ranges of TPE-RF and TPE-AdaBoost were also within the range of the training data, the predictions of the minimum and maximum values were not accurate, being far less than those of the GBDT model. As for MLR, the prediction accuracy was poor due to the limitations of the algorithm, and the reference value of its mapping results was low. The map of TPE-GBDT (Figure 6a) shows an appropriate trend in the transition from high to low soil water content, and no abnormal situation arose due to a cliff-like decline in soil water content. Additionally, the spatial change was relatively smooth and natural, which is in line with the natural variation in farmland soil water content in the study area. However, the color change in the TPE-RF and TPE-AdaBoost mapping results was too abrupt, and the sense of boundary was strong, showing that the methods did not capture the trend in the transition of soil water content. Based on the spatial distribution of soil water content produced by the TPE-GBDT model, the soil water content in the study area from light blue to red represents a gradual increasing trend. The soil water content of the farmland in the southwest of the study area was low, and the areas with high soil water content appeared in the form of scattered spots along the Yellow River. The northeastern part of the study area is close to the estuary of Yellow River, where the number of tributaries increases, so the soil water content substantially increases as well.

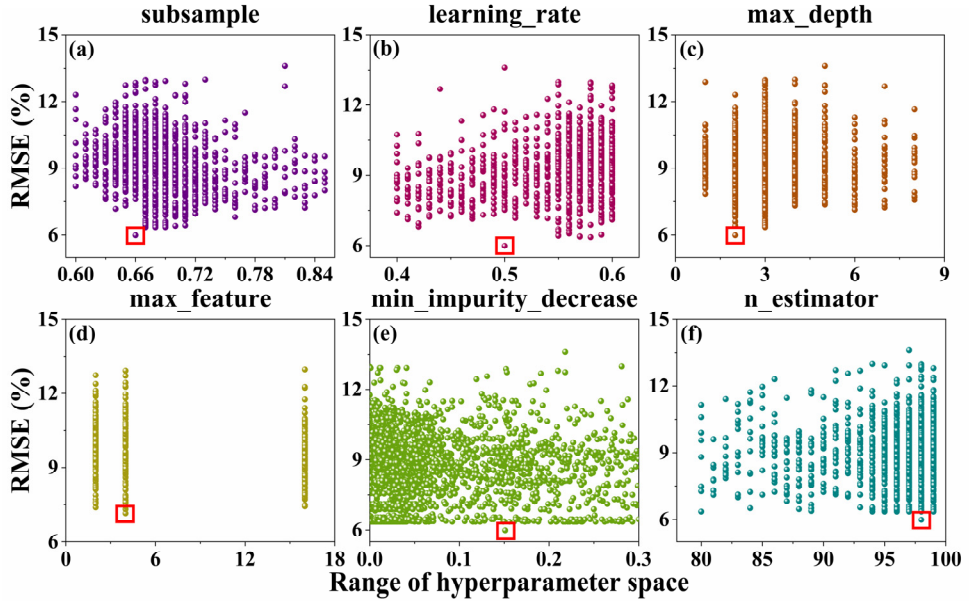

Figure 5. TPE optimization process of GBDT model hyperparameters (including (**a**) subsample, (**b**) learning_rate, (**c**) max_depth, (**d**) max_feature, (**e**) min_impurity_decrease, (**f**) n_estimators). The horizontal axis is the sampling value of the hyperparameter, and the vertical axis is the RMSE obtained by the hyperparameter. The points marked in red are the optimal hyperparameters obtained after multiple iterations.

3.4. Importance of Environmental Factors Using SHAP Analysis

As shown in Figure 7a, the five environmental factors that contributed the most to soil water content prediction were PC2 (1.7), PC5 (1.3), S2REP (1.2), PC4 (1.12), and MNDVI (0.89). The red points of PC2, which had the highest contribution, are all distributed on the positive side of the SHAP value, and the blue-purple points are mostly distributed on the negative side, indicating that PC2 was positively correlated with soil water content (Figure 7b). Similarly, PC4 and MNDVI also showed a positive correlation with soil water content. Regarding the environmental factors (i.e., PC5, OSAVI, IRECI, RENDVI, PSRI, etc.), as the eigenvalue increases, the target value decreases, indicating that these features negatively impact soil water content. In particular, although the SHAP value of the S2REP factor is higher, both the high and low eigenvalues appear on the negative influence side of the SHAP value; whether it is positive or negative is unclear. Mutual influences exist among the various factors. In some samples, the effect of other factors on soil water content may be stronger than that of S2REP, so the correlation of S2REP is not sufficiently clear. However, the remote sensing images were obtained in autumn. Decreases in the chlorophyll contents of plants lead to the lack of effective information in the S2REP factor, which produced the ambiguity regarding positive and negative effects on soil water content.

