

Article

Construction of Yunnan Flue-Cured Tobacco Yield Integrated Learning Prediction Model Driven by Meteorological Data

Yunshuang Wang ¹ , Jinheng Zhang ², Xiaoyi Bai ², Mengyan Zhao ¹, Xianjin Jin ¹ and Bing Zhou ^{2,*} 

¹ College of Big Data, Yunnan Agricultural University, Kunming 650500, China

² College of Science, Yunnan Agricultural University, Kunming 650500, China

* Correspondence: bingzhoukm@126.com; Tel.: +86-15368097838

Abstract

The timely and accurate prediction of flue-cured tobacco yield is crucial for its stable yield and income growth. Based on yield and meteorological data from 2003 to 2023 (from the NASA POWER database) of Yunnan Province, this study constructed a coupled framework of polynomial regression and a Stacking ensemble model. Four trend yield separation methods were compared, with polynomial regression selected as being optimal for capturing long-term trends. A total of 135 meteorological features were built using flue-cured tobacco's growth period data, and 17 core features were screened via Pearson's correlation analysis and Recursive Feature Elimination (RFE). With Random Forest (RF), Multi-Layer Perceptron (MLP), and Support Vector Regression (SVR) as base models, a ridge regression meta-model was developed to predict meteorological yield. The final results were obtained by integrating trend and meteorological yields, and core influencing factors were analyzed via SHapley Additive exPlanations (SHAP). The results showed that the Stacking model had the best predictive performance, significantly outperforming single models; August was the optimal prediction lead time; and the day–night temperature difference in the August maturity stage and the solar radiation in the April transplantation stage were core yield-influencing factors. This framework provides a practical yield prediction tool for Yunnan's flue-cured tobacco areas and offers important empirical support for exploring meteorology–yield interactions in subtropical plateau crops.



Academic Editor: Alberto San Bautista

Received: 11 September 2025

Revised: 4 October 2025

Accepted: 10 October 2025

Published: 21 October 2025

Citation: Wang, Y.; Zhang, J.; Bai, X.; Zhao, M.; Jin, X.; Zhou, B.

Construction of Yunnan Flue-Cured Tobacco Yield Integrated Learning Prediction Model Driven by Meteorological Data. *Agronomy* **2025**, *15*, 2436. <https://doi.org/10.3390/agronomy15102436>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Keywords: flue-cured tobacco; stacking model; SHAP; yield forecasting

1. Introduction

China is the world's largest tobacco producer and consumer, and its yield and sales rank first in the world, accounting for about 30% of global tobacco production [1,2]. As an important pillar of the national economy, the tobacco industry had a total industrial and commercial tax and profit of 1600.8 billion CNY in 2024 (accounting for 9.24% of the national tax revenue), and a total fiscal revenue of 1544.6 billion CNY. Its stable development is crucial to economic and fiscal revenue growth [3–5].

In the global tobacco industry chain, flue-cured tobacco is the core raw material for cigarette production, and its yield and quality directly determine the trajectory of industrial development, thus becoming the core research object of tobacco agronomy. As the core production area of China's flue-cured tobacco, Yunnan Province has a long history of planting and extensive regional distribution. The 12 prefectures and cities across the province carry out large-scale planting annually. Statistics in 2024 showed that Yunnan Province's tobacco industry firmly ranked first in the country, accounting for 39.8% of

the total contribution to the national economy, making it an important area to ensure the security of the national tobacco supply and industrial stability [6].

Meteorological conditions are the key drivers for regulating the production and quality of flue-cured tobacco [7]. Yunnan Province has a subtropical plateau monsoon climate [8]. Although the growth period of flue-cured tobacco field (April–September) has the basic advantages of sufficient light and appropriate temperature, extreme meteorological events such as regional droughts and phased rainstorms caused by strong convective weather in summer occur frequently, causing significant interference to the survival rate of flue-cured tobacco during the transplantation stage and long-term nutrient accumulation, and the formation of quality in the mature period, resulting in outstanding inter-annual fluctuations in the yield. Therefore, accurately analyzing the correlation mechanism between meteorological factors and flue-cured tobacco production and building a scientific yield prediction model have important theoretical and practical significance for guiding the optimization of flue-cured tobacco production layout and the formulation of disaster prevention and mitigation measures, and ensuring the stable development of the industry. Crop yield formation is the result of the synergistic action of multiple factors such as agricultural scientific and technological progress (such as variety improvement, cultivation technology upgrade) [9], fluctuations in meteorological conditions [10], soil characteristics [11], and field management measures [12]. Among them, meteorological factors have become the core variables in yield prediction research due to their strong inter-annual variability and outstanding uncontrollability [13].

In recent years, the research on the relationship between meteorological factors and crop yield has become increasingly in-depth. Based on temperature and precipitation data, Zhao et al. used regression models to quantify the degree of influence of meteorological factors on China's grain yield [14]; Ma et al. combined trend yield and meteorological yield to construct a time series prediction model, providing method support for the dynamic prediction of crop yields [15]; Bognár et al. revealed the significant impact of seasonal climate fluctuations on crop yields such as corn and winter wheat through partial least squares regression [16]; and Didari et al. used the Lasso regression model to confirm the key role of extreme temperature and precipitation events in the estimation of dry wheat yields [17]. All the above studies show that the nonlinear correlation mechanism between analytical meteorological factors and crop yield is the core breakthrough for achieving accurate prediction. Current crop yield prediction methods are mainly divided into three categories: crop model methods, statistical model methods, and decomposition model methods. The crop model method realizes yield estimation by simulating the physiological and biochemical processes of crops, but relies on fine field parameters. Due to the high cost of parameter acquisition and poor timeliness, it is limited in large-scale practical applications [18]. The statistical model method predicts yield by constructing a mathematical relationship between meteorological factors and yield, but the ability to analyze complex nonlinear relationships is limited, making it difficult to capture the mutation effect of extreme meteorological events [19]. The decomposition model method separates the actual yield into trend yield (reflecting long-term stability factors such as scientific and technological progress) and meteorological yield (reflecting the short-term impact of climate fluctuations) and can realize dynamic prediction based on real-time meteorological data, and its application value is more prominent in agricultural production [20]. Commonly used trend yield extraction methods include the moving average method, the exponential smoothing method, the high-pass filtering method, etc., but the existing methods still need to be optimized to improve the capture accuracy of long-term trends and are susceptible to short-term random fluctuations.

With the outstanding advantages of dealing with nonlinear relationships, machine learning algorithms have been widely used in crop yield prediction [21]. However, a single model is limited by its own assumptions and data adaptability, and prediction accuracy is susceptible to extreme meteorological events [22,23]. When comparing various machine learning and deep learning models, Sharma et al. found that single models such as decision trees and convolutional neural networks have obvious shortcomings in crop yield prediction, and their errors are significantly larger than those of stochastic forests and other ensemble models [24]; Luo et al. [25] further confirmed in corn yield prediction research under drought conditions that traditional single algorithms reduce their prediction capabilities in extreme climate events. Even if remote sensing indicator optimization is introduced, their robustness cannot meet the demand. By integrating the prediction results of multiple basic models, the ensemble learning model can effectively reduce the deviation and variance of a single model and improve the prediction stability [26]. Among them, as an advanced integration method, the Stacking model is outstanding in integrating the feature information of different basic models through meta-models. Islam et al. [27] found in rice yield estimation that after Stacking multiple tree-based regression models, their prediction accuracy was significantly better than linear regression or individual machine learning models, especially when handling nonlinear coupling problems between meteorological and remote sensing data.

Based on meteorological data and flue-cured tobacco yield data, with 12 long-term major flue-cured tobacco-producing prefectures and cities in Yunnan Province as the study area, this study constructs a coupled flue-cured tobacco yield prediction framework integrating polynomial regression and the Stacking model. It focuses on exploring the applicability of this framework in the accurate prediction of flue-cured tobacco yield and its ability to interpret key meteorological factors. Using this coupled framework allows for the accurate separation of trend yield and meteorological yield, mitigating the interference of short-term fluctuations on prediction results. Meanwhile, it improves the accuracy of interpreting the nonlinear effects of meteorological factors during key growth stages and optimizes the efficiency of regional flue-cured tobacco yield prediction. The specific research objectives are as follows: (1) Verify the feasibility of the polynomial regression–Stacking coupled framework for flue-cured tobacco yield prediction in Yunnan. (2) Interpret the influence mechanism of meteorological factors during key growth stages on flue-cured tobacco yield using the SHAP (SHapley Additive exPlanations) method. (3) Identify the optimal lead time for flue-cured tobacco yield prediction, so as to provide a basis for formulating production scheduling and disaster prevention/mitigation measures in tobacco-growing areas. This study can provide methodological support for flue-cured tobacco yield prediction in subtropical plateau tobacco-growing areas and offer technical references for the application of machine learning in crop yield simulation.

