ISSN: 2710-0588

Soil Moisture Prediction for Intelligent Irrigation: An XGBoost-based Model with Multi-Dimensional Feature Engineering

Jinqiao Liang*

Business School, Shandong Normal University, Jinan, 250014, China

*18565450647@163.com

Abstract

Accurate soil moisture prediction is essential for intelligent irrigation and optimal agricultural water resource allocation. To address the limitation of traditional linear models in capturing complex nonlinear relationships between meteorological factors and soil moisture, this study developed an Extreme Gradient Boosting (XGBoost)-based soil moisture prediction model using hourly meteorological and soil moisture data from a 1-hectare multi-crop farm, achieving high-precision prediction through in-depth data preprocessing, multi-dimensional feature engineering, and systematic model training and validation. Results indicate that with 44 optimal features selected, the model achieves a coefficient of determination (R^2) of 0.673 on the test set, with low mean squared error (MSE) and mean absolute error (MAE), outperforming traditional linear models significantly; feature importance analysis identifies daily average temperature, previous day's soil moisture, and total daily precipitation as key driving factors (consistent with soil water evaporation and replenishment mechanisms); and the model predicts 5 cm depth soil moisture of 0.2435 (meeting the minimum crop survival threshold) for the target date. This model provides reliable data support for dynamic decision-making in intelligent irrigation systems, with great practical value for improving agricultural water use efficiency and advancing precision agriculture.

Keywords

Soil Moisture Prediction, Intelligent Irrigation, Meteorological Factors, Gradient Boosting Decision Tree.

1. Introduction

1.1. Research Background

Global water scarcity has become a critical bottleneck restricting the sustainable development of agriculture. Statistics show that agricultural water use accounts for 70% of the world's total freshwater consumption, among which approximately 50% is wasted due to extensive irrigation decision-making. Intelligent irrigation systems collect real-time data such as soil moisture and meteorological parameters through IoT sensors and achieve precise regulation by integrating data analysis, providing an effective solution to address the problem of agricultural water waste[1].However, as the core parameter for irrigation decisions, soil moisture exhibits significant non-linearity and spatiotemporal heterogeneity in its dynamic changes, which are comprehensively influenced by meteorological conditions, soil properties, and crop water consumption. Traditional linear prediction models struggle to capture these complex relationships, resulting in insufficient prediction accuracy that fails to meet the needs of precision irrigation.

In farmland ecosystems, soil moisture at a depth of 5 cm directly affects the water absorption efficiency of crop roots and serves as a key indicator reflecting the availability of soil moisture[2]. The variation in soil moisture at this depth has a particularly close response

ISSN: 2710-0588

relationship with meteorological factors: increased temperature accelerates the evaporation of soil moisture, precipitation directly supplements soil moisture, and relative humidity and wind speed indirectly alter soil moisture by influencing the evapotranspiration process[3]. Nevertheless, most existing studies focus on the relationship between a single meteorological factor and soil moisture[4][5][6], lacking systematic consideration of the synergistic effects of multiple factors, temporal variation characteristics, and lag effects. This leads to weak model generalization ability and large prediction errors under dynamic meteorological conditions. Therefore, constructing a soil moisture prediction model that can integrate multi-dimensional meteorological information and capture complex non-linear relationships has become a key link in promoting the practical application of intelligent irrigation systems.

1.2. Significance of the Study

The academic value and practical significance of the soil moisture prediction model constructed in this study are mainly reflected in three aspects. First, from the methodological perspective, this study verifies the applicability of the XGBoost algorithm[7] in the temporal prediction of agricultural environments. By conducting in-depth feature engineering to explore the potential correlations between meteorological factors and soil moisture, it provides a reference technical framework for similar prediction problems in agricultural environments. Second, from the practical application perspective, the high-precision prediction results of the model can directly serve as a decision-making basis for intelligent irrigation systems. Farmers can adjust the timing and amount of irrigation according to the predicted soil moisture, thereby reducing water waste and lowering agricultural production costs. Third, from the perspective of environmental sustainability, precision irrigation can avoid soil salinization and groundwater pollution caused by over-irrigation, while reducing ineffective water consumption and improving the efficiency of agricultural water use, which is in line with the requirements of green and sustainable agricultural development.