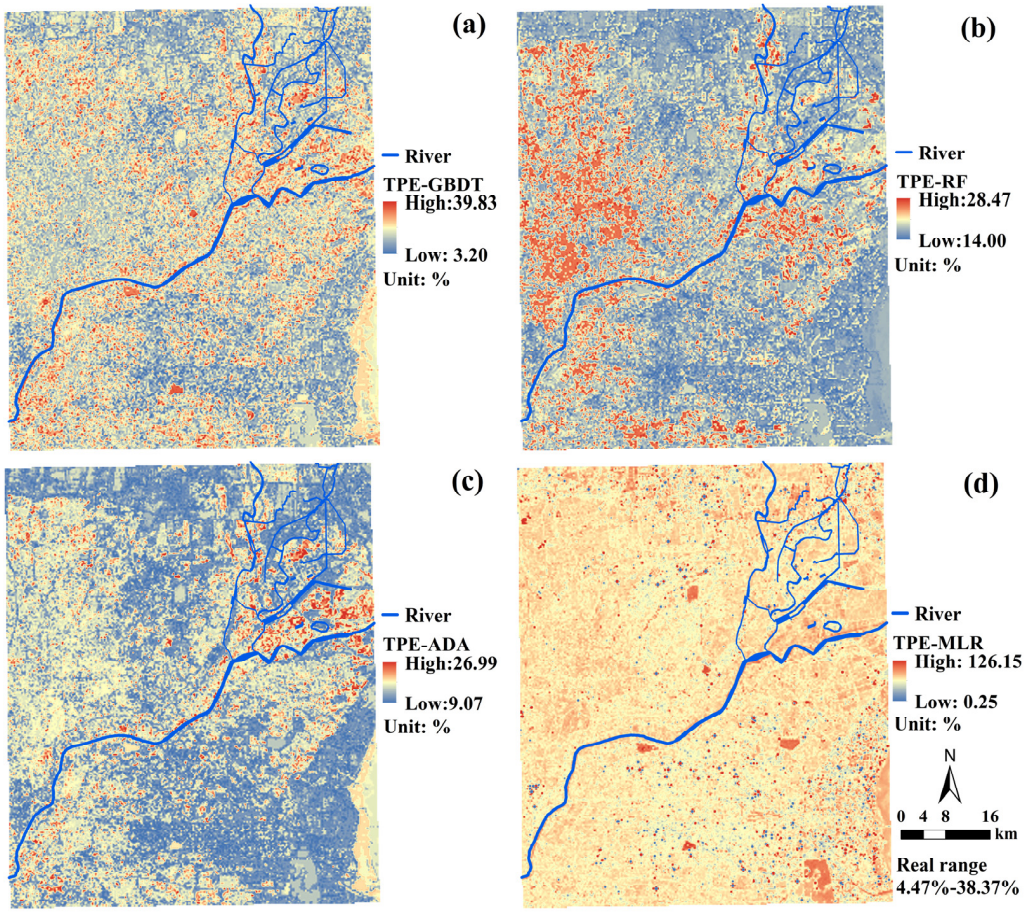

Figure 6. Spatial distribution of soil water content (0–10cm) in the study area (**a**) TPE-GBDT, (**b**) TPE-RF, (**c**) TPE-AdaBoost, and (**d**) MLR models.

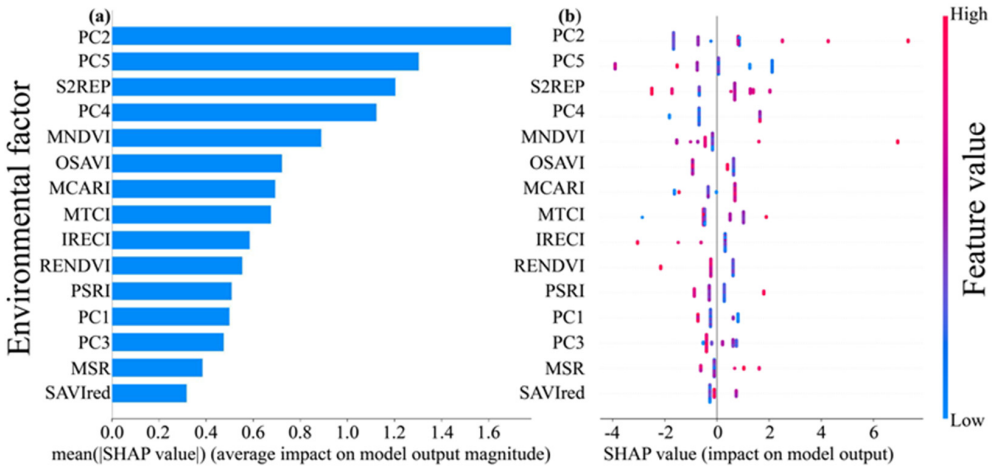

Figure 7. SHAP analysis: (**a**) the contribution ranking of 15 environmental factors of the TPE-GBDT model, and (**b**) further analyzed by the scatter plot of the sample features. The positive SHAP value represents a positive influence on the target and vice versa. When the color is close to red, the eigenvalue is higher, and when it is close to blue, the eigenvalue is lower.

4. Discussion

4.1. Advantage of Fitting Mechanism of TPE-GBDT Model

The TPE-GBDT model demonstrated an absolute advantage among the considered models in both R^2 and RMSE. The key reason for this is that the GBDT model was optimized with the independent variable–dependent variable fitting mechanism. The prediction effect of a single weak learner may not be ideal, so the GBDT model establishes multiple trees, and each tree in the model is in a series. Each time a new tree is constructed, the new tree fits the negative gradient value of the loss function [45,54] and continuously iterates, thereby continuously reducing the loss value. It simplifies the process of solving the objective function and increases the fitting accuracy of the model to the real data [44]. Different from the RF model, which is based on the bagging algorithm, all learners of the GBDT model are in series [55] and are weighted to integrate all learners. Considering the results of all learners, its prediction performance is better than that of the RF model. Compared with the AdaBoost model, which is also a boosting algorithm, the GBDT model uses gradient descent as the basis for optimizing the model rather than the weight of incorrectly divided samples, so is not easily affected by outliers [56]. A similar study has also shown that the AdaBoost model may be influenced by abnormal samples during training, which leads to a poor final prediction ability [35]. In other words, the GBDT model produces a stronger generalization performance, as demonstrated by obtaining the lowest RMSE and the highest R^2 in the prediction of soil water content.