2. Materials and Methods

2.1. Flue-Cured Tobacco Yield Forecast Process

The forecast process for flue-cured tobacco yield is shown in Figure 1, which mainly includes the following steps: (1) Collect meteorological data for the flue-cured tobacco growth period (April–September), construct basic meteorological characteristics, derivative meteorological characteristics, and interactive meteorological characteristics through feature engineering, combine Pearson’s correlation analysis with Recursive Feature Elimination (RFE) to screen key features, synchronize polynomial regression to strip non-meteorological influences from the yield data, and obtain meteorological yield as the target variable. (2) Taking the screened meteorological characteristics as the input, basic Random Forest (RF), Multi-Layer Perceptron (MLP), and Support Vector Regression (SVR) models were

trained. The prediction results were generated after 5-fold cross-validation (5-fold CV). The ridge regression meta-model was input to build a Stacking integrated model that was tuned using Grid Search, while predicting the yield several months in advance. A time series verification strategy was adopted, where the data from 2003 to 2020 were used as training data and the data from 2021 to 2023 were used as prediction data, and the model performance was evaluated through indicators such as R^2 and root mean square error (RMSE). (3) Use feature analysis methods to analyze the impact of key variables on yield. First, use SHAP values to analyze the overall importance of meteorological characteristics to yield; second, combine the division of flue-cured tobacco breeding periods to analyze the response laws of key characteristics at different growth stages, and reveal the differences in the stage impact of meteorological conditions.

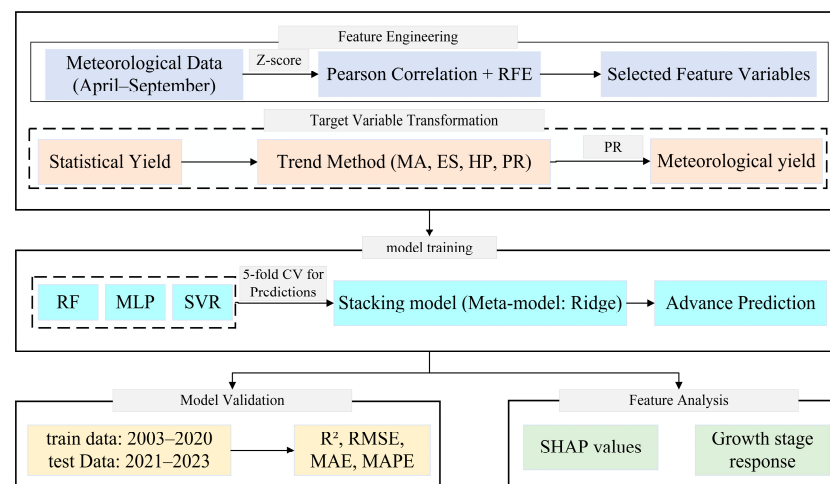


Figure 1. Prediction process of tobacco yield.

2.2. Overview of the Study Area

Yunnan Province is located by the southwestern border of China (97°31′–106°11′ E, 21°8′–29°15′ N), with a complex and diverse geographical environment and significant regional climate differences (Figure 2). It is not only a typical three-dimensional climate zone in the country, but also a province with the widest area of tobacco and the highest total yield in China. The 12 prefectures and cities in the province are annual flue-cured tobacco cultivation areas. The growth period of flue-cured tobacco fields in this area (April to September) has sufficient sunshine resources, appropriate temperature conditions, and a balanced precipitation distribution, forming the optimal climate combination conducive to the growth and development of flue-cured tobacco, making it the core representative production area of flue-cured tobacco cultivation in the province. However, complex and changeable climatic conditions such as seasonal extreme precipitation and regional drought pose a significant threat to flue-cured tobacco production. Therefore, prediction research on flue-cured tobacco yield based on meteorological factors is of great practical significance to ensure the stable production and increased income of flue-cured tobacco.

2.3. Data Source

2.3.1. Production Data

This study collected total flue-cured tobacco yield (unit: ton) and planting area data (unit: hectare) from 12 prefectures and cities in the study area from 2003 to 2023, including Kunming, Qujing, Yuxi, Chuxiong, Honghe, Dali, Zhaotong, Baoshan, Lijiang, Puer, Lincang, and Wenshan. A total of 504 data points were obtained, and all data were derived from the *Yunnan Statistical Yearbook* compiled by the Yunnan Provincial Bureau of Statistics.

over the years [28]. Among them, the yield data and meteorological data from 2021 to 2023 are specifically used for the independent verification of prediction models.

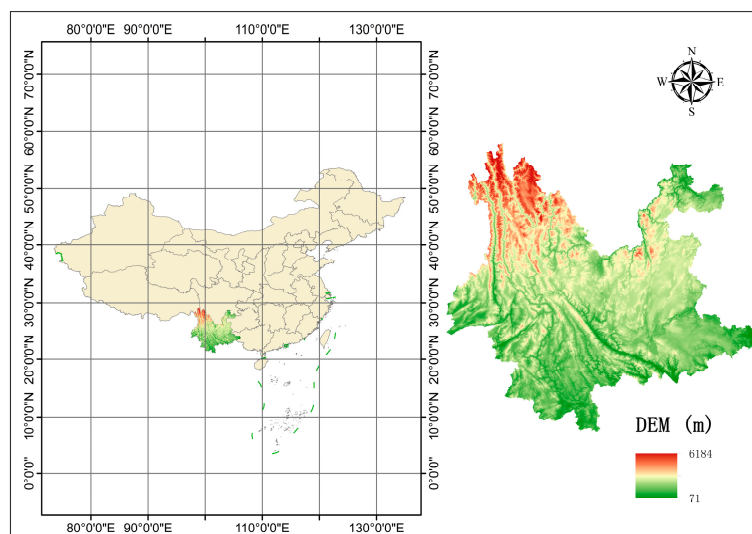


Figure 2. Location map of the study area and Digital Elevation Model (DEM) distribution.

2.3.2. Meteorological Data

The growth period of Yunnan tobacco fields can be divided into three key fertility stages: transplanting and rooting period (April–May), vigorous growth period (June–July), and mature period (August–September) [29]. In order to systematically explore the correlation mechanism between flue-cured tobacco yield and climatic conditions, this study was conducted based on the division of the fertility stages of flue-cured tobacco fields. The daily average meteorological data from 2003 to 2023 were obtained from the NASA POWER database (<https://power.larc.nasa.gov/>, accessed on 5 June 2025), with a spatial resolution of $0.5^\circ \times 0.5^\circ$, and 6 core meteorological factors in the field growth period (April to September) were screened, including the maximum temperature ($^\circ\text{C}$), the minimum temperature ($^\circ\text{C}$), the cumulative precipitation (mm), the cumulative solar radiation (kJ/m^2), the wind speed (m/s), and the water vapor pressure (hPa).

2.4. Data Preprocessing Method

2.4.1. Production Data Preprocessing

Based on the total yield of flue-cured tobacco from prefecture- and municipal-level regions from 2003 to 2023, the yield of flue-cured tobacco is calculated (Y_s). The calculation formula is as follows:

$$Y = \frac{Y_s}{A} \quad (1)$$

The yield data are processed using the decomposition model method. This method decomposes crop yield into three parts: trend yield, meteorological yield, and random yield. The expression is

$$Y = Y_t + Y_m + Y_r \quad (2)$$

where Y is the actual yield (kg/ha); Y_t is the trend yield (kg/ha), which mainly reflects the impact of long-term stability factors such as agricultural scientific and technological progress on yield; Y_m is the meteorological yield (kg/ha), which reflects the effect of inter-annual fluctuations in climatic conditions on yield; and Y_r is the random yield (kg/ha), which represents the random error in the model, and this variable is usually dimensionless. Since the proportion of random yield in this study is extremely low and the impact on

the results is negligible, the actual yield can be simplified to the sum of trend yield and meteorological yield as follows:

$$Y = Y_t + Y_m \quad (3)$$

In this study, the trend yield (Y_t) is obtained by fitting the yield sequence, and then the final yield prediction model is constructed with the meteorological yield (Y_m) prediction model. Four common trend yield separation methods are used for comparison and analysis. The moving average method smooths short-term fluctuations by calculating the average value of the observed values in a specific window in the time series, using the moving average as the trend yield, and the residuals between the original sequence and the trend yield are the meteorological yield. The exponential smoothing method improves the moving average method based on the principle of weight decay, and captures the yield trend by dynamically adjusting the historical data weight, which is suitable for the processing of non-stationary yield sequences. The high-pass (HP) filtering method separates the high-frequency perturbation components and the low-frequency trend components of the time series based on the state space model, and uses the low-frequency component as the trend yield. Polynomial regression (PR) fits the long-term yield trend through the n -degree polynomial function, and the expression is

$$Y_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \cdots + \beta_n t^n + \varepsilon \quad (4)$$

In the formula, t is a time variable; $\beta_0, \beta_1, \cdots, \beta_n$ is a polynomial coefficient; and ε is a random error term. By estimating the polynomial coefficient, the characteristics of the growth rate and inflection point of the yield trend can be analyzed.