2. Model Formulation

2.1. Data Source and Preprocessing

The data in this study were derived from on-site farm measurements, including hourly meteorological data such as soil moisture and precipitation, with a time span covering from May 1 to July 31, 2021. Among them, the meteorological data comprised 8 indicators in total: hourly temperature (T, °C), sea-level pressure (Po, hPa), station pressure (P, hPa), water vapor pressure (Pa, kPa), relative humidity (U, %), wind direction (DD), wind force (Ff), and precipitation (RR, mm). The soil moisture data referred to the absolute moisture at a depth of 5 cm (%), which was calculated as the percentage of the mass of water in the soil relative to the dry mass of the soil. The formula is as follows:

$$AbsoluteMoisture(\%) = (\frac{M_w}{M_d}) \bullet 100\% \tag{1}$$

where M_w represents the mass of soil water and M_d denotes the dry soil mass. The dry soil mass per unit area M_d was 1500 kg/m^3.

The data preprocessing workflow mainly included three parts: missing value handling, outlier detection, and time series reconstruction. For missing values in meteorological data and soil moisture data, linear interpolation was adopted for filling: for single-point missing in continuous time series, supplementary filling was conducted through linear fitting of data at adjacent time points before and after; for missing at multiple consecutive time points, an

ISSN: 2710-0588

interpolation method based on similar-day meteorological patterns was used to ensure the temporal continuity of the data.

Outlier detection was performed using the interquartile range (IQR) method. First, the upper quartile (Q3) and lower quartile (Q1) of each indicator were calculated. Data beyond the range of $[Q_1 - 1.5IQR, Q_3 + 1.5IQR]$ were identified as outliers and replaced with the 7-day moving average of the corresponding indicator, so as to avoid the interference of outliers on model training.

Considering that the daily-scale variation of soil moisture is the core basis for irrigation decision-making, this study reconstructed the hourly meteorological data into daily-scale data: for continuous indicators such as temperature (T), pressure (Po, P), water vapor pressure (Pa), and relative humidity (U), their daily average value, maximum value, minimum value, and intra-day range were calculated; for precipitation (RR), its daily total accumulation was calculated; for wind direction (DD) and wind force (Ff), qualitative indicators were converted into quantitative data through wind direction angle conversion and wind force grading quantification. Finally, a daily-unit sample dataset was formed, where each sample contained the daily meteorological features and the corresponding soil moisture value.

2.2. Feature Engineering

Feature engineering is a crucial step in enhancing the predictive performance of models. Based on the physical significance of meteorological data and the physiological mechanism of soil moisture variation, this study constructs four categories of feature sets and selects the optimal feature subset through feature selection.

In terms of feature construction, first, intra-day statistical features are built to reflect the overall level and fluctuation range of meteorological conditions on a given day. These features include daily average temperature, daily maximum temperature, daily minimum temperature, intra-day temperature range, daily average relative humidity, and daily total precipitation. Among them, daily total precipitation is directly related to the amount of soil moisture supplement, while daily average temperature and temperature range serve as core factors because they affect the evaporation rate of soil moisture.

Meanwhile, temporal variation features are constructed to capture the dynamic trends of meteorological conditions, such as hourly temperature change rate and intra-day air pressure change slope.

Considering the temporal memory effect of soil moisture, lag features of 1–3 days are built to capture the moisture accumulation effect and slow variation process, thereby reducing the model's over-sensitivity to short-term meteorological fluctuations.

In addition, based on the synergistic effect of meteorological factors on soil moisture, interaction features are also constructed.

For feature selection, a method combining forward sequential feature selection and F-test is adopted. First, the F-test is used to calculate the correlation score between each candidate feature and the target variable, and features with extremely low correlation are eliminated. Then, features are gradually added to the model starting from the one with the highest correlation. After adding each feature, 5-fold time-series cross-validation is conducted to evaluate the model performance. When the number of features increases to 44, the R^2 of the model on the test set reaches a peak of 0.673; a further increase in the number of features leads to a slight decline in performance. Therefore, 44 features are determined as the optimal feature subset.