As an advanced automatic machine learning algorithm, the GBDT model can accurately learn the relationship between features and soil water content with less training data, and it overcomes the limitations of statistical linear fitting, so it can more easily manage multicollinearity and tasks with missing values [57]. For example, Liu et al. [58] used a GBDT model to predict the distribution of air pollution in the streets of Wuhan and found that pollutants are not easily diffused when the building reaches a certain height. Zhang et al. [44] also found that the GBDT model can be used for almost all regression problems. Whether linear or nonlinear regression, it can often provide accurate predictions. In addition, the TPE optimization algorithm can maximize the mechanism advantages of the independent variable–dependent variable fitting of the GBDT model. It is similar to the manual parameter adjustment, which analyzes the previous soil water content sample point and determines the selection of the next sample point. In the process of TPE optimizing the GBDT model, the loss function gets smaller and smaller, and the hyperparameters of the model must exist in the most dense direction of the sampling points [59]. The TPE algorithm can quickly capture the most suitable hyperparameters for the GBDT model in the space of millions of parameter combinations, which ensures efficiency and reduces the loss function. Finally, the prediction performance of the GBDT model optimized by TPE provides substantial improvements compared with previous models in the spatial prediction of soil water content.

4.2. Effect of Environmental Factors in Driving Soil Water Content

Soil texture and vegetation were found to be important factors driving the soil water content in the study area. PC2, PC5, and PC4 played an important role in the spatial prediction of soil water content. PC2, PC5, and PC4 represented the soil texture information after DIS, CORR, and MEAN transformation, respectively, indicating that soil texture strongly influences soil water content [60]. The hydrophysical properties of soil (i.e., texture, structure, and particle size distribution) are essentially determined by soil formation processes [61], water retention, and water holding capacity [62]. In other words, in a certain area, the average soil moisture is determined by soil formation conditions and the environment, which can lead to a balance in the supply and use of soil moisture [63]. Zhu et al. [64] also found a clear correlation between soil water content and the soil's dielectric properties.

Vegetation indexes such as NDVI directly impact soil moisture through vegetation coverage, root function, and transpiration. First, vegetation cover can reduce the solar radiation reaching the soil surface and reduce the evaporation of soil moisture [65,66].

Additionally, vegetation can intercept rainwater, reduce the speed of surface water flow, allow more water to penetrate the soil, and increase the soil water content [67]. Moreover, the soil moisture in the whole profile is strongly affected by long-term changes in the leaf area index (LAI), whereas the surface soil moisture is only affected by the short-term change in the LAI. Therefore, the effect of seasonal variation of the LAI on soil moisture dynamics in the whole profile may be stronger, because the soil moisture dynamics in the whole profile have a long memory [68].

Second, Liu and Shao et al. found that plant roots function in soil and water conservation, where the larger the leaf area index of the plant, the more conducive to the production of litter [69]. An increase in litter can improve the soil's properties, such as increasing soil organic matter [70], which increases soil water holding capacity [71]. The roots of vegetation can also change the structural characteristics of the surrounding soil. An increase in porosity near the root encourages the stem to grow deeper and beyond the rooting zone [67]. Third, plants can change the surface soil water content and evaporation through transpiration and more effectively produce indirect feedback [72]. For example, in the case of complete vegetation coverage, approximately 64% of precipitation enters the ecosystem through evaporation, whereas on the bare soil, only approximately 36% of the precipitation evaporates [73]. Vegetation restoration also affects the soil moisture content and the formation of soil heterogeneity [74,75]. Moreover, topographic factors such as altitude and slope substantially affect the spatial variability in soil water content [76–78]. However, the altitude of our study area is close to sea level and the terrain is flat; these have little effect on the soil water content of the farmland.

4.3. Limitations and Implications

In this study, TPE-GBDT showed strong performance and good mapping results in spatial prediction of soil water content. However, there are still some limitations and implications for future research. Firstly, the increase of temperature usually significantly reduces the soil water content in farmland [79,80]. For example, the increase of temperature will affect the amount of snowfall and other factors [81], and rising temperatures drive surface evaporation and reduce soil water content [82]. Secondly, wind speed, solar radiation, and other conditions are important factors affecting the dynamic changes of soil water content [83,84]. Some studies found that the soil water content was different due to the various incident angles of the sun affected by the slope direction [85]. Inversely, some studies also found that the response of surface soil water to temperature is low under warm climate conditions, but it is affected by local factors such as land cover and soil type [80]. Moreover, at a smaller spatial scale, the variability of soil water content is mainly controlled by conditions such as topography and soil texture [86,87], while climate variables have a significant effect on soil water content in a large spatial scale and over a long period of time [88]. Thirdly, the surface soil is more affected by large-scale precipitation, irrigation, and surface runoff in comparison with the deep soil [89–91]. Some scholars have carried out studies on multi-layer soil water content, and found that 6-h cumulative rainfall is important in predicting surface soil water content, while 24-h cumulative rainfall is more important in predicting deep soil water content [92]. In addition, the relationship between surface and deep soil moisture depends on the kind of transition between soil horizons [8].