2.4.2. Meteorological Data Preprocessing

(1) Construction of meteorological factors

In the meteorological data preprocessing stage, first, based on the study's regional administrative boundary division, the daily meteorological observation data from 2003 to 2023 were aggregated by geographical attributes and converted into a monthly regional climate data set in order to eliminate the random fluctuations of single-day meteorological data and more accurately reflect the overall climate characteristics of the months during which flue-cured tobacco grows, providing a standardized and regionalized data basis for subsequent meteorological factor construction.

In the process of meteorological factor construction, this study was based on 6 original meteorological observation indicators and expanded the systemic feature engineering to form a total of 135 flue-cured climate factor data sets in 4 categories, aiming to comprehensively analyze the mechanism of influence of meteorological conditions on flue-cured tobacco yield, and specifically included the following:

Basic meteorological characteristics: These characteristics cover the original indicators such as the maximum temperature, minimum temperature, cumulative precipitation, cumulative solar radiation, wind speed, and water vapor pressure obtained by direct observation, providing basic data support for climate factor analysis.

Derived meteorological characteristics: The extension of the quantitative attributes of meteorological elements is achieved through the mathematical transformation and statistical calculation of the original meteorological data, including active accumulated temperature (the cumulative value of effective temperature in crop growth), effective accumulated temperature (the sum of temperatures higher than the lower limit of biological), and temperature difference (highest and lowest temperature difference value) [30,31].

Meteorological suitability indexes: These indexes are constructed in combination with the biological characteristics of tobacco and a growth threshold model, including evaluation

indicators, such as temperature suitability, lighting suitability, and precipitation suitability, to quantify the degree of adaptation between meteorological conditions and flue-cured tobacco growth needs [32,33].

Meteorological interaction characteristics: These characteristics are generated through multi-factor coupling operations, including light–precipitation coupling terms, temperature–light coupling terms, etc., to capture the joint impact of the synergistic effects between meteorological factors on yield formation [34].

The above four types of characteristics range from basic observation, attribute extension, adaptability assessment to synergistic effect capture, and the multi-dimensional coverage of the comprehensive effect of meteorological conditions on the growth and development of flue-cured tobacco, providing systematic feature support for the analysis of subsequent yield impact mechanisms.

(2) Z-score standardization

In order to eliminate the impact of dimension differences and regional yield heterogeneity, the meteorological characteristics and yield data are standardized. The formula is

$$X^* = \frac{X - \bar{X}}{S} \quad (5)$$

where X is the original data; X^* is the standardized data; $\bar{X}$ is the mean of the original data; and S is the standard deviation of the original data. The standardized data presents distribution characteristics with a mean of 0 and a standard deviation of 1, which not only retains the relative relationship between variables, but also effectively weakens the interference of outliers.

(3) Feature filtering method

Pearson's correlation coefficient screening uses the quantification of linear correlation strength between features to eliminate redundant features. Calculate the Pearson correlation coefficient (r_{ij}) for any two features as follows:

$$r_{ij} = \frac{\sum_{k=1}^m (x_{ik} - \bar{x}_i)(x_{jk} - \bar{x}_j)}{\sqrt{\sum_{k=1}^m (x_{ik} - \bar{x}_i)^2 \sum_{k=1}^m (x_{jk} - \bar{x}_j)^2}} \quad (6)$$

In the formula, $\bar{x}_{ik}$ and $\bar{x}_{jk}$ are the values of features k and i in the j th sample; x_{ik} and x_{jk} are the mean values of features i and j , respectively; and m is the sample size. Set the correlation coefficient threshold to filter highly correlated feature pairs and retain features with a higher correlation with the target variable (meteorological yield) (correlation intensity ($r_t = |r_{ty}|$), where r_{ty} is the correlation coefficient of feature (t) and target variable (y)).

In this study, Recursive Feature Elimination (RFE) is based on a Random Forest and retains key features through multiple rounds of screening. Each round calculates the importance of features based on the reduction in node impurity, removes the least important features, and retrains the model until the preset number of features or the model performance is stable. The subset of features retained by round t is

$$F_t = \frac{F_{t-1}}{f_t} \quad (7)$$

In the formula, F_{t-1} is a subset of features retained by the $t - 1$ -th wheel; f_t is a feature removed by the t -th wheel. The model performance evaluation index is the mean square error of the verification set, calculated as follows:

$$MSE_t = \frac{1}{m} \sum_{k=1}^m (y_k - \hat{y}_{t,k})^2 \quad (8)$$

where y_k is the real label of the verification set; $\hat{y}_{t,k}$ is the predicted value of the t round model.

2.5. Construction of Meteorological Yield Forecast Model

To verify the practical operability of the model, all model training in this study was completed on a conventional scientific research workstation (Dell Inc., Round Rock, TX, USA; configuration: Intel Xeon Silver 4214R processor, 128 GB memory). To further quantify the computational complexity of each prediction model, this study tested and counted the core complexity indicators of the models (Table 1). Among the single models, Ridge Regression exhibits the lowest complexity due to its linear structure; the complexity of MLP, RF, and SVR is consistent with their nonlinear or ensemble architecture characteristics. Although the Stacking model integrates multiple base models, it still maintains good computational efficiency and is suitable for conventional research environments.

Table 1. Computational Complexity Indicators of Different Prediction Models.

Model	Number of Parameters (k)	FLOPS (Million Operations)	Time (s)
RF	214.20	32.13	222.55
MLP	0.38	0.11	54.49
SVR	0.16	0.38	0.78
Ridge	0.02	0.04	0.51
Stacking Model	214.74	195.72	10.23

2.5.1. Single Prediction Model

Random Forest (RF): Multiple training sample sets are generated through Bootstrap sampling, and each sample set is used to train a decision tree. The final prediction result is determined by multiple trees in a comprehensive manner, with a strong generalization ability and anti-overfitting characteristics.

Multi-Layer Perceptron (MLP): A feed-forward neural network composed of an input layer, several hidden layers, and an output layer. Neurons of each layer are fully connected. The hidden layer maps input features through a nonlinear activation function, optimizes weights through a backpropagation algorithm, and is good at capturing complex nonlinear relationships.

Support Vector Regression (SVR): The core of the application of support vector machines in regression problems is to find the optimal hyperplane so that the samples fall within the interval band with a width of 2ϵ as much as possible; by mapping data to a high-dimensional space via a kernel function, it can fit nonlinear data, which has strong processing capabilities for high-dimensional data and uses anti-noise interference.

Ridge regression: An improved linear regression method, introducing L2 regularization terms into the loss function, controlling the complexity of the model by punishing coefficients, alleviating the multicollinearity problem, and improving the stability of parameter estimation.

2.5.2. Stacking Model Construction

Stacking models are an ensemble learning method that improves performance by integrating multi-base models to predict results. Its core logic is to input the prediction results of the basic model as new features into the meta-model to realize secondary learning. This study uses RF, MLP, and SVR as the base models, takes their prediction results as input features, and uses ridge regression as the meta-model to construct a Stacking model for predicting the meteorological yield of flue-cured tobacco.

2.5.3. Hyperparameter Optimization Method

Grid Search (GridSearchCV) is employed for the hyperparameter tuning of the aforementioned machine learning models. This method is suitable for scenarios with moderately dimensional and well-defined hyperparameter spaces; it evaluates the performance of each hyperparameter combination by traversing a predefined parameter grid and integrating 5-fold cross-validation (random_state = 42), with the negative mean squared error (neg_MSE) as the objective function. It achieves “exploration” by fully covering the predefined parameter space (avoiding the omission of potential high-quality hyperparameter combinations) and “exploitation” by selecting the combination with the optimal validation performance (focusing on effective parameter intervals). Ultimately, it efficiently identifies the optimal hyperparameter combination within the predefined range, reducing the uncertainties caused by subjective parameter tuning while enhancing the model performance and experimental reproducibility.

2.6. SHAP Interpretability Analysis

The SHapley Additive exPlanations (SHAP) method is used to explain the model output, and the Shapley value in cooperative game theory quantifies the contribution of each feature to the prediction results and clarifies the direction and intensity of influence of each feature. The calculation formula is as follows:

$$\text{SHAP}_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|! \cdot (|N| - |S| - 1)!}{|N|!} [f(S \cup \{i\}) - f(S)] \quad (9)$$

In the formula, SHAP_i is the SHAP value of feature (i), indicating its contribution to the prediction result; S is a subset without feature i; N is the complete set of features; and $f(S)$ is the prediction result of the model on the subset S. The application of the SHAP method can clarify the key meteorological factors that affect flue-cured tobacco production and their degree of action.

