

Article

A Multi-Temporal Agricultural Remote Sensing Framework for Sustainable Crop Yield Estimation with Economic Impact

Shengyuan Tang ¹, Chenlu Jiang ^{1,2}, Jingdan Zhang ², Mingran Tian ¹, Yang Zhang ¹, Yating Yang ¹ and Min Dong ^{2,*}

¹ National School of Development, Peking University, Beijing 100871, China

² China Agricultural University, Beijing 100083, China

* Correspondence: dongmin@cau.edu.cn

Abstract

Under the intensifying impacts of climate change, tightening agricultural resource constraints, and escalating food security pressures, the development of high-accuracy and interpretable crop yield estimation methods has become a critical technical issue in sustainable agricultural engineering. In this study, multi-temporal and multi-spectral remote sensing imagery are utilized as the core input. A multi-scale visual feature extraction module is designed to characterize canopy texture, field structure, and regional heterogeneity, while a temporal growth modeling module captures the dynamic evolution of crops from emergence to maturity. Yield regression is further integrated with economic mapping and explainability mechanisms, thereby forming an end-to-end prediction framework. Experimental results across multiple regions and years demonstrate that the proposed method outperforms various representative models. In the primary regression experiment, the framework achieves approximately $R^2 = 0.76$, with MAE reduced to 0.60 and MSE to 0.62, representing an error reduction of over 25% compared with traditional regression approaches and classical machine learning models. In classification experiments for yield-grade evaluation, the model attains an accuracy of approximately 0.85, with both precision and recall exceeding 0.82, demonstrating its effectiveness in both continuous yield prediction and stable yield-level region identification. Cross-region and cross-year validation further indicate strong generalization capability, with R^2 remaining above 0.65 in unseen regions and around 0.67 under cross-year prediction settings. Ablation studies confirm the synergistic contributions of multi-scale spatial modeling, temporal growth modeling, and explainability constraints, as performance consistently declines when any individual module is removed. Overall, the results highlight that the proposed framework provides reliable data support for precision agricultural management, resource optimization, and agricultural engineering decision-making, while also offering a scalable and reproducible pathway for sustainable agricultural engineering development.



Academic Editors: Nikolay Kiktev,
Taras Lendiel and Oleksii Opryshko

Received: 21 January 2026

Revised: 26 February 2026

Accepted: 27 February 2026

Published: 3 March 2026

Copyright: © 2026 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 \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: sustainable agriculture; multi-temporal remote sensing; agricultural engineering for sustainable development; crop yield estimation; explainable artificial intelligence; agro-chemical sustainability

1. Introduction

Under the intensifying impacts of global climate change, population growth, and increasing constraints on arable land and water resources, agricultural systems face the critical challenge of sustaining yield growth under mounting resource and environmental

pressure [1]. Agricultural yield is closely linked to food security and sustainable development and serves as a key indicator influencing farmers' income, regional disparities, and agricultural policy formulation [2]. High-accuracy, spatially explicit yield estimation is therefore essential for precision agriculture, resource optimization, and agricultural modernization [3]. However, yield data are often difficult to obtain, updated slowly, and limited in spatial resolution, restricting fine-grained decision-making and long-term monitoring [4]. Constructing a system capable of accurately characterizing large-scale, continuous spatiotemporal yield patterns thus holds significant theoretical and practical value for sustainable agriculture [5].

From the perspective of sustainable agricultural engineering, yield formation is regulated not only by climate and management but also by agro-chemical processes such as soil nutrient cycling, fertilizer utilization efficiency, and crop biochemical metabolism. Variations in nitrogen availability, chlorophyll synthesis, and water-carbon interactions directly affect productivity and environmental sustainability. Integrating agro-chemical indicators into engineering-oriented yield estimation frameworks is therefore crucial for promoting resource-efficient and environmentally sustainable agricultural development.

Traditional yield estimation methods rely on statistical surveys, field measurements, and empirical models [6]. Although ground-based approaches achieve high accuracy at small scales, they are costly, time-consuming, and unsuitable for large-area applications, while failing to reflect spatial heterogeneity [7]. Remote sensing-based empirical regression and vegetation index methods such as NDVI and EVI enable large-scale estimation but assume simplified relationships and struggle to capture complex nonlinear growth dynamics [8]. Because yield is jointly influenced by climate, soil, and management with strong regional and interannual variability, traditional models often exhibit poor cross-regional and cross-temporal generalization [9]. These limitations constrain the practical value of remote sensing-based yield estimation in precision agriculture and policy analysis [10].