The TPE-GBDT model has great potential in the spatial prediction of soil water content. However, it is mainly suitable for stable climatic conditions in the short term. In future studies, the space-time generalization of the TPE-GBDT model on the prediction of long-term soil water content with different depths needs to be widely considered. In these scenarios, climate and soil depth conditions should be added into the training of the TPE-optimized machine learning models. When it comes to rainy and snowy seasons, considering the influence of climatic factors can theoretically improve the accuracy of models predicting soil water content. Similarly, due to the characteristics of the soil and the influence of long-term cumulative rainfall, studies also need to pay attention to the grading prediction of water content with various soil depths. Therefore, based on the

TPE-optimized machine learning models in the study, it is essential to develop a model framework of time and soil depth fusion for the improvement of performance of models predicting soil water content, which may obtain more interesting results.

5. Conclusions

In this study, we developed a machine learning model based on the TPE hyperparameter optimization algorithm, which we combined with soil texture features and vegetation indices, to predict spatial soil water content distribution in the Yellow River Delta. Based on the prediction results produced by the TPE-GBDT model, we found that the model accurately fitted the nonlinear relationship between soil water content and environmental factors (R^2 = 0.71 and RMSE = 6.02%). The TPE optimization algorithm can quickly capture the optimal hyperparameters of the GBDT model, reduce the loss function, and ensure prediction accuracy. The results of the spatial mapping of soil water content showed that the GBDT model is highly reliable. The areas with high soil water content were mainly distributed along the banks of the Yellow River and near the estuary. Furthermore, the results of SHAP analysis showed that soil texture and vegetation (vegetation coverage, root function, and transpiration) are important forces driving the migration and transformation of farmland soil water content in the study area. Because soil water content is affected by seasonal climate, a hyperparameter optimization machine learning model with time attributes should be developed in the future, and environmental factors with a finer time resolution should be incorporated to predict soil water content in both time and space.

Author Contributions: Conceptualization, Y.S.; methodology, D.Z.; software, D.Z. and Z.S.; validation, Y.M. and W.D.; formal analysis, Y.S. and Z.D.; investigation, D.Z.; resources, Y.S. and K.Y.; data curation, M.Y. and W.D.; writing—original draft, Y.S. and D.S.; writing—review and editing, D.Z. and Y.S.; visualization, D.Z.; supervision, Y.S.; project administration, Y.S.; funding acquisition, Y.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Shandong Provincial Natural Science Foundation (ZR2020QD013).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The soil testing Part III: Determination of soil mechanical composition (NY/T 1121.3-2006) comes from the Ministry of Agriculture and Rural Affairs of the People's Republic of China (https://www.moa.gov.cn, (accessed on 18 January 2022)). The Sentinel-2 remote sensing image (23 September 2022) comes from the European Space Agency (https://scihub.copernicus.eu/, (accessed on 12 November 2022)).

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

References

1. Li, Q.; Li, Z.; Shangguan, W.; Wang, X.; Li, L.; Yu, F. Improving soil moisture prediction using a novel encoder-decoder model with residual learning. *Comput. Electron. Agric.* **2022**, *195*, 106816. [CrossRef]
2. Li, Q.; Wang, Z.; Shangguan, W.; Li, L.; Yao, Y.; Yu, F. Improved daily SMAP satellite soil moisture prediction over China using deep learning model with transfer learning. *J. Hydrol.* **2021**, *600*, 126698. [CrossRef]
3. Xu, Z.; Man, X.; Duan, L.; Cai, T. Improved subsurface soil moisture prediction from surface soil moisture through the integration of the (de)coupling effect. *J. Hydrol.* **2022**, *608*, 127634. [CrossRef]
4. Wendroth, O.; Rogasik, H.; Koszinski, S.; Ritsema, C.J.; Dekker, L.W.; Nielsen, D.R. State-space prediction of field-scale soil water content time series in a sandy loam. *Soil Till Res.* **1999**, *50*, 85–93. [CrossRef]
5. Gao, X.; Wu, P.; Zhao, X.; Wang, J.; Shi, Y.; Zhang, B.; Tian, L.; Li, H. Estimation of spatial soil moisture averages in a large gully of the Loess Plateau of China through statistical and modeling solutions. *J. Hydrol.* **2013**, *486*, 466–478. [CrossRef]
6. Huang, X.; Shi, Z.H.; Zhu, H.D.; Zhang, H.Y.; Ai, L.; Yin, W. Soil moisture dynamics within soil profiles and associated environmental controls. *Catena* **2015**, *136*, S1242748033. [CrossRef]
7. Zhu, H.D.; Shi, Z.H.; Fang, N.F.; Wu, G.L.; Guo, Z.L.; Zhang, Y. Soil moisture response to environmental factors following precipitation events in a small catchment. *Catena-Giess. Amst.* **2014**, *120*, 73–80. [CrossRef]