2.7. Model Evaluation Indicators

This study used four indicators to evaluate the prediction accuracy, including the determination coefficient (R^2), root mean square error (RMSE), average absolute error (MAE), and average absolute percentage error (MAPE).

The coefficient (R^2) measures the model’s ability to interpret data variations. The value range is 0, 1. The closer it is to 1, the better the fitting effect.

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (10)$$

The root mean square error (RMSE) reflects the standard deviation of the prediction error. The smaller the value, the higher the accuracy.

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (11)$$

The average absolute error (MAE) reflects the mean of the absolute value of the prediction error. The smaller the value, the higher the accuracy.

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (12)$$

The average absolute percentage error (MAPE) reflects the mean of relative error, and the smaller the value indicates the higher the accuracy.

$$\text{MAPE} = \frac{1}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right| \times 100\% \quad (13)$$

In the formula, y_i is the actual measured value of i ; $\hat{y}_i$ is the predicted value of i ; $\bar{y}$ is the average value of the measured value; and n is the sample size.

3. Results

3.1. Comparison of Yield Decomposition Results for Flue-Cured Tobacco

To clarify the differences in the application of different trend decomposition methods in flue-cured tobacco yield analysis, the flue-cured tobacco yield in Honghe from 2003 to 2020 was used as the study subject, and the moving average method, exponential smoothing method, high-pass filtering method, and polynomial regression method were adopted to conduct the separation experiment of trend yield and meteorological yield.

The results show that there are significant differences in the decomposition performance of the four methods for flue-cured tobacco yield sequences: the trend yield extracted by the moving average method and exponential smoothing method shows obvious high-frequency fluctuations. Because long-term influence factors such as technological iteration and policy-driven factors in flue-cured tobacco yield fluctuations have continuous logic, these two methods are too sensitive to short-term fluctuations, and it is difficult to stably characterize the long-term trend, resulting in frequent fluctuations in the trend yield curve, which cannot accurately reflect the long-term development trends of the industry. The high-pass filtering method is based on the principle of frequency domain separation and has a strong separation ability for meteorological disturbances (high-frequency signals). It can be seen from Figure 3 that its trend yield fluctuation amplitude is lower than that of the moving average method and the exponential smoothing method, but the trend yield extracted by this method shows the characteristics of “rise first and then fall”, which is contrary to the realistic logic of continuous progress in flue-cured production technology and the steady increase in industrial investment. Driven by agricultural technology iteration and continuous policy support, the trend yield should show a continuous increase or gradual steady-state evolution. Due to an excessive focus on the separation of high-frequency meteorological signals, the low-frequency long-term components of trend yield are cut into segmented forms, which easily lead to phased deviations in trend analysis and are difficult to adapt to the continuous evolution process of “technology accumulation–continuous gain” in industrial development; in comparison, the trend yield curve fitted by the polynomial regression method is smooth and continuous. It effectively captures the long-term evolution trend of flue-cured tobacco production by constructing mathematical models, adapts to the cumulative gain effect brought about by technological progress, and effectively weakens short-term random fluctuations interference, which is highly consistent with the development laws of the tobacco industry and can better accurately reflect the driving effect of long-term factors on yield.

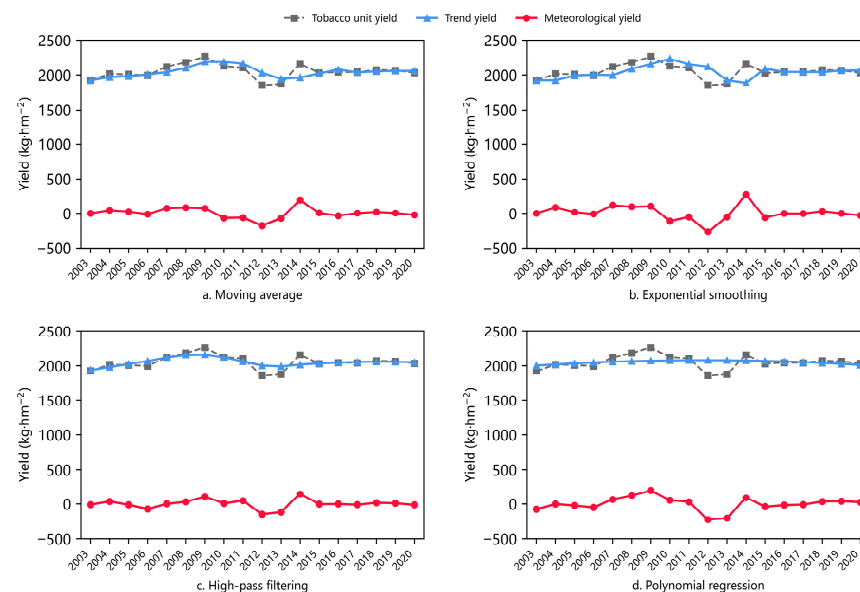


Figure 3. Comparative results of different decomposition models of flue-cured tobacco production in Honghe area from 2003 to 2020.

The separation accuracy of meteorological yield depends on the accuracy of trend yield extraction. Different methods have significant differences in the analysis of meteorological yield: the polynomial regression method enables accurate trend analysis, so the meteorological yield separated by this method more truly reflects the impact of climate fluctuations. In 2012, the meteorological yield of flue-cured tobacco of -202 kg/ha was directly related to the extreme precipitation in Yunnan that year, causing diseases in tobacco fields. Extreme climate events destroy the growth environment of tobacco plants, inhibit photosynthesis and nutrient absorption processes, and ultimately cause yield loss. This result reflects the driving effect of meteorological factors on flue-cured tobacco production [35]. In contrast, other methods, such as the moving average method and the exponential smoothing method, overshadow meteorological signals due to trend fluctuations, resulting in the inability to clearly distinguish climate disturbances and trend noise from meteorological yield. The high-pass filtering method deviates from industrial logic due to trend segmentation, causing meteorological yield to be doped with false fluctuating components, which both cause “signal distortion” problems and cannot provide accurate data support for the research of climate and yield response mechanisms.

In summary, the polynomial regression method has significant advantages in the trend decomposition of flue-cured tobacco yield. By fitting the smoothing curve, it not only captures the long-term increase in production driven by technological progress but also weakens the interference of abnormal meteorological events on trend separation, and builds a solid data foundation for the subsequent prediction of flue-cured meteorological yield. At the same time, precisely separated meteorological yield can quantify the degree of impact of climate factors on yield, provide method support for the industry to formulate disaster avoidance planting plans and improve meteorological insurance strategies, and help the flue-cured tobacco industry improve climate adaptability and risk response capabilities.

3.2. Meteorological Characteristics Screening

In order to eliminate the influence of multicollinearity among variables and improve the model’s computing efficiency and prediction accuracy, the characteristic factors with a high correlation were eliminated through Pearson’s correlation analysis, and 83 characteristic factors were retained for subsequent screening. These characteristic factors specifically

include 16 basic meteorological characteristics, 33 derivative meteorological characteristics, 19 meteorological suitability indexes, and 15 meteorological interaction characteristics.

In the process of Recursive Feature Elimination (RFE), Random Forest was used as the base model, the model performance was evaluated through cross-validation scores, and the optimal number of subsets of features was determined based on the mean square error (MSE). The results show that with the increase in the number of input variables, MSE shows a phased change: the intervals of 1–6 variables show a significant downward trend, the intervals of 6–17 variables fluctuate, and the intervals of 17–83 variables rise slowly (Figure 4). Taking into account both the model performance and variable redundancy, we finally selected 17 variables for model construction. This subset not only retains the core information of the original data set, but also effectively avoids the model overfitting problem caused by redundant variables.

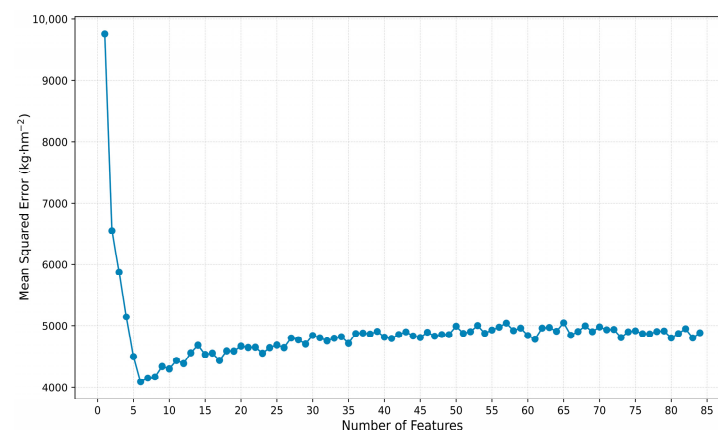


Figure 4. Trend of RFE feature selection accuracy with the number of features.