Advances in remote sensing and computing power have enabled deep learning-based yield estimation [11], which has been widely applied in crop type identification, growth monitoring [12], and disaster assessment [13]. Nevertheless, high-precision prediction of continuous variables such as remote sensing imagery-crop yield remains relatively underexplored [14]. Some studies rely on single-temporal imagery and neglect growth dynamics [15], while others emphasize predictive accuracy but overlook interpretability and socioeconomic implications [16], limiting their value for decision support and social science research [17]. Representative studies include Li et al. [18], Jian Li et al. [19], Zhangxin Liu et al. [20], Yedukondalu N. et al. [21], and Sungwook Yoon [22].

Building upon this background, a comprehensive remote sensing and deep learning-driven yield estimation framework integrating multi-temporal information is proposed. As shown in Figure 1, deep learning models automatically extract crop growth features from multi-source, multi-scale imagery to establish high-precision mappings between remote sensing features and yield, enabling spatially continuous prediction. The estimated yields are further linked with farmers' income and regional economic indicators to evaluate the socioeconomic implications of spatial variability. This design improves predictive accuracy while extending the practical value of remote sensing-based yield estimation within sustainable agriculture. Agro-chemical and biochemical indicators are also incorporated to interpret productivity from the perspective of nutrient management and resource-use efficiency.