8. Hagen, K.; Berger, A.; Gartner, K.; Geitner, C.; Niedertscheider, K. Event-based dynamics of the soil water content at Alpine sites (Tyrol, Austria). *Catena* **2020**, *194*, 104682. [CrossRef]
9. Li, X.; Zhang, S.; Peng, H.; Hu, X.; Ma, Y. Soil water and temperature dynamics in shrub-encroached grasslands and climatic implications: Results from Inner Mongolia steppe ecosystem of north China. *Agric. For. Meteorol.* **2013**, *s171–172*, 20–30. [CrossRef]
10. Pignotti, G.; Rathjens, H.; Chaubey, I.; Williams, M.; Crawford, M. Strong sensitivity of watershed-scale, ecohydrologic model predictions to soil moisture. *Environ. Modell. Softw.* **2021**, *144*, 105162. [CrossRef]
11. Rawat, K.S.; Sehgal, V.K.; Singh, S.K.; Ray, S.S. Soil moisture estimation using triangular method at higher resolution from MODIS products. *Phys. Chem. Earth Parts A/B/C* **2022**, *126*, 103051. [CrossRef]
12. Filgueiras, R.; Almeida, T.S.; Mantovani, E.C.; Dias, S.H.B.; Fernandes-Filho, E.I.; Da Cunha, F.F.; Venancio, L.P. Soil water content and actual evapotranspiration predictions using regression algorithms and remote sensing data. *Agric. Water Manag.* **2020**, *241*, 106346. [CrossRef]
13. Schönauer, M.; Prinz, R.; Väätäinen, K.; Astrup, R.; Pszenny, D.; Lindeman, H.; Jaeger, D. Spatio-temporal prediction of soil moisture using soil maps, topographic indices and SMAP retrievals. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *108*, 102730. [CrossRef]
14. Fuentes, I.; Padarian, J.; Vervoort, R.W. Towards near real-time national-scale soil water content monitoring using data fusion as a downscaling alternative. *J. Hydrol.* **2022**, *609*, 127705. [CrossRef]
15. Chaudhary, S.K.; Srivastava, P.K.; Gupta, D.K.; Kumar, P.; Prasad, R.; Pandey, D.K.; Das, A.K.; Gupta, M. Machine learning algorithms for soil moisture estimation using Sentinel-1: Model development and implementation. *Adv. Space Res.* **2022**, *69*, 1799–1812. [CrossRef]
16. Babaeian, E.; Paheding, S.; Siddique, N.; Devabhaktuni, V.K.; Tuller, M. Estimation of root zone soil moisture from ground and remotely sensed soil information with multisensor data fusion and automated machine learning. *Remote Sens. Environ.* **2021**, *260*, 112434. [CrossRef]
17. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. *ACM* **2016**, 785–794.
18. Sraitih, M.; Jabrane, Y.; Hajjam, E.L.; Hassani, A. An Automated System for ECG Arrhythmia Detection Using Machine Learning Techniques. *J. Clin. Med.* **2021**, *10*, 5450. [CrossRef]
19. Yu, J.; Zheng, W.; Xu, L.; Meng, F.; Li, J.; Zhangzhong, L. TPE-CatBoost: An adaptive model for soil moisture spatial estimation in the main maize-producing areas of China with multiple environment covariates. *J. Hydrol.* **2022**, *613*, 128465. [CrossRef]
20. He, Y.; Zhang, Z.; Xiang, R.; Ding, B.; Du, R.; Yin, H.; Chen, Y.; Ba, Y. Monitoring salinity in bare soil based on Sentinel-1/2 image fusion and machine learning. *Infrared Phys. Technol.* **2023**, *131*, 104656. [CrossRef]