3.3. Results and Analysis of Yield Prediction Modeling

In order to verify the predictive effectiveness of different models on flue-cured tobacco yield, this study used the data of 216 flue-cured tobacco yields from 12 regions in Yunnan Province (2003–2020) as the sample set, and randomly divided them into a training set (151 samples) and a test set (65 samples) at a 7:3 ratio; retained the data from 2021 to 2023 for model performance verification; and compared the prediction results of the Random Forest (RF), Multi-Layer Perceptron (MLP), Support Vector Regression (SVR), ridge regression (Ridge), and Stacking models (Table 2).

Table 2. Prediction results of the flue-cured tobacco yield test set for different prediction models.

Model	R ²	RMSE/(kg/ha)	MAE/(kg/ha)	MAPE/%
RF	0.83	66.44	50.10	2.47
MLP	0.85	62.96	48.42	2.36
SVR	0.82	67.78	51.68	2.52
Ridge Regression	0.82	69.25	47.22	2.28
Stacking Model	0.87	60.04	47.14	2.29

Note: variables are defined as follows—R²: determination coefficient (measures the model's ability to interpret data variations); RMSE: root mean square error (reflects the standard deviation of prediction errors); MAE: mean absolute error (reflects the mean of absolute prediction errors); MAPE: mean absolute percentage error (reflects the mean of relative prediction errors).

The results show that the R² of each model exceeds 0.82, with the Stacking model exhibiting the best performance. In a single model, the RF has a strong generalization ability due to the decision tree integration strategy, but it lacks the ability to capture nonlinear responses to extreme meteorological events. Although MLP is good at nonlinear fitting,

it is easy to overfit under small samples. SVR and ridge regression are insufficient in high-dimensional feature processing. The Stacking model integrates the predictions of the base models via the ridge regression meta-model with equilibrium deviation and variance, and its RMSE is significantly lower than that of the MLP, which verifies the advantages of ensemble learning in scenarios involving complex meteorological factor coupling.

3.4. Advance Prediction of Flue-Cured Tobacco Yield

In view of the phenological characteristics of the maturation of flue-cured tobacco in September in the study area, based on the 2003–2020 data set, this study constructed a yield prediction model by combining polynomial regression and the Stacking ensemble model with monthly meteorological data. By analyzing the dynamic changes in prediction accuracy across different months during the growth period, the optimal lead time for yield prediction was determined (Table 3).

Table 3. Comparison of prediction accuracy of flue-cured tobacco yield across different months.

Month	MAPE/%	MAE/(kg/ha)	RMSE/(kg/ha)	R ²
September (Original features)	2.29	47.14	60.04	0.87
August	2.92	60.38	76.83	0.78
July	3.26	67.18	90.88	0.69
June	4.09	84.11	109.40	0.55
May	4.80	98.49	130.71	0.36
April	4.21	86.93	116.61	0.49

From the perspective of prediction accuracy dynamics, the model accuracy showed a gradual optimization trend with the progression of the growth period: the prediction performance was the worst in the early stage of the vigorous growth period (May), improved in July, and was significantly optimized after entering the maturity period (August–September). Among them, the accuracy was the highest in September (maturity period), and the accuracy in August was close to that in September.

The results of the verification of early prediction performance showed that the Stacking model performed best when it was approximately one month before the maturity period (August, the end of the vegetative growth stage). At this time, the model R² reached 0.78, the MAPE was 2.92%, and the error level was close to the prediction results of the maturity period (September) (MAPE = 2.29%, R² = 0.87). This result confirms that August can be used as the best lead time for short-term forecasts of flue-cured tobacco production. This model can effectively capture the correlation characteristics of meteorological factors and yield formation during the fertility period; provide reliable decision-making support for production areas to formulate baking scheduling, disaster prevention and mitigation, and other management measures in advance; and has significant practical application value.

3.5. Verification and Analysis of the Model's Prediction Performance

In order to further verify the actual predictive effectiveness of the model, this study selected flue-cured tobacco yield data from four typical regions in Yunnan Province (2021–2023) for independent sample validation and systematically compared the deviation characteristics of the predicted yield and the actual yield (Table 4).

Table 4. Actual and predicted yield results of flue-cured tobacco in selected regions of Yunnan Province (2021–2023).

Region	Year	Actual Yield/(kg/ha)	Predicted Yield/(kg/ha)	MAPE/%
Chuxiong	2021	2171.25	2145.43	1.19
	2022	2049.19	2061.93	0.62
	2023	2163.27	2107.60	2.57
Honghe	2021	2125.00	2101.64	1.10
	2022	2116.62	2134.76	0.86
	2023	2082.89	1989.98	4.46
Kunming	2021	2033.80	2031.57	0.11
	2022	2100.72	2118.14	0.83
	2023	2389.69	2228.91	6.73
Baoshan	2021	2087.28	2142.60	2.65
	2022	2043.90	2136.56	4.53
	2023	2071.77	2056.22	0.75

The results showed that the overall consistency between the predicted yield and the actual yield in each region was high. The Chuxiong area exhibited low prediction errors during the validation period, with the best accuracy in 2022 and a slight increase in error in 2023 due to abnormal fluctuations in meteorological conditions; the Honghe area had controllable overall error during the verification period, with the smallest error in 2021 and a significant increase in error in 2023 due to regional drought events that caused the actual yield to deviate from the conventional fluctuation range. The forecast error in Kunming showed significant inter-annual differences. The MAPE in 2021 was only 0.11%, and the forecast results were close to the actual value. In 2023, due to extreme high-temperature weather, the actual yield was abnormally high, resulting in the MAPE rising to 6.73%. The forecast stability in the Baoshan area is relatively good, with a three-year MAPE between 0.75% and 4.53%. In 2022, there was a lag in the model’s response to local flooding events, and the error is relatively high.

Overall, except for Kunming in 2023, the prediction error (MAPE) of the model in each region and year is controlled within 5%, indicating that the model has good applicability and stability in the actual prediction of tobacco production, and can effectively support the dynamic monitoring and accurate early warning of regional tobacco production.

3.6. SHAP Feature Contribution Analysis

In order to accurately analyze the driving mechanism of the yield results of the flue-cured tobacco yield prediction model according to meteorological factors, this study uses the SHAP method to conduct a quantitative analysis of feature importance. As the core indicator of model interpretability analysis, the SHAP value can quantitatively disassemble the contribution of each meteorological factor to the predicted results, clearly define its influence direction and intensity, and provide a theoretical basis for the in-depth explanation of the response relationship of “meteorological conditions–production formation” [36].

3.6.1. Analysis of Single Factor and Synergistic Effects

Based on the analysis of the distribution characteristics of the SHAP value, the discrete span of Day and Night Temperature Difference in August (TDIFF8) is significantly greater than that of other factors (Figure 5). It can be determined that it is the core meteorological variable that affects the formation of tobacco yield and model prediction accuracy. Its driving mechanism can be explained by the physiological process of photosynthesis. An

appropriate high temperature during the day can activate the key enzyme activities of photosynthetic carbon assimilation and promote the generation of photosynthetic products; a moderate low temperature at night can inhibit respiration and reduce material loss [37]. The synergistic effect of photothermal conditions dominated by TDIFF8 improves the efficiency of dry matter accumulation by optimizing the dynamic balance of “photosynthetic accumulation–respiratory consumption”, verifying the adaptability of the model to physiological mechanisms.

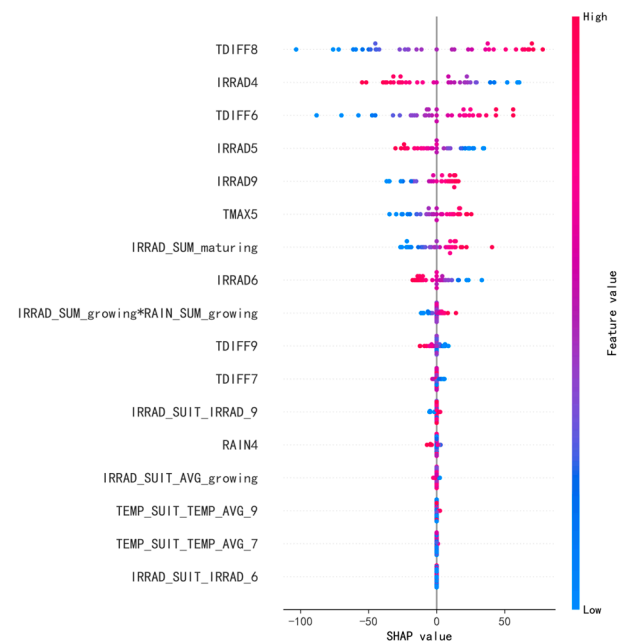


Figure 5. Feature importance analysis plot of the flue-cured tobacco yield prediction model based on SHAP. The x-axis represents SHAP values, indicating the feature contribution to prediction results (positive values for positive impact and negative values for negative impact). The y-axis shows meteorological and phenological variables input into the model, where variables in the format of “TDIFF + number” (e.g., TDIFF8, TDIFF6) denote day–night temperature difference in the corresponding month (the number represents the month). The meanings of other variable acronyms are as follows: IRRAD (solar irradiance), TMAX (daily maximum temperature), RAIN (precipitation), IRRAD_SUM_maturing (total solar irradiance during the maturing stage), IRRAD_SUIT_IRRAD_9 (suitable solar irradiance in September), IRRAD_SUIT_AVG_growing (average suitable solar irradiance during the growing stage), TEMP_SUIT_TEMP_AVG_9/7/6 (average suitable temperature in September/July/June), and IRRAD_SUM_growing + RAIN_SUM_growing (total solar irradiance + precipitation during the growing stage). The color gradient from blue (low value) to red (high value) represents the variable value levels. This figure quantifies the magnitude and direction of each variable’s impact on flue-cured tobacco yield prediction.