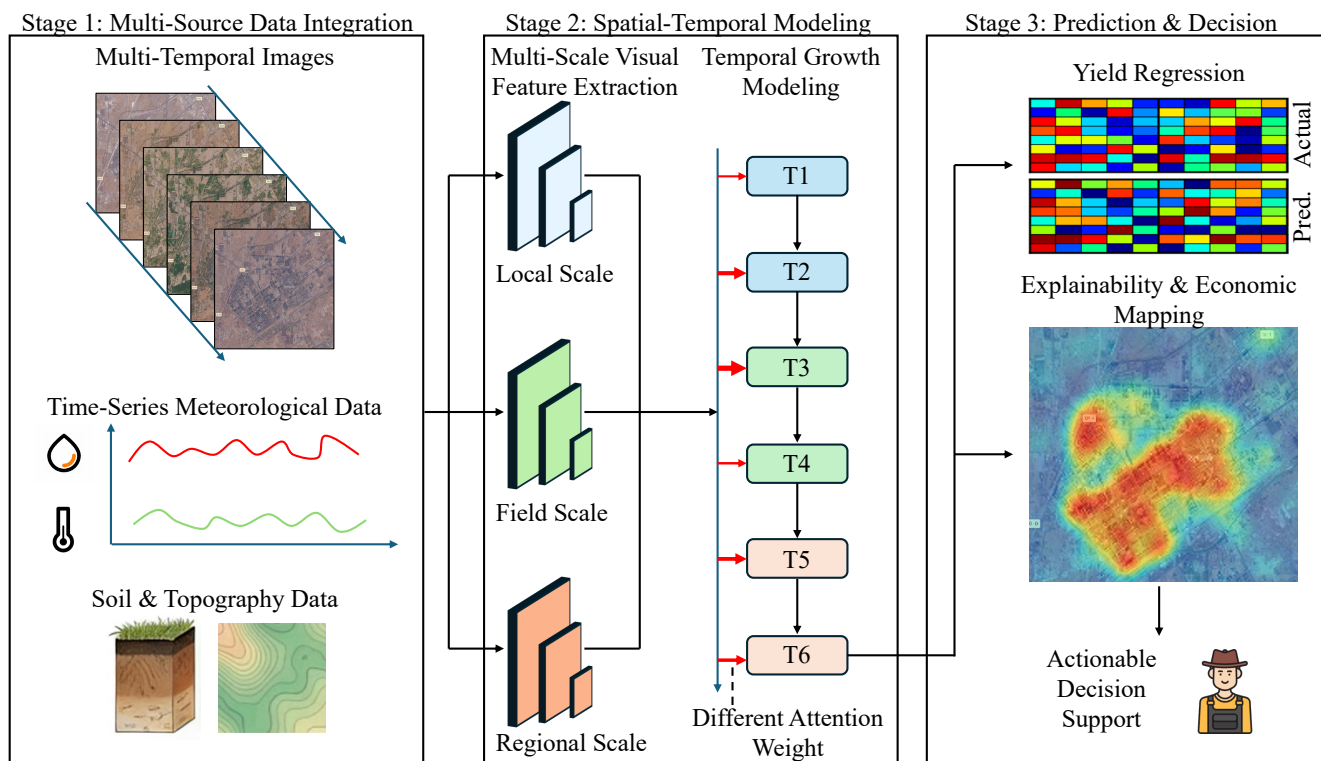


Figure 1. Overview of the proposed holistic yield estimation framework. The pipeline consists of three phases: Stage 1 integrates multi-temporal remote sensing imagery with meteorological and topological data; Stage 2 employs a multi-scale spatial feature extraction and temporal attention mechanism to model crop growth dynamics; Stage 3 generates high-precision yield regression maps and interpretable economic insights to provide actionable decision support for agricultural management.

The main contributions of this study are as follows:

1. A deep learning-based yield estimation method integrating multi-temporal remote sensing imagery is proposed, effectively capturing nonlinear crop growth dynamics and improving prediction accuracy and stability.
2. Multi-scale feature extraction and temporal modeling strategies enhance adaptability across regions, crop types, and growth conditions, thereby improving model generalization.
3. An interpretability analysis mechanism is introduced to quantify the contributions of growth stages and features to yield, facilitating transparent explanation of model predictions.
4. High-resolution yield outputs are integrated with farmers' income and regional development indicators to support agricultural policy formulation, resource optimization, and coordinated regional development.
5. The potential of intelligent yield estimation models in agricultural education, farmer training, and sustainable development curricula is explored, highlighting their broader societal value.

2. Related Work

2.1. Applications of Remote Sensing in Agricultural Yield Estimation

Remote sensing-based agricultural yield estimation is grounded in the principle that crop growth status can be indirectly characterized through spectral information [23]. Across growth stages, variations in leaf area index, canopy structure, biomass accumulation, and water content substantially influence solar radiation absorption and reflection, particularly in the visible and near-infrared bands [24]. Periodic remote sensing observation therefore enables non-contact acquisition of spectral information reflecting crop conditions

for yield estimation [25]. Based on this mechanism, vegetation index methods became dominant in early studies, where indices such as NDVI and EVI were widely used to characterize crop growth and establish statistical relationships with final yield [26]. In practice, vegetation indices or multispectral features are treated as independent variables, with linear or multiple regression models applied for yield prediction [27]. These methods support regional and national yield trend analysis with relatively low computational cost [28]. However, several limitations have been identified. First, index saturation forms a critical bottleneck: NDVI loses sensitivity when LAI exceeds 3, reducing responsiveness to biomass variation in high-yield crops [29]. Second, the assumption that “greenness equals yield” fails during reproductive stages, when photosynthetic activity shifts to ears and leaf senescence occurs despite increasing grain weight [30]. Third, pixel-wise spectral ratios neglect spatial texture information such as planting density and row structure. Finally, climatic, soil, and management heterogeneity causes identical index values to correspond to different yield levels, limiting cross-regional generalization [31]. These constraints highlight the need for methods capable of modeling nonlinear spatio-temporal dynamics beyond simple spectral indices.

2.2. Deep Learning and Computer Vision in Agriculture

Deep learning has been increasingly applied in agricultural remote sensing, including crop type identification, planting structure mapping [32], growth monitoring, and disaster assessment [33]. Compared with traditional approaches, these methods demonstrate stronger feature representation and robustness in complex environments [34]. Recent advances further show their effectiveness in enhancing satellite spatial resolution for fine-grained agricultural characterization [35]. Nevertheless, yield estimation as a continuous prediction task remains less explored than classification-oriented applications [36]. Existing deep learning yield studies face two key limitations. First, spatial-temporal modeling is often structurally decoupled. CNN-based models emphasize static spatial textures but overlook cumulative growth dynamics [37], whereas recurrent models (e.g., LSTM) capture temporal sequences while losing field-level spatial context. Second, end-to-end models suffer from “black-box” interpretability issues. Although physics-informed designs have emerged in related domains such as soil moisture mapping [38], most yield models lack attribution mechanisms to verify whether predictions rely on agronomically meaningful features, reducing trust in decision-making contexts [39]. Developing unified, interpretable spatio-temporal frameworks therefore remains a critical challenge [40].