2.3. Model Selection

In this study, the XGBoost algorithm was selected to construct a soil moisture prediction model. As an improved version of the Gradient Boosting Decision Tree (GBDT), this algorithm builds

ISSN: 2710-0588

an ensemble of decision trees through iteration. It can effectively handle non-linear relationships, avoid overfitting, and possesses the capability to evaluate feature importance.

2.3.1. Principle of the XGBoost Algorithm

The core idea of XGBoost is to train weak learners sequentially, such that each newly constructed decision tree can correct the prediction residuals of the previous round of the model. Finally, a strong predictive model is obtained by integrating all weak learners. Its objective function consists of a loss function and a regularization term, with the mathematical expression as follows:

$$Obj(\Theta) = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t)}\right) + \sum_{k=1}^{t} \Omega(f_k) \tag{2}$$

Where:

- $Obj(\Theta)$ denotes the objective function;
- $l(y_i, \hat{y}_i) = (y_i - \hat{y}_i)^2$ is the loss function for the t-th iteration;
- $\hat{y}_i^{(t)}$ represents the predicted value of sample i in the t-th iteration;
- $\Omega(f_k)$ is the complexity regularization term for the k-th decision tree, which is used to prevent overfitting. Its expression is given by:

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{3}$$

In this formula:

- T is the number of leaf nodes in the decision tree;
- w_j is the weight of the $j-th$ leaf node;
- γ and λ are regularization parameters, which control the number of leaf nodes and the penalty intensity of weights, respectively.

2.3.2. Demonstration of Model Advantages

Compared with traditional linear models, single decision trees, and neural networks, XGBoost exhibits significant advantages in soil moisture prediction:

1) It can automatically capture the non-linear relationships between meteorological factors and soil moisture without requiring complex non-linear transformation of data, making it suitable for scenarios where the interaction of meteorological factors is complex.

2) It effectively controls model complexity through regularization terms γ and λ, thereby avoiding overfitting and improving the generalization ability of the model.

3) It has strong robustness to missing values. There is no need for complex imputation of missing data, as missing features can be directly handled through the splitting rules of decision trees.

4) It can automatically evaluate feature importance, providing a basis for analyzing the mechanism by which meteorological factors affect soil moisture.

5) It features fast training speed and supports parallel computing, making it applicable to medium-scale time-series datasets.

ISSN: 2710-0588

2.4. Model Training and Validation

2.4.1. Training Strategy

Time-series cross-validation was employed for model training and hyperparameter tuning to avoid the "data leakage" issue associated with traditional random cross-validation. The dataset was divided into 5 consecutive subsets in chronological order; each subset served as the validation set once, while the remaining 4 subsets were used as the training set. This training and validation process was repeated sequentially 5 times, and the average value of the 5 validation results was taken as the model performance metric.

Grid search was adopted for hyperparameter tuning, with a focus on optimizing the following key parameters: learning rate, number of decision trees, tree depth, subsample ratio, and column sample ratio. The optimal hyperparameter combination determined through grid search was as follows: learning_rate = 0.1, n_estimators = 100, max_depth = 3, subsample = 0.8, colsample_bytree = 0.8, and regularization parameters $\gamma = 0.1$ and $\lambda = 1$.

2.4.2. Performance Evaluation and Validation Methods

The model performance was evaluated using three metrics: coefficient of determination R^2, RMSE, MAE and AUC with their calculation formulas provided below:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{4}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{5}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right| \tag{6}$$

$$AUC = \frac{1}{n_+ n_-}\sum_{i:y_i=+1}\sum_{j:y_j=-1}\mathbb{I}(f(x_i) > f(x_j)) + \frac{1}{2}\mathbb{I}(f(x_i) = f(x_j)) \tag{7}$$

Where y_i denotes the actual value, $\hat{y}_i$ represents the predicted value, $\bar{y}$ is the average of the actual values, n is the number of samples, n_+ is the number of positive samples (e.g., samples with "high soil moisture" when soil moisture is binarized), n_- is the number of negative samples (e.g., samples with "low soil moisture" when soil moisture is binarized), $f(x_i)$ is the predicted score for sample i, $f(x_j)$ is the predicted score for sample j, and $\mathbb{I}(\cdot)$ is an indicator function that takes the value 1 if the condition inside is true, and 0 otherwise. A R^2 value closer to 1 and RMSE/MAE values closer to 0 indicate better predictive performance of the model.