The discrete characteristics of radiation factors such as April solar radiation (IRRAD4) and temperature factors such as May maximum temperature (TMAX5) are also prominent. Among them, radiation factors, as the energy source of photosynthesis, directly determine the amount of photosynthetic products; temperature and humidity factors indirectly affect the formation of yield by regulating physiological metabolism and material distribution processes [38–40]. The synergistic mechanism of photothermal resources is presented in a concrete way through the quantitative analysis of the model, laying the foundation for subsequent research on multi-factor coupling effects.

3.6.2. Stage Specificity of Fertility-Related Factors

The SHAP value of the accumulated meteorological factors during the growth period represented by IRRAD_SUM_maturing (the accumulation amount of radiation in maturity)

shows significant fertility stage specificity. When the radiation conditions are suitable, the transport and accumulation of photosynthetic products to the blade can be accelerated. At this time, the SHAP value is positive; if the radiation is abnormal, it will interfere with the distribution and accumulation process of substances, resulting in negative fluctuations in the SHAP value. The above phenomenon verifies the model's ability to capture the cumulative effect of key meteorological factors during the fertility period, which is consistent with the modeling logic of “focusing on the environment in the fertility period–yield response relationship”. From a methodological perspective, this also proves the scientific nature of feature screening (selecting relevant factors for the growth period such as transplantation, prosperity, maturity, etc.) and model architecture (adapting to nonlinear response relationships), ensuring that the analysis results are consistent with the physiological laws of flue-cured tobacco growth.

3.6.3. Global Quantification and Verification of Feature Importance

With the help of the SHAP feature importance map, the explanatory power of TDIFF8 on yield prediction is the most prominent, confirming the core regulatory value of the stability of the day–night temperature difference in flue-cured tobacco maturity period (Figure 6). The stable temperature difference can optimize the photothermal synergy effect and promote dry matter accumulation. The influence of IRRAD4 (April solar radiation) is secondary, and April corresponds to the transplanted rooting period. Adequate radiation can accelerate seedling establishment and promote root development. Insufficient radiation will delay the growth rhythm and reduce stress resistance. TDIFF6 (Day and Night Temperature Difference in June) and IRRAD6 (June solar radiation) jointly regulate the long-term growth rhythm: abnormal temperature differences often induce leaf deformity and insufficient radiation will limit the accumulation of photosynthetic products. The two are related to the final yield by affecting the growth rate and leaf development.

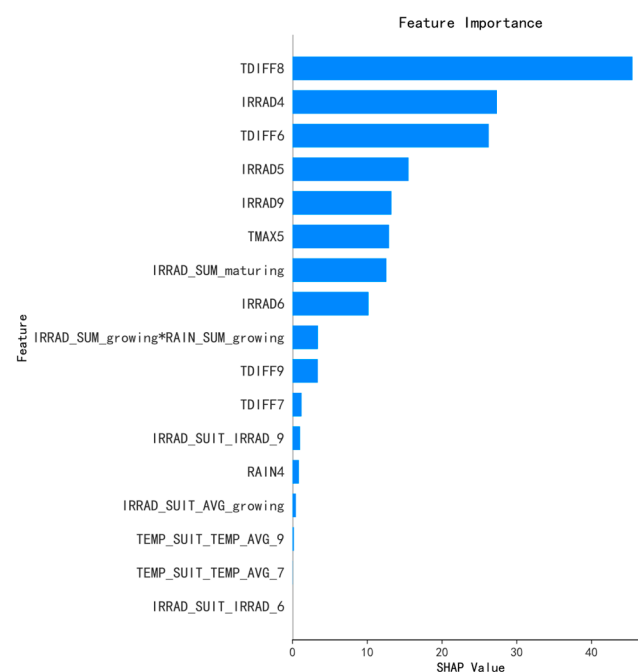


Figure 6. SHAP feature importance.

In addition, features such as IRRAD5 (May solar radiation) and IRRAD_SUM_maturing (total solar radiation accumulation during the maturity stage) have also been included in the core influence collection, and from the dimensions of post-transplantation growth connection and maturity material accumulation, it confirms the multi-fertility driving effect

of meteorological factors on yield. The action law of factors revealed by SHAP analysis in this study forms mutual verification with the model prediction results. Through the inversion verification of an abnormal year (for an abnormal TDIFF8 year, the deviation rate of the predicted yield and actual yield of the model is higher than that of conventional years), the effectiveness of the analysis is further verified, which is consistent with the conclusion that “extreme temperature difference leads to negative contribution” in the SHAP graph.

4. Discussion

The coupling method of polynomial regression and the Stacking model proposed in this study provides a valuable approach to alleviate the inherent limitations of single models in analyzing the nonlinear relationship between meteorological factors and flue-cured tobacco yield. In the yield decomposition experiment, polynomial regression exhibited the best performance in trend yield fitting; its smooth long-term trend curve can accurately capture the cumulative effect of agricultural technological progress, laying a reliable data foundation for the precise separation of meteorological yield. This method is highly consistent with Ji et al.’s [41] conclusion that “trend yield should accurately reflect the long-term impact of technological progress,” further verifying the universality of polynomial regression in agricultural yield trend extraction. It also offers insights for the improvement of Ma et al.’s [15] crop yield time series prediction framework; while Ma et al.’s framework focuses on the overall prediction process, the optimization of the trend extraction step in this study can provide references for the application of this framework to subtropical plateau crops. Compared with the moving average method, exponential smoothing method, and high-pass filtering method, polynomial regression effectively avoids the over-sensitivity of the moving average method to short-term fluctuations and the trend segmentation bias of the high-pass filtering method caused by frequency domain separation [26], making it more aligned with the long-term evolution law of “technological iteration–yield increase” in the flue-cured tobacco industry. Additionally, this study focuses on the stage-specific effects of meteorological factors during the growth period, expanding Zhao et al.’s [14] linear regression-based analysis of meteorological–yield relationships. By breaking through the limitations of linear models, it captures the nonlinear coupling effects of multi-stage factors.

The feature selection results reveal the multi-dimensional driving mechanism of meteorological factors on flue-cured tobacco yield. Temperature-derived features and light suitability features during the growth period contribute significantly to yield prediction, which is consistent with the biological characteristic of flue-cured tobacco—“preferring temperature and light but being intolerant to extreme high temperatures.” This mechanism can be supported by Didari et al.’s [17] Lasso regression study on dryland wheat yield: Didari et al. confirmed the key role of extreme temperature and precipitation events in yield estimation, while this study further identifies the growth stage of flue-cured tobacco sensitive to extreme factors (extreme temperature differences in the August maturity stage may affect yield), which can provide references for studies on crop-specific meteorological responses. Critically, the inclusion of interaction terms between growing season solar radiation (IRRAD_growing) and precipitation (RAIN_growing) in the core feature set confirms that the photo–water coupling effect is a key mechanism regulating flue-cured tobacco yield formation. This finding echoes Tang et al.’s [7] conclusion in their Honghe flue-cured tobacco study that “climatic factors affect yield and quality through synergistic effects,” and realizes an extension at the quantitative level—even under drought conditions, it can still capture synergistic effects well. Meanwhile, this result is also consistent with Guo et al.’s [34] view that “multi-factor coupling operations can capture joint meteorological impacts”. Guo et al.’s study focuses on drought dynamics, while this study applies this

concept to yield prediction scenarios, providing a new scenario for the application of multi-factor coupling in agricultural prediction. Furthermore, significant regional differences in model applicability were observed: the model performed best in the Chuxiong tobacco-growing area (average MAPE = 1.46% during 2021–2023) due to the stable local climate; the relatively high prediction error in Kunming in 2023 (MAPE = 6.73%) was mainly associated with extreme high temperatures (August average temperature higher than normal) and the inability of NASA POWER data ($0.5^\circ \times 0.5^\circ$ spatial resolution) to reflect mountain microclimate heterogeneity. Compared with the winter wheat studied by Zhang et al. [30], flue-cured tobacco, as a leaf crop, is more sensitive to temperature–humidity synergy during the maturity stage. This difference stems from the inherent biological characteristics of crops and can provide a basis for comparative studies on meteorological response laws of different crops.