2.3. Integrated Studies from Agricultural Yield, Economic Development, and Education Perspectives

From economic and social science perspectives, agricultural yield is a key determinant of farmers’ income, food security, and regional development disparities [41]. Spatial yield distribution directly affects income stability, input returns, and regional economic structures, informing agricultural policy and planning [42]. Traditional analyses rely on statistical yearbooks and survey datasets, which, despite macro-level authority, lack spatial granularity and temporal continuity and are constrained by the Modifiable Areal Unit Problem (MAUP) [43]. Integrated studies still face notable theoretical gaps between technical metrics and socioeconomic applicability. First, most remote sensing yield models optimize statistical accuracy rather than economic decision value, limiting their utility for production function analysis [44]. Second, the “black-box” nature of deep learning hinders causal interpretation required for policy formulation [45]. Third, limited pedagogical visualization restricts their integration into agricultural education and training [46]. Addressing these gaps requires frameworks that combine high-accuracy prediction with interpretability to support economic analysis and educational practice.

3. Materials and Method

3.1. Data Collection

A comprehensive dataset integrating remote sensing imagery, ground-based yield data, and meteorological and soil information was constructed to support the training and analysis of the yield estimation model, and the data were mainly collected from Wuyuan County, Bayannur City, Inner Mongolia, as well as from online sources, as shown in Table 1. Remote sensing data included multispectral satellite imagery from Sentinel-2 and Landsat 8/9, and multi-temporal images covering the main growing seasons of the study area were acquired, as illustrated in Figure 2. The temporal range covered key growth stages of crops, including sowing, emergence, jointing, heading, and maturity. The spatial resolution of the imagery ranged from 10–30 m and included visible, near-infrared, and shortwave infrared bands, which can be used to characterize canopy structure, changes in leaf area index, and biomass accumulation. All images were processed through radiometric calibration, atmospheric correction, and geometric correction, and cloud and shadow pixels were removed using cloud detection and de-clouding algorithms to ensure data quality and temporal continuity.

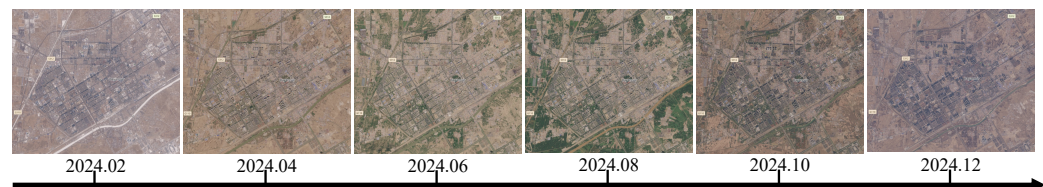


Figure 2. Examples of multispectral satellite imagery acquired by Sentinel-2 and Landsat-8/9 over Wuyuan County, Bayannur City, Inner Mongolia, China. The figure shows true-color images acquired in February, April, June, August, October, and December 2024.

Ground-based yield data were organized at the township or village level and included information on crop type, sown area, and total yield. Field-measured data were obtained by setting sampling plots in representative fields, and actual yield values were acquired through manual harvesting, weighing, and moisture content correction, which were used to supplement and validate the statistical data. All yield data were converted into unified units and were spatially located using plot boundaries or administrative vector data, so that accurate matching with remote sensing imagery could be achieved. To support the yield grade classification experiment mentioned in this study, the continuous yield data were further discretized into three categories based on the regional yield distribution: “Low Yield” (<6000 kg/ha), “Medium Yield” (6000–9000 kg/ha), and “High Yield” (>9000 kg/ha). The dataset exhibits a balanced distribution among these classes, with approximately 30% classified as Low Yield, 45% as Medium Yield, and 25% as High Yield, ensuring that the model is trained and evaluated on a representative sample of yield levels. These variables also reflect agro-chemical sustainability conditions, including nutrient availability, soil fertility status, and fertilizer response potential.