21. Gomez, C.; Dharumarajan, S.; Féret, J.-B.; Lagacherie, P.; Ruiz, L.; Sekhar, M. Use of Sentinel-2 Time-Series Images for Classification and Uncertainty Analysis of Inherent Biophysical Property: Case of Soil Texture Mapping. *Remote Sens.* **2019**, *11*, 565. [CrossRef]
22. Zhou, T.; Geng, Y.; Ji, C.; Xu, X.; Wang, H.; Pan, J.; Bumberger, J.; Haase, D.; Lausch, A. Prediction of soil organic carbon and the C:N ratio on a national scale using machine learning and satellite data: A comparison between Sentinel-2, Sentinel-3 and Landsat-8 images. *Sci. Total Environ.* **2020**, *755*, 142661. [CrossRef] [PubMed]
23. Frampton, W.J.; Dash, J.; Watmough, G.; Milton, E.J. Evaluating the capabilities of Sentinel-2 for quantitative estimation of biophysical variables in vegetation. *Isprs J. Photogramm. Remote Sens.* **2013**, *82*, 83–92. [CrossRef]
24. Daughtry, C.; Walthall, C.L.; Kim, M.S.; Colstoun, E.; Iii, M.M. Estimating Corn Leaf Chlorophyll Concentration from Leaf and Canopy Reflectance. *Remote Sens. Environ.* **2000**, *74*, 229–239. [CrossRef]
25. Fernández-Manso, A.; Fernández-Manso, O.; Quintano, C. SENTINEL-2A red-edge spectral indices suitability for discriminating burn severity. *Int. J. Appl. Earth Obs. Geoinf.* **2016**, *50*, 170–175. [CrossRef]
26. Tucker, C.J. Red and photographic infrared linear combinations for monitoring vegetation. *Remote Sens. Environ.* **1979**, *8*, 127–150. [CrossRef]
27. Rondeaux, G.; Steven, M.; Baret, F. Optimization of soil-adjusted vegetation indices. *Remote Sens. Environ.* **1996**, *55*, 95–107. [CrossRef]
28. Huete, A.R. A soil-adjusted vegetation index (SAVI). *Remote Sens. Environ.* **1988**, *25*, 295–309. [CrossRef]
29. Wu, C.; Niu, Z.; Tang, Q.; Huang, W. Estimating chlorophyll content from hyperspectral vegetation indices: Modeling and validation. *Agric. For. Meteorol.* **2008**, *148*, 1230–1241. [CrossRef]
30. Wu, C.; Wu, J.; Luo, Y.; Zhang, L.; DeGloria, S.D. Spatial Prediction of Soil Organic Matter Content Using Cokriging with Remotely Sensed Data. *Soil Fertil. Plant Nutr.* **2009**, *73*, 1202–1208. [CrossRef]
31. Odhiambo, B.O.; Kenduiywo, B.K.; Were, K. Spatial prediction and mapping of soil pH across a tropical afro-montane landscape. *Appl. Geogr.* **2020**, *114*, 102129. [CrossRef]
32. Idowu, O.; Johnson, F.; Adefemi, O. Modeling cation exchange capacity and soil water holding capacity from basic soil properties. *Eurasian J. Soil Sci.* **2016**, *5*, 266–274.
33. Freund, Y. Experiment With a New Boosting Algorithm. *Morgan Kaufmann* **1996**, *96*, 148–156.
34. Liu, H.; Chen, C. Spatial air quality index prediction model based on decomposition, adaptive boosting, and three-stage feature selection: A case study in China. *J. Clean Prod.* **2020**, *265*, 121777. [CrossRef]
35. Tian, Z.; Xiao, J.; Feng, H.; Wei, Y. Credit Risk Assessment based on Gradient Boosting Decision Tree. *Procedia Comput. Sci.* **2020**, *174*, 150–160. [CrossRef]
36. Yang, J.; Zhao, C.; Yu, H.; Chen, H. Use GBDT to Predict the Stock Market. *Procedia Comput. Sci.* **2020**, *174*, 161–171. [CrossRef]