Model comparison experiments fully verify the advantages of ensemble learning in modeling complex agricultural systems. By integrating the algorithmic characteristics of RF, MLP, SVR, and ridge regression, the Stacking model significantly outperforms single models in key performance metrics. This result is consistent with Islam et al.'s [27] findings in rice yield prediction, that ensemble models can improve prediction stability by balancing the bias and variance of base models. The difference between the two lies in the following: Islam et al.'s model required remote sensing data to achieve an R^2 of 0.85, while the model in this study achieved a higher accuracy using only meteorological data. This characteristic highlights its efficiency advantage in meteorology-driven yield prediction and helps reduce the concerns about “relatively high remote sensing data acquisition costs” mentioned by Sharma et al. [24] in crop yield prediction studies.

During the verification period, the model showed a stable predictive performance at the regional scale, but relatively high errors were observed in local production areas. This difference can be attributed to three clear limitations: (1) Non-meteorological factors such as soil physical and chemical properties, field management measures, and biological stresses were not included. This situation echoes Bukowiecki et al.'s [42] finding that “soil texture and green area index (GAI) can improve wheat yield estimation accuracy”, suggesting that the supplementation of non-meteorological factors may provide room for model accuracy improvement. (2) The insufficient spatial resolution of meteorological data, as the $0.5^\circ \times 0.5^\circ$ resolution of NASA POWER data makes it difficult to capture the small-scale heterogeneity of Yunnan's “three-dimensional climate”, which is consistent with Lecerf et al.'s [13] concern that “coarse-resolution meteorological data may limit the spatial precision of crop yield prediction”. (3) The adaptability to extreme climates beyond historical data needs to be enhanced, as the model's training data (2003–2020) did not include events such as “20 consecutive days of drought” or “temperatures exceeding 40°C ”, leading to increased prediction errors during the 2023 extreme high-temperature event in Kunming. This limitation is consistent with the phenomenon of “limited response ability of single algorithms to extreme climates” found by Luo et al. [25] in maize drought prediction, suggesting that the introduction of extreme climate indices may help optimize model performance. Based on the above observations, corresponding future research directions may include: (1) Fusing high-resolution remote sensing data and IoT data to supplement non-meteorological information. (2) Introducing extreme climate indices to optimize Stacking base models. (3) Exploring the application potential of this framework in subtropical plateau crops such as maize and rapeseed by adjusting growth stage divisions and core features.

5. Conclusions

Based on the flue-cured tobacco yield and meteorological data of Yunnan Province from 2003 to 2023, this study constructed a coupled framework of polynomial regression and a Stacking ensemble model ($R^2 = 0.87$). Its core contribution lies in providing a new methodological path of “trend separation–ensemble prediction” for subtropical plateau crop yield prediction, while offering a practical basis for precise flue-cured tobacco cultivation by analyzing the physiological regulatory mechanisms of key meteorological factors (e.g., day–night temperature difference in the August maturity stage, solar radiation in the April transplanting stage). The framework not only provides empirical support for studying the meteorology–crop yield interaction mechanism, but also facilitates the formulation of disaster prevention and mitigation plans in advance for Yunnan’s tobacco-growing areas. It should be noted that the framework currently relies solely on meteorological data (without integrating non-meteorological factors such as soil physicochemical properties and field management), making it more suitable for Yunnan and climatically similar subtropical plateau tobacco-growing areas; the future integration of multi-source data can further enhance its applicability.

Author Contributions: Conceptualization, Y.W. and M.Z.; Methodology, Y.W. and M.Z.; Software, Y.W.; Validation, Y.W.; Formal Analysis, Y.W.; Investigation, X.J.; Resources, Y.W. and M.Z.; Data Curation, Y.W. and M.Z.; Writing—original draft preparation, Y.W.; Writing—review and editing, B.Z. and X.B.; Visualization, Y.W.; Supervision, X.B.; Project Administration, J.Z.; Funding Acquisition, B.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Major Science and Technology Special Project of Yunnan Province titled “Integrated Research on Key Technologies of Smart Agriculture”, grant number 202302AE09002003.

Data Availability Statement: The data presented in this study can be obtained from the following public resources: Basic statistical data are from the Statistical Yearbooks of Yunnan Province (2004–2024) published on the Official Website of Yunnan Provincial Bureau of Statistics, China (<https://stats.yn.gov.cn/>); Meteorological data are from the NASA POWER Database (<https://power.larc.nasa.gov/>). The above data are directly accessible in the public domain and have no additional reference numbers. If further clarification on the scope of data extraction or parameters is needed, please feel free to contact the author for supplementary information.

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

References

1. Wang, Z.; Zheng, Q.; Gartner, C.; Chan, G.C.K.; Ren, Y.; Wang, D.; Thai, P.K. Comparison of tobacco use in a university town and a nearby urban area in China by intensive analysis of wastewater over one year period. *Water Res.* **2021**, *206*, 117733. [CrossRef] [PubMed]
2. Li, F.; Wang, Y.L.; Zhang, J.D.; Lu, Y.C.; Zhu, X.; Chen, X.Y.; Yan, J.J. Toxic metals in top selling cigarettes sold in China: Pulmonary bioaccessibility using simulated lung fluids and fuzzy health risk assessment. *J. Clean. Prod.* **2020**, *275*, 124131. [CrossRef]
3. State Tobacco Monopoly Administration. Tobacco Industry Achieves Record-High Total Tax Profit and Fiscal Revenue in 2024. Available online: <http://www.tobacco.gov.cn/gjyc/hyyw/202503/9e95b1a1bde1498aa3799495a199ff6d.shtml> (accessed on 6 March 2025).
4. National Bureau of Statistics of China. Profits of Industrial Enterprises Above Designated Size Nationwide Decreased by 3.3% in 2024. Available online: https://www.stats.gov.cn/sj/zxfb/202501/t20250127_1958485.html (accessed on 27 January 2025).
5. State Taxation Administration of China. Annual Report on China’s Taxation. 2024. Available online: <https://www.chinatax.gov.cn/chinatax/c102292/c5240191/5240191/files/1cd3cd9a33b14969a3e37dd87707029b.pdf> (accessed on 25 June 2025).
6. Yunnan Provincial Bureau of Statistics. Yunnan Statistical Yearbook 2024. Available online: http://stats.yn.gov.cn/pages_22_6933.aspx (accessed on 23 May 2025).
7. Tang, Z.X.; Chen, L.L.; Chen, Z.B.; Fu, Y.L.; Sun, X.L.; Wang, B.B.; Xia, T.Y. Climatic factors determine the yield and quality of Honghe flue-cured tobacco. *Sci. Rep.* **2020**, *10*, 19868. [CrossRef] [PubMed]