Meteorological data mainly included daily variables such as temperature, precipitation, solar radiation, and relative humidity. Soil data were obtained from the national soil database and regional soil survey results and included soil type, organic matter content, texture, and water-holding capacity. These data were used to describe the environmental background of crop growth and to provide support for the economic and ecological interpretation of model results. All data were unified to the same spatial resolution as the remote sensing imagery through spatial interpolation or resampling, and temporal alignment methods were applied to match meteorological data with remote sensing phases and growth stages. Finally, a sample dataset with the core structure of “field–growth

stage–remote sensing features–environmental variables–yield” was formed, providing a reliable data foundation for subsequent model training and evaluation.

Table 1. dataset composition and scale statistics.

Data Type	Time Range	Data Scale
multi-temporal remote sensing imagery	growing seasons from 2019–2024	1250 images
ground-based yield statistics	2019–2024	3600 administrative units
field-measured yield samples	2021–2024	420 sampling plots
meteorological data	2019–2024	about 2200 days at daily scale
soil data	static data	one regional raster dataset

3.2. Data Preprocessing and Augmentation Strategy

Remote sensing imagery constitutes the core input for agricultural yield estimation, and its quality directly affects the reliability of learned features. Owing to complex imaging conditions, sensor discrepancies, and atmospheric disturbances, raw data often suffer from cloud contamination, noise interference, and spatial resolution inconsistency. Directly feeding such data into deep learning models can introduce systematic errors, causing misinterpretation of non-crop information and degrading prediction accuracy and generalization. Therefore, systematic preprocessing is essential to ensure effective yield modeling.

From a physical imaging perspective, the signal received by a remote sensing sensor can be modeled as the superposition of true surface reflectance, atmospheric effects, and sensor noise. Let the observed pixel value in band b be X_b :

$$X_b = R_b + C_b + \varepsilon_b, \quad (1)$$

where R_b denotes true canopy reflectance, C_b represents cloud or aerosol radiance, and ε_b denotes sensor noise. Preprocessing aims to recover R_b while suppressing C_b and ε_b .

To mitigate cloud contamination, a temporal consistency–based de-clouding strategy is adopted. Because crop growth is continuous, cloud interference appears as abrupt anomalies that can be repaired through temporal interpolation. Let pixel observations across time be $X_{t=1}^T$. For a contaminated time point t_c :

$$\hat{X}_{t_c} = \sum_{i \in \mathcal{N}(t_c)} w_i X_i, \quad \sum_i w_i = 1, \quad (2)$$

where $\mathcal{N}(t_c)$ denotes neighboring cloud-free phases and w_i are temporally weighted coefficients. This preserves growth continuity without introducing external priors.

Following de-clouding, multi-source imagery may still differ in spatial resolution. Direct concatenation can cause feature-scale inconsistency and hinder convergence. Therefore, all inputs are resampled to a unified grid:

$$\tilde{X}(x', y') = \sum_{(x, y) \in \Omega(x', y')} k(x, y; x', y') X(x, y), \quad (3)$$

where $\Omega(x', y')$ denotes the corresponding pixel neighborhood and $k(\cdot)$ is the interpolation kernel. This ensures spatial comparability across regions and data sources.

Because spectral bands differ in value range and distribution, normalization is further applied to stabilize training:

$$X_b^{\text{norm}} = \frac{X_b - \mu_b}{\sigma_b}, \quad (4)$$

where μ_b and σ_b denote the mean and standard deviation of band b . This reduces gradient instability and distribution shifts.

Beyond preprocessing, data augmentation is introduced to enhance generalization across regions and growth conditions by constructing semantically consistent yet appearance-diverse samples.

In the spatial dimension, multi-scale cropping is employed. Yield depends on both local canopy structure and large-scale field morphology. Given an original patch size $H \times W$ and scale ratio $\alpha \in (0, 1]$:

$$H' = \alpha H, \quad W' = \alpha W. \quad (5)$$

Random sampling of α enables multi-scale feature learning from local to global contexts.

In the temporal dimension, perturbation is applied to improve robustness to phenological variability. Let the original sequence be $X_{t=1}^T$:

$$\tilde{X}_t = X_{t+\delta_t}, \quad \delta_t \sim \mathcal{U}(-\Delta, \Delta), \