The dataset was split into a training set and a test set in chronological order: the training set covered the period from May 1, 2021, to June 20, 2021, accounting for 70% of the total samples, while the test set spanned from June 21, 2021, to July 31, 2021, making up 30% of the total samples. This division ensured that the test set data was temporally later than the training set, simulating the "predicting the future" scenario of the model in practical applications and thereby verifying the generalization ability of the model.

ISSN: 2710-0588

3. Research Results

3.1. Model Performance Analysis

Considering the coefficients of determination R^2, RMSE and AUC collectively, the model performance exhibited distinct patterns of change as the number of features increased. From the perspective of traditional regression metrics, the results of forward sequential feature selection showed that the R^2 of the model on the test set first increased, then stabilized, and finally decreased with the growth in the number of features. When the number of features rose from 1 to 44, the test set R^2 gradually increased from 0.321 to 0.673, while the test set RMSE decreased from 0.052 to 0.038 and the MAE dropped from 0.038 to 0.029, indicating continuous optimization of the model's quantitative prediction accuracy. However, when the number of features exceeded 44, the R^2 began to decline slightly, and both RMSE and MAE increased marginally-this phenomenon suggested that an excessive number of redundant features led to overfitting and a reduction in generalization ability. Therefore, 44 features were determined as the optimal number under regression metrics: at this point, the training set R^2 reached 0.892, the test set R^2 stood at 0.673, and the test set RMSE and MAE were 0.038 and 0.029, respectively, demonstrating a good balance between the model's fitting performance and generalization ability.