8. Wei, T.P.; Sun, Z.B.; Huang, J.C.; Zou, C.Q. Methods of Causal Inference in Surface Process Research: A Review and Experiment. *Sci. Geogr. Sin.* **2025**, *45*, 1986–1999. [CrossRef]
9. Kishk, A.; Chang, X.H.; Wang, D.M.; Wang, Y.J.; Yang, Y.S.; Zhao, G.C.; Tao, Z.Q. Evolution of varieties and development of production technology in Egypt wheat: A review. *J. Integr. Agric.* **2019**, *18*, 483–495. [CrossRef]
10. Kelly, C.I.; Boateng, E.F.; Zibrila, A.; Andam-Akorful, S.A.; Quaye-Ballard, J.A.; Laari, P.B.; Damoah-Afari, P. Understanding hydrometeorological conditions and their relationship with crop production in the upper east region, Ghana. *Agric. Water Manag.* **2025**, *312*, 109434. [CrossRef]
11. Sainju, U.M.; Liptzin, D.; Jabro, J.D. Relating soil physical properties to other soil properties and crop yields. *Sci. Rep.* **2022**, *12*, 22025. [CrossRef]
12. Andualem, Z.A.; Meshesha, D.T.; Hassen, E.E. The impacts of watershed management practices on crop yield potential in Yezat Watershed, North West, Ethiopia. *Environ. Sci. Pollut. Res.* **2025**, *32*, 16395–16412. [CrossRef]
13. Lecerf, R.; Ceglar, A.; López-Lozano, R.; Van Der Velde, M.; Baruth, B. Assessing the information in crop model and meteorological indicators to forecast crop yield over Europe. *Agric. Syst.* **2019**, *168*, 191–202. [CrossRef]
14. Zhao, Y.; Zheng, R.; Zheng, F.; Zhong, K.; Fu, J.; Zhang, J.; Flanagan, D.C.; Xu, X.; Li, Z. Spatiotemporal distribution of agrometeorological disasters in China and its impact on grain yield under climate change. *Int. J. Disaster Risk Reduct.* **2023**, *95*, 103823. [CrossRef]
15. Ma, P.; Zhang, N.; Yang, Y.; Wang, Z.; Li, G.; Fu, Z. Sugarcane Yield Prediction in Chongzuo, Guangxi—An LSTM Model Based on the Fusion of Trend Yield and Meteorological Yield. *Agronomy* **2024**, *14*, 2512. [CrossRef]
16. Bognár, P.; Kern, A.; Pásztor, S.; Steinbach, P.; Lichtenberger, J. Testing the Robust Yield Estimation Method for Winter Wheat, Corn, Rapeseed, and Sunflower with Different Vegetation Indices and Meteorological Data. *Remote Sens.* **2022**, *14*, 2860. [CrossRef]
17. Didari, S.; Talebnejad, R.; Bahrami, M.; Mahmoudi, M.R. Dryland farming wheat yield prediction using the Lasso regression model and meteorological variables in dry and semi-dry region. *Stoch. Environ. Res. Risk Assess.* **2023**, *37*, 3967–3985. [CrossRef]
18. Zhao, C.; Liu, B.; Xiao, L.; Hoogenboom, G.; Boote, K.J.; Kassie, B.T.; Pavan, W.; Shelia, V.; Kim, K.S.; Hernandez-Ochoa, I.M.; et al. A SIMPLE crop model. *Eur. J. Agron.* **2019**, *104*, 97–106. [CrossRef]
19. Sun, F.; Mejia, A.; Che, Y. Disentangling the Contributions of Climate and Basin Characteristics to Water Yield Across Spatial and Temporal Scales in the Yangtze River Basin: A Combined Hydrological Model and Boosted Regression Approach. *Water Resour. Manag.* **2019**, *33*, 3449–3468. [CrossRef]
20. Wang, J.; Liu, W.; Yin, D. Impacts of Integrated Meteorological and Agricultural Drought on Global Maize Yields. *Agric. Water Manag.* **2025**, *318*, 109727. [CrossRef]
21. Meroni, M.; Waldner, F.; Seguini, L.; Kerdiles, H.; Rembold, F. Yield forecasting with machine learning and small data: What gains for grains? *Agric. For. Meteorol.* **2021**, *308–309*, 108555. [CrossRef]
22. Eddamiri, S.; Bassine, F.Z.; Ongoma, V.; Epule, T.E.; Chehbouni, A. An automatic ensemble machine learning for wheat yield prediction in Africa. *Multimed. Tools Appl.* **2024**, *83*, 66433–66459. [CrossRef]
23. Shahhosseini, M.; Hu, G.P.; Archontoulis, S.V. Forecasting Corn Yield With Machine Learning Ensembles. *Front. Plant Sci.* **2020**, *11*, 1120. [CrossRef]
24. Sharma, P.; Dadheech, P.; Aneja, N.; Aneja, S. Predicting Agriculture Yields Based on Machine Learning Using Regression and Deep Learning. *IEEE Access* **2023**, *11*, 111255–111264. [CrossRef]
25. Luo, Y.; Wang, H.; Cao, J.; Li, J.; Tian, Q.; Leng, G.; Niyogi, D. Evaluation of machine learning-dynamical hybrid method incorporating remote sensing data for in-season maize yield prediction under drought. *Precis. Agric.* **2024**, *25*, 1982–2006. [CrossRef]
26. Lin, S.; Liang, Z.; Zhao, S.; Dong, M.; Guo, H.; Zheng, H. A comprehensive evaluation of ensemble machine learning in geotechnical stability analysis and explainability. *Int. J. Mech. Mater. Des.* **2024**, *20*, 331–352. [CrossRef]
27. Islam, M.D.; Di, L.; Qamer, F.M.; Shrestha, S.; Guo, L.; Lin, L.; Mayer, T.J.; Phalke, A.R. Rapid Rice Yield Estimation Using Integrated Remote Sensing and Meteorological Data and Machine Learning. *Remote Sens.* **2023**, *15*, 2374. [CrossRef]
28. Yunnan Provincial Bureau of Statistics. Yunnan Statistical Yearbook. Available online: <https://stats.yn.gov.cn/> (accessed on 23 May 2025).
29. Sun, X.F.; Huang, Z.Y.; Xie, M.E.; Xie, X.Q.; Dai, K. Climatic causes of planting on chemical quality and style characteristics of flue-cured tobacco in Yunnan. *Chin. J. Eco-Agric.* **2025**, *33*, 1371–1382.
30. Zhang, Z.; Zhou, N.; Xing, Z.; Liu, B.; Tian, J.; Wei, H.; Gao, H.; Zhang, H. Effects of Temperature and Radiation on Yield of Spring Wheat at Different Latitudes. *Agriculture* **2022**, *12*, 627. [CrossRef]
31. Gao, B.W.; Ma, Y.Z.; Yang, P.G.; Fu, Y.H.; Dong, B.D.; Zhou, Y.F.; Chen, Q.R.; Qiao, Y.Z. Enhanced early growth rates in high cumulative temperature requirement maize (*Zea mays* L.) varieties drive superior production potential in rainfed North China Plain. *J. Agric. Food Res.* **2025**, *22*, 102044. [CrossRef]
32. Tendeng, B.; Asselin, H.; Imbeau, L. Moose (*Alces americanus*) habitat suitability in temperate deciduous forests based on Algonquin traditional knowledge and on a habitat suitability index. *Écoscience* **2016**, *23*, 77–87. [CrossRef]

33. Monish, N.T.; Rehana, S. Suitability of distributions for standard precipitation and evapotranspiration index over meteorologically homogeneous zones of India. *J. Earth Syst. Sci.* **2020**, *129*, 25. [\[CrossRef\]](#)
34. Guo, W.; Huang, S.; Zhao, Y.; Li, X.; Zhang, Q. Quantifying the effects of nonlinear trends of meteorological factors on drought dynamics. *Nat. Hazards* **2023**, *117*, 2505–2526. [\[CrossRef\]](#)
35. Han, Y.J. Analysis of Spatiotemporal Characteristics of Climate Change and Construction of Black Shank Early Warning Model in Yunnan Tobacco-Growing Areas. Master's Thesis, Chinese Academy of Agricultural Sciences, Beijing, China, 2024.
36. Wang, Y.; Wang, P.X.; Tansey, K.; Liu, J.M.; Delaney, B.; Quan, W.T. An Interpretable Approach Combining Shapley Additive Explanations and LightGBM Based on Data Augmentation for Improving Wheat Yield Estimates. *Comput. Electron. Agric.* **2025**, *229*, 109758. [\[CrossRef\]](#)
37. Li, Y.; Ren, K.; Hu, M.; He, X.; Gu, K.; Hu, B.; Su, J.; Jin, Y.; Gao, W.; Yang, D.; et al. Cold stress in the harvest period: Effects on tobacco leaf quality and curing characteristics. *BMC Plant Biol.* **2021**, *21*, 131. [\[CrossRef\]](#)
38. Zhang, L.H.; Shen, M.G.; Jiang, N.; Lv, J.X.; Liu, L.C.; Zhang, L. Spatial Variations in the Response of Spring Onset of Photosynthesis of Evergreen Vegetation to Climate Factors Across the Tibetan Plateau: The Roles of Interactions Between Temperature, Precipitation, and Solar Radiation. *Agric. For. Meteorol.* **2023**, *335*, 109440. [\[CrossRef\]](#)
39. Ha, S.; Kim, Y.T.; Im, E.S.; Hur, J.; Jo, S.; Kim, Y.S.; Shim, K.M. Impacts of Meteorological Variables and Machine Learning Algorithms on Rice Yield Prediction in Korea. *Int. J. Biometeorol.* **2023**, *67*, 1825–1838. [\[CrossRef\]](#)
40. Lal, N.; Kumar, A.; Pandey, S.D.; Sahu, N. Impact of Climatic Variability on Litchi Yields in the Tarai Region of India. *Appl. Fruit Sci.* **2024**, *66*, 2371–2374. [\[CrossRef\]](#)
41. Ji, Y.H.; Zhou, G.S.; Wang, L.X.; Wang, S.D.; Li, Z.S. Identifying climate risk causing maize (*Zea mays* L.) yield fluctuation by time-series data. *Nat. Hazards* **2019**, *96*, 1213–1222. [\[CrossRef\]](#)
42. Bukowiecki, J.; Rose, T.; Kage, H. Assessment of the Impact of Accurate Green Area Index, Water Regime and Harvest Index on Site-Specific Wheat Yield Estimation. *Comput. Electron. Agric.* **2024**, *226*, 109429. [\[CrossRef\]](#)

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.