37. Sun, B.; Chen, S.; Wang, J.; Chen, H. A robust multi-class AdaBoost algorithm for mislabeled noisy data. *Knowl.-Based Syst.* **2016**, *102*, 87–102. [CrossRef]
38. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
39. Pham, D.T.; Yokoya, N.; Nguyen, T.T.T.; Le, N.N.; Ha, N.T.; Xia, J.; Takeuchi, W.; Pham, T.D. Improvement of Mangrove Soil Carbon Stocks Estimation in North Vietnam Using Sentinel-2 Data and Machine Learning Approach. *Gisci. Remote Sens.* **2021**, *58*, 68–87. [CrossRef]
40. Musial, J.P.; Bojanowski, J.S. Comparison of the Novel Probabilistic Self-Optimizing Vectorized Earth Observation Retrieval Classifier with Common Machine Learning Algorithms. *Remote Sens.* **2022**, *14*, 378. [CrossRef]
41. Cutler, D.R.; Edwards, T.C., Jr.; Beard, K.H.; Cutler, A.; Hess, K.T.; Gibson, J.; Lawler, J.J. Random forests for classification in ecology. *Ecology* **2007**, *88*, 2783–2792. [CrossRef] [PubMed]
42. Jiang, F.; Kutia, M.; Sarkissian, A.J.; Lin, H.; Wang, G. Estimating the Growing Stem Volume of Coniferous Plantations Based on Random Forest Using an Optimized Variable Selection Method. *Sensors* **2020**, *20*, 7248. [CrossRef] [PubMed]
43. Friedman, J. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [CrossRef]
44. Zhang, H.; Wu, W.; Wu, H. TOC prediction using a gradient boosting decision tree method: A case study of shale reservoirs in Qinshui Basin. *Geoenergy Sci. Eng.* **2023**, *221*, 111271. [CrossRef]
45. Wu, W.; Wang, J.; Huang, Y.; Zhao, H.; Wang, X. A novel way to determine transient heat flux based on GBDT machine learning algorithm. *Int. J. Heat Mass Transf.* **2021**, *179*, 121746. [CrossRef]
46. Zhou, J.; Huang, S.; Tao, M.; Khandelwal, M.; Dai, Y.; Zhao, M. Stability prediction of underground entry-type excavations based on particle swarm optimization and gradient boosting decision tree. *Undergr. Space.* **2023**, *9*, 234–249. [CrossRef]
47. Chen, C.; Seo, H. Prediction of rock mass class ahead of TBM excavation face by ML and DL algorithms with Bayesian TPE optimization and SHAP feature analysis. *Acta Geotech.* **2023**, *9*, 234–249. [CrossRef]
48. Ozaki, Y.; Tanigaki, Y.; Watanabe, S.; Onishi, M. Multiobjective tree-structured parzen estimator for computationally expensive optimization problems. *ACM* **2020**, *9*, 533–541.
49. Bergstra, J.; Bardenet, R.; Bengio, Y.; Kégl, B. Algorithms for Hyper-Parameter Optimization. *Adv. Neural Inf. Process. Syst.* **2011**, *24*, 2546–2554.
50. 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. *Agric. Water Manag.* **2021**, *245*, 106547. [CrossRef]
51. Fan, J.; Wu, L.; Ma, X.; Zhou, H.; Zhang, F. Hybrid support vector machines with heuristic algorithms for prediction of daily diffuse solar radiation in air-polluted regions. *Renew. Energy.* **2020**, *145*, 2034–2045. [CrossRef]
52. Lundberg, S.; Lee, S.I. A Unified Approach to Interpreting Model Predictions. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 1–10.
53. Wilding, L.P. Spatial variability: Its documentation, accomodation and implication to soil surveys. *Spat. Var.* **1985**, *1985*, 166–194.
54. Li, D.; Ma, C. Research on lane change prediction model based on GBDT. *Phys. A: Stat. Mech. Its Appl.* **2022**, *608*, 128290. [CrossRef]
55. Zhao, L.; Zhao, X.; Zhou, H.; Wang, X.; Xing, X. Prediction model for daily reference crop evapotranspiration based on hybrid algorithm and principal components analysis in Southwest China. *Comput. Electron. Agric.* **2021**, *190*, 106424. [CrossRef]
56. An, R.; Tong, Z.; Ding, Y.; Tan, B.; Wu, Z.; Xiong, Q.; Liu, Y. Examining non-linear built environment effects on injurious traffic collisions: A gradient boosting decision tree analysis. *J. Transp. Health.* **2022**, *24*, 101296. [CrossRef]
57. Xiao, L.; Lo, S.; Liu, J.; Zhou, J.; Li, Q. Nonlinear and synergistic effects of TOD on urban vibrancy: Applying local explanations for gradient boosting decision tree. *Sust. Cities Soc.* **2021**, *72*, 103063. [CrossRef]
58. Liu, M.; Chen, H.; Wei, D.; Wu, Y.; Li, C. Nonlinear relationship between urban form and street-level PM2.5 and CO based on mobile measurements and gradient boosting decision tree models. *Build. Environ.* **2021**, *205*, 108265. [CrossRef]
59. Chen, B.; Zheng, H.; Luo, G.; Chen, C.; Bao, A.; Liu, T.; Chen, X. Adaptive estimation of multi-regional soil salinization using extreme gradient boosting with Bayesian TPE optimization. *Int. J. Remote Sens.* **2022**, *43*, 778–811. [CrossRef]
60. Lin, B.B.; Egerer, M.H.; Liere, H.; Jha, S.; Philpott, S.M. Soil management is key to maintaining soil moisture in urban gardens facing changing climatic conditions. *Sci Rep.* **2018**, *8*, 17565. [CrossRef]
61. Cosby, B.J.; Hornberger, G.M.; Clapp, R.B.; Ginn, T.R. A Statistical Exploration of the Relationships of Soil Moisture Characteristics to the Physical Properties of Soils. *Water Resour. Res.* **1984**, *20*, 682–690. [CrossRef]
62. Quan, Q.; Tian, D.; Luo, Y.; Zhang, F.; Niu, S. Water scaling of ecosystem carbon cycle feedback to climate warming. *Sci. Adv.* **2019**, *5*, v1131. [CrossRef] [PubMed]
63. Zhao, W.; Yu, X.; Xu, C.; Li, S.; Wu, G.; Yuan, W. Dynamic traceability effects of soil moisture on the precipitation–vegetation association in drylands. *J. Hydrol.* **2022**, *615*, 128645. [CrossRef]
64. Zhu, Z.; Bo, Y.; Sun, T. Spatial downscaling of satellite soil moisture products based on apparent thermal inertia: Considering the effect of vegetation condition. *J. Hydrol.* **2023**, *616*, 128824. [CrossRef]
65. Qiu, L.; Zhang, X.; Cheng, J.; Yin, X. Effects of black locust (Robinia pseudoacacia) on soil properties in the loessial gully region of the Loess Plateau, China. *Plant Soil* **2010**, *332*, 207–217. [CrossRef]
66. Kou, M.; Garcia-Fayos, P.; Hu, S.; Jiao, J. The effect of Robinia pseudoacacia afforestation on soil and vegetation properties in the Loess Plateau (China): A chronosequence approach. *For. Ecol. Manag.* **2016**, *375*, 146–158. [CrossRef]
67. Metzger, J.C.; Wutzler, T.; Valle, N.D.; Filipzik, J.; Hildebrandt, A. Vegetation impacts soil water content patterns by shaping canopy water fluxes and soil properties. *Hydrol. Process.* **2017**, *31*, 3783–3795. [CrossRef]

68. Chen, M.; Willgoose, G.R.; Saco, P.M. Investigating the impact of leaf area index temporal variability on soil moisture predictions using remote sensing vegetation data. *J. Hydrol.* **2015**, *522*, 274–284. [CrossRef]
69. Liu, B.; Shao, M.A. Modeling soil–water dynamics and soil–water carrying capacity for vegetation on the Loess Plateau, China. *Agric. Water Manag.* **2015**, *159*, 176–184. [CrossRef]
70. Angst, Š.; Mueller, C.W.; Cajthaml, T.; Angst, G.; Lhotáková, Z.; Bartuška, M.; Špaldoňová, A.; Frouz, J. Stabilization of soil organic matter by earthworms is connected with physical protection rather than with chemical changes of organic matter. *Geoderma.* **2017**, *289*, 29–35. [CrossRef]
71. Li, R.; Kan, S.; Zhu, M.; Chen, J.; Ai, X.; Chen, Z.; Zhang, J.; Ai, Y. Effect of different vegetation restoration types on fundamental parameters, structural characteristics and the soil quality index of artificial soil. *Soil Tillage Res.* **2018**, *184*, 11–23. [CrossRef]
72. Liu, Z.; Notaro, M.; Gallimore, R. Indirect vegetation–soil moisture feedback with application to Holocene North Africa climate. *Glob. Change Biol.* **2010**, *16*, 1733–1743. [CrossRef]
73. Ivanov, V.Y.; Fatichi, S.; Jenerette, G.D.; Espeleta, J.F.; Troch, P.A.; Huxman, T.E. Hysteresis of soil moisture spatial heterogeneity and the "homogenizing" effect of vegetation. *Water Resour. Res.* **2015**, *46*, 1–15. [CrossRef]
74. D'Odorico, P.; Caylor, K.; And, G.; Scanlon, T.M. On soil moisture–vegetation feedbacks and their possible effects on the dynamics of dryland ecosystems. *J. Geophys. Res. Biogeosciences* **2007**, *112*, 1–10. [CrossRef]
75. Zkan, U.; Kbulak, F.G. Effect of vegetation change from forest to herbaceous vegetation cover on soil moisture and temperature regimes and soil water chemistry. *Catena* **2017**, *149*, 158–166.
76. Wang, S.; Fu, B.; Gao, G.; Zhou, J.; Jiao, L.; Liu, J. Linking the soil moisture distribution pattern to dynamic processes along slope transects in the Loess Plateau, China. *Environ. Monit. Assess.* **2015**, *187*, 778. [CrossRef] [PubMed]
77. Melliger, J.J.; Niemann, J.D. Effects of gullies on space–time patterns of soil moisture in a semiarid grassland. *J. Hydrol.* **2010**, *389*, 289–300. [CrossRef]
78. Gao, X.; Zhao, X.; Wu, P.; Brocca, L.; Zhang, B. Effects of large gullies on catchment-scale soil moisture spatial behaviors: A case study on the Loess Plateau of China. *Geoderma* **2016**, *261*, 1–10. [CrossRef]
79. Soares, P.M.M.; Lima, D.C.A. Water scarcity down to earth surface in a Mediterranean climate: The extreme future of soil moisture in Portugal. *J. Hydrol.* **2022**, *615*, 128731. [CrossRef]
80. Fan, K.; Slater, L.; Zhang, Q.; Sheffield, J.; Gentine, P.; Sun, S.; Wu, W. Climate warming accelerates surface soil moisture drying in the Yellow River Basin, China. *J. Hydrol.* **2022**, *615*, 128735. [CrossRef]
81. Cook, B.I.; Mankin, J.S.; Anchukaitis, K.J. Climate Change and Drought: From Past to Future. *Curr. Clim. Chang. Rep.* **2018**, *4*, 164–179. [CrossRef]
82. Berg, A.; Sheffield, J. Climate Change and Drought: The Soil Moisture Perspective. *Curr. Clim. Chang. Rep.* **2018**, *4*, 180–191. [CrossRef]
83. Zheng, W.; Zhangzhong, L.; Zhang, X.; Wang, C.; Zhang, S.; Sun, S.; Niu, H. *A Review on the Soil Moisture Prediction Model and Its Application in the Information System*; Springer International Publishing: Cham, Switzerland, 2019; pp. 352–364.
84. Liu, Y.; Liu, Y.; Wang, W.; Fan, X.; Cui, W. Soil moisture droughts in East Africa: Spatiotemporal patterns and climate drivers. *J. Hydrol. Reg. Stud.* **2022**, *40*, 101013. [CrossRef]
85. Famiglietti, J.S.; Rudnicki, J.W.; Rodell, M. Variability in surface moisture content along a hillslope transect: Rattlesnake Hill, Texas. *J. Hydrol.* **1998**, *210*, 259–281. [CrossRef]
86. Crow, W.T.; Berg, A.A.; Cosh, M.H.; Loew, A.; Mohanty, B.P.; Panciera, R.; Rosnay, P.D.; Ryu, D.; Walker, J.P. Upscaling sparse ground-based soil moisture observations for the validation of coarse-resolution satellite soil moisture products. *Rev. Geophys.* **2012**, *50*, 1–20. [CrossRef]
87. Gaur, N.; Mohanty, B.P. Land-surface controls on near-surface soil moisture dynamics: Traversing remote sensing footprints. *Water Resour. Res.* **2016**, *52*, 6365–6385. [CrossRef]
88. Peng, D.; Zhou, Q.; Tang, X.; Yan, W.; Chen, M. Changes in soil moisture caused solely by vegetation restoration in the karst region of southwest China. *J. Hydrol.* **2022**, *613*, 128460. [CrossRef]
89. Rockstrm, J.; Karlberg, L.; Wani, S.P.; Barron, J.; Qiang, Z. Managing Water in Rainfed Agriculture—The Need for a Paradigm Shift. *Agric. Water Manag.* **2010**, *97*, 543–550. [CrossRef]
90. Seneviratne, S.I.; Corti, T.; Davin, E.L.; Hirschi, M.; Jaeger, E.B.; Lehner, I.; Orlowsky, B.; Teuling, A.J. Investigating soil moisture–climate interactions in a changing climate: A review. *Earth Sci. Rev.* **2010**, *99*, 125–161. [CrossRef]
91. Tugwell-Wootton, T.; Skrzypek, G.; Dogramaci, S.; McCallum, J.; Grierson, P.F. Soil moisture evaporative losses in response to wet-dry cycles in a semiarid climate. *J. Hydrol.* **2020**, *590*, 125533. [CrossRef]
92. Bartels, G.K.; Castro, N.M.D.R.; Pedrollo, O.; Collares, G.L. Soil moisture estimation in two layers for a small watershed with neural network models: Assessment of the main factors that affect the results. *Catena* **2021**, *207*, 105631. [CrossRef]