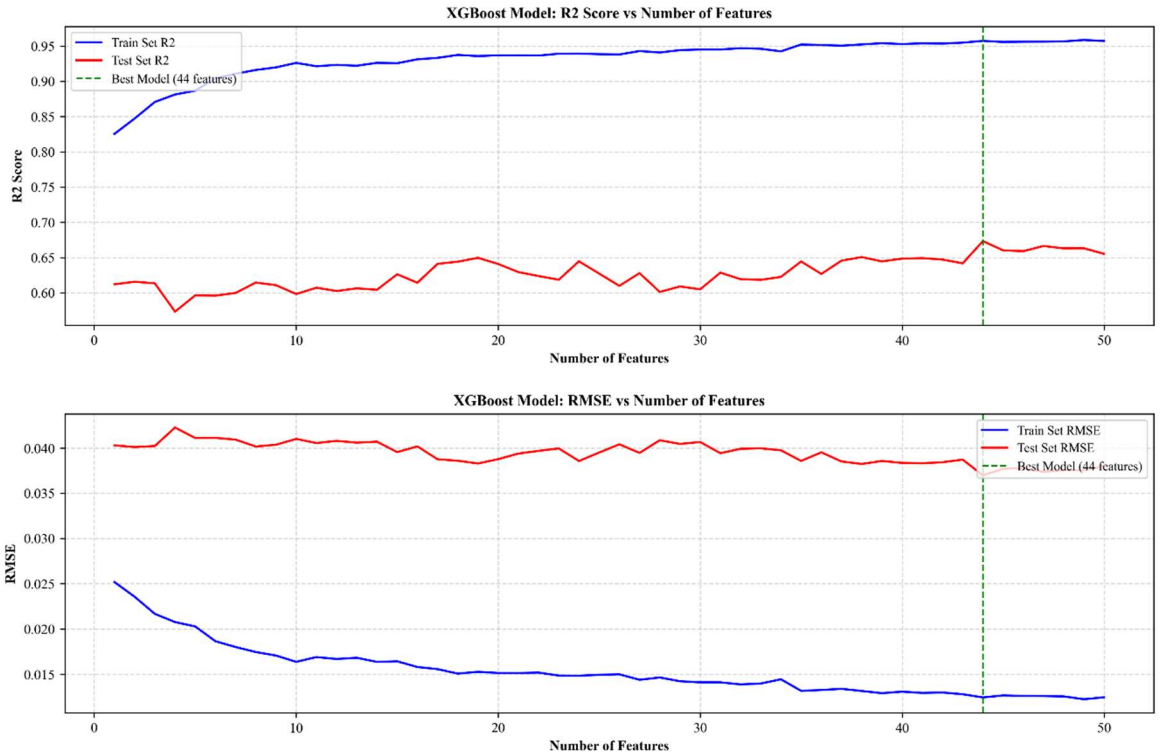

Figure 1. XGBoost Model: RMSE vs Number of Features

From the perspective of AUC, a metric for classification tasks, its trend of change complemented that of the regression metrics. The training set AUC continued to increase as the number of features grew, rising gradually from 0.938 (with 1 feature) to 0.982 (with 50 features), which indicated that the model's ability to distinguish between "high humidity" and "low humidity" categories-divided based on the median-was continuously enhanced during the training phase. The test set AUC showed a trend of "first increasing and then stabilizing": when the number of

ISSN: 2710-0588

features was 44, the test set AUC reached 0.815, and although the number of features continued to increase afterward, the AUC remained at a relatively high level of approximately 0.81. This result confirmed that with 44 features, the model not only achieved the highest accuracy in continuous-value prediction but also reached the peak of its ability to judge binary classification scenarios such as "whether soil moisture meets the standard," further verifying the optimality of 44 features, as this number ensured both the accuracy of regression prediction and strong performance in classification decision-making.

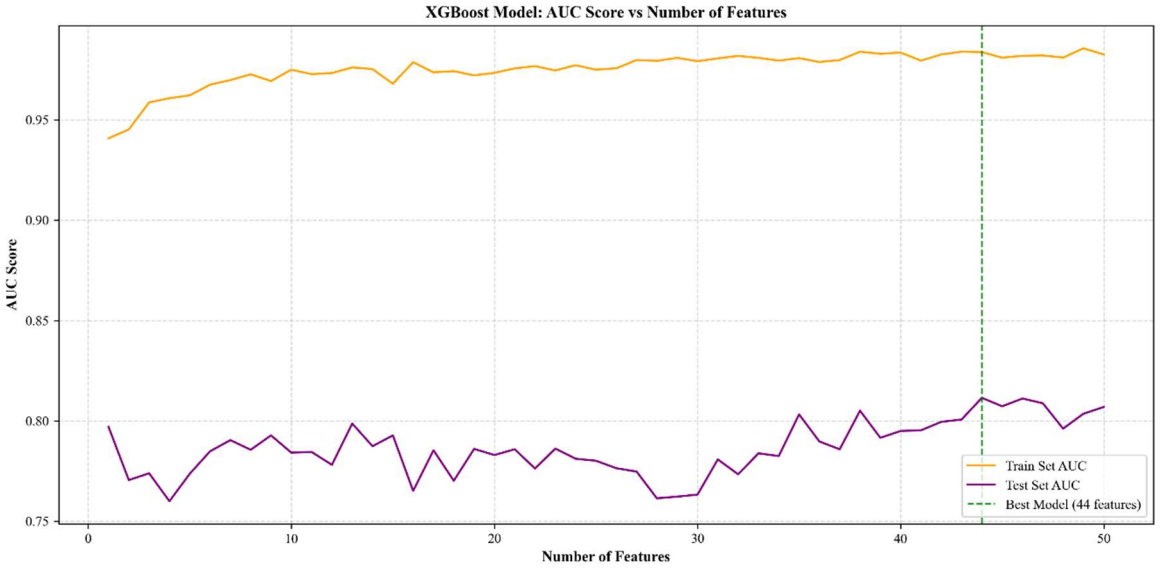

Figure 2. XGBoost Model: AUC Score vs Number of Features

3.2. Learning Curve Analysis

The learning curves revealed the impact of feature engineering on model performance from multiple dimensions, including R^2, RMSE, and AUC. When the number of features was small (≤20), the model suffered from underfitting: both the training set and test set R^2 values were low, and the gap between them was small, indicating that the model failed to fully capture the correlations between meteorological factors and soil moisture. When the number of features ranged from 20 to 44, both the training set and test set R^2 values increased rapidly, while the gap between them remained within a reasonable range-this suggested that the newly added features effectively improved the model's expressive capacity without causing obvious overfitting. When the number of features exceeded 44, the training set R^2 continued to rise slightly, but the test set R^2 decreased and the gap between the two expanded, meaning the model began to overfit to noise in the training data and its generalization ability declined.

3.3. Feature Importance Analysis

The results of feature importance evaluation from the XGBoost model showed that the key driving factors affecting 5 cm soil moisture, ranked by importance, were as follows: daily average temperature, 5 cm soil moisture on the previous day, daily total precipitation, daily average relative humidity, intra-day temperature range, total precipitation on the previous day, and the interaction term of temperature × relative humidity.

ISSN: 2710-0588

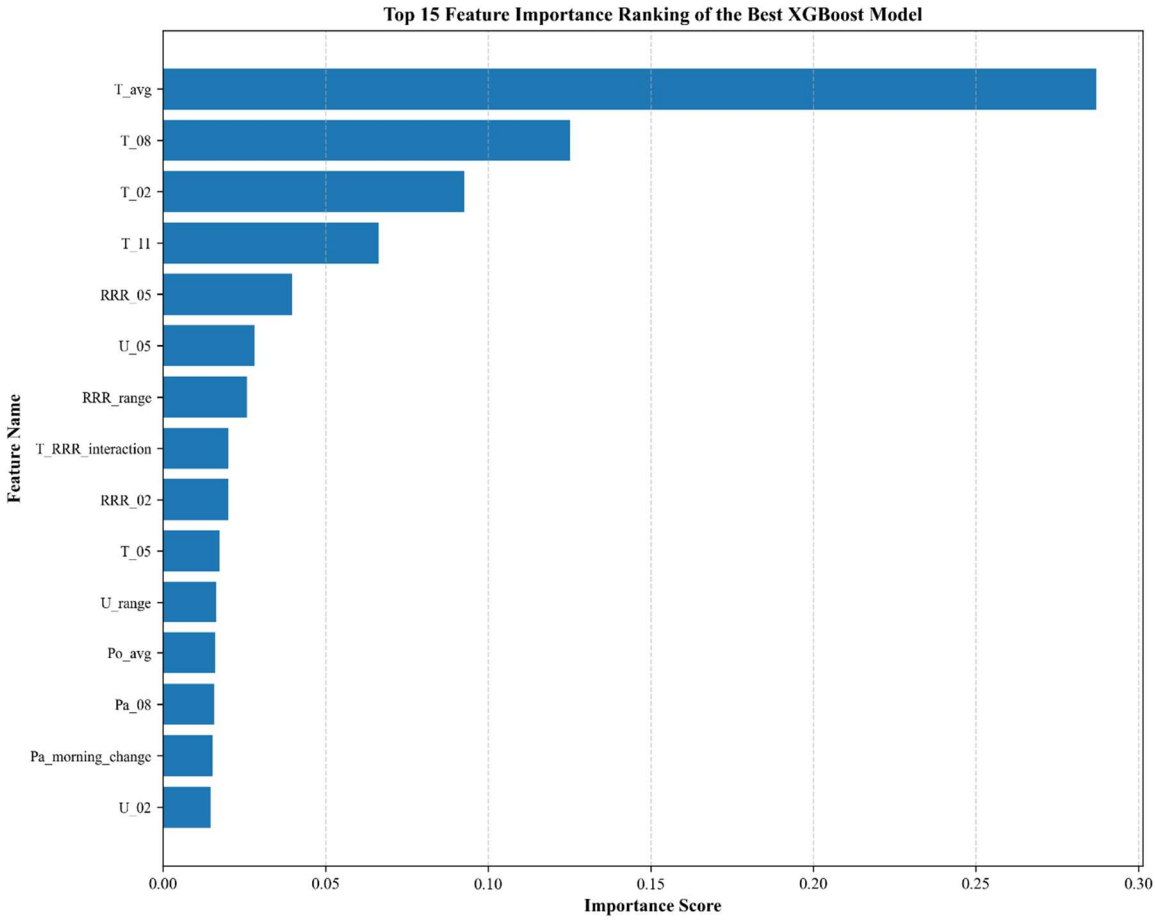

Figure 3. Top 15 Feature Importance Ranking of the Best XGBoost Model

The importance ranking of these key features is highly consistent with the physical mechanism of soil moisture variation, as elaborated below:

First, daily average temperature, as the most important feature, directly affects the evaporation rate of soil moisture. An increase in temperature accelerates the evaporation of moisture from the soil surface and simultaneously enhances the transpiration rate of crops, leading to a decrease in 5 cm soil moisture. Conversely, under low-temperature conditions, evaporation and transpiration are weakened, allowing soil moisture to be retained for a longer period and resulting in higher 5 cm soil moisture.

Second, soil moisture on the previous day reflects the temporal memory effect of soil moisture. The consumption and replenishment of soil moisture is a continuous process, and the moisture level on the previous day directly determines the initial moisture state of the current day—hence, this feature holds high importance.

Third, daily total precipitation is the most direct source of soil moisture replenishment. A larger amount of precipitation leads to more sufficient soil moisture replenishment and thus higher 5 cm soil moisture. When precipitation is 0, soil moisture is mainly affected by evaporation and crop water consumption, showing a downward trend.

Fourth, daily average relative humidity and intra-day temperature range also play significant roles. Relative humidity indirectly affects the evaporation rate by influencing the air vapor pressure deficit: when relative humidity is low, the air vapor pressure deficit is large, resulting in a high evaporation rate and a rapid decline in soil moisture. The intra-day temperature range reflects the intensity of diurnal temperature fluctuations; larger fluctuations lead to higher

ISSN: 2710-0588

instability in soil moisture evaporation, which exerts a significant impact on the dynamic changes of 5 cm soil moisture.

In addition, the importance of interaction features indicates that the impact of meteorological factors on soil moisture is not independent but involves synergistic effects. For instance, under high-temperature and low-humidity conditions, the evaporation rate is much higher than that under high-temperature and high-humidity conditions. Such synergistic effects are effectively captured by the model through interaction features.

4. Discussion

4.1. Result Interpretation and Mechanism Analysis

The XGBoost model constructed in this study achieved favorable performance in soil moisture prediction $(with\ R^2 = 0.673)$ and the rationality of this result can be explained from two aspects: physical mechanisms and model characteristics.

From the perspective of physical mechanisms, the key features identified by the model are directly related to the "replenishment-consumption" process of soil moisture. Precipitation serves as the primary source of soil moisture replenishment; temperature and relative humidity determine the consumption of soil moisture by influencing the evaporation rate; and the soil moisture of the previous day reflects the continuity of moisture changes. The high importance of these features is consistent with the dynamic balance law of soil moisture in farmland ecosystems.

From the perspective of model characteristics, the gradient boosting mechanism of XGBoost can effectively capture the nonlinear relationships between meteorological factors and soil moisture. For example, when precipitation is low (< 5 mm), the amount of soil moisture replenishment is insufficient to offset the consumption caused by evaporation, leading to a significant decrease in 5 cm soil moisture as temperature rises. When precipitation is high (> 20 mm), the soil moisture reaches saturation, and the impact of temperature on 5 cm soil moisture weakens. Such nonlinear relationships are accurately learned by the model through the integration of multiple decision trees, whereas traditional linear models fail to capture this threshold effect, resulting in lower prediction accuracy.

In addition, an R^2 of 0.673 holds high practical value in agricultural soil moisture prediction. The accuracy requirement of agricultural irrigation decisions for soil moisture is not absolutely strict-As long as the deviation between the predicted value and the actual value is within an acceptable range, the model can provide effective references for irrigation decisions. For instance, when the model predicts that the 5 cm soil moisture is below the threshold, farmers can activate the irrigation system; when the predicted value is above the threshold, irrigation can be postponed to avoid water waste.

4.2. Method Advantages and Innovations

Compared with existing studies on soil moisture prediction, the advantages and innovations of this study are mainly reflected in the following three aspects:

First, the innovation in in-depth feature engineering. Most existing studies adopt a single type of meteorological feature, while this study constructs a multi-dimensional feature set including intra-day statistical features, temporal variation features, lag features, and interaction features. In particular, the introduction of lag features and interaction features effectively captures the temporal memory effect of soil moisture and the synergistic effect of meteorological factors. For example, the inclusion of the previous day's soil moisture enables the model to consider the cumulative effect of moisture, and the introduction of the temperature × relative humidity interaction term allows the model to distinguish the impact of different temperature-humidity

ISSN: 2710-0588

combinations on evaporation. The construction of these features significantly improves the predictive performance of the model.

Second, the rationality in model selection. Compared with traditional linear models and single decision trees, the regularization mechanism of XGBoost effectively prevents overfitting; its robustness to missing values reduces the complexity of data preprocessing; and its feature importance evaluation function provides a basis for analyzing the influence mechanism of meteorological factors. For example, through the ranking of feature importance, it can be confirmed that daily average temperature is the most critical factor affecting soil moisture, which provides a key focus direction for subsequent agricultural meteorological research.

Third, the scientificity in validation methods. The adoption of time-series cross-validation and the chronological division of training-test sets avoids the problem of "data leakage" and ensures that the model can "predict the future" in practical applications rather than fitting the noise in historical data. This validation method is more in line with the actual application scenario of intelligent irrigation systems, enhancing the practical value of the research results.

4.3. Limitations Analysis

This study still has the following limitations, which need to be further addressed in future research:

First, data limitations. The time span of the research data only covers May to July 2021, lacking data from non-growing seasons such as winter and spring. This means the model's generalization ability across different seasons has not been verified. Meanwhile, the data is sourced from a single farm, with no data from areas with different soil types or different crop plantations. Thus, the applicability of the model to other farms requires further testing.

Second, idealization of model assumptions. The model assumes that the soil properties of the farm are uniform; however, in actual farms, soil texture may exhibit spatial heterogeneity. This heterogeneity affects the dynamic changes of soil moisture, leading to increased prediction errors of the model in areas with heterogeneous soil.

Third, failure to consider differences in crop water consumption. The model does not distinguish between the differences in water consumption characteristics of different crops. Nevertheless, different crops consume different amounts of water even in the same growth stage, which exerts varying impacts on soil moisture. For example, the water demand of corn is higher than that of soybeans; under the same meteorological conditions, soil moisture in corn-growing areas decreases more rapidly. The model's failure to account for such differences may result in reduced prediction accuracy.

References

[1] Li, Mengna, et al. "Climate-smart irrigation strategy can mitigate agricultural water consumption while ensuring food security under a changing climate." Agricultural Water Management 292 (2024): 108663.

[2] Maan, Cynthia, Marie-Claire ten Veldhuis, and Bas JH van de Wiel. "Dynamic root growth in response to depth-varying soil moisture availability: a rhizobox study." Hydrology and Earth System Sciences 27.12 (2023): 2341-2355.

[3] Jovani, Sancho AJ, S. Brosnan, and K. A. Byrne. "Partitioning of soil respiration in a first rotation beech plantation." Biology and Environment 117.2 (2017): 91-105.

[4] Gu, Xinyi, et al. "Response of Soil Moisture to Precipitation in the Source Region of the Yellow River." Advances in Atmospheric Sciences (2025): 1-20.

[5] Dai, Licong, et al. "Soil moisture variations in response to precipitation across different vegetation types on the northeastern Qinghai-Tibet plateau." Frontiers in Plant Science 13 (2022): 854152.

ISSN: 2710-0588

[6] Zhong, Shi, et al. "Temporal and spatial variations of soil moisture–Precipitation feedback in East China during the East Asian summer monsoon period: A sensitivity study." Atmospheric Research 213 (2018): 163-172.

[7] Chen, Tianqi, and Carlos Guestrin. "Xgboost: A scalable tree boosting system." Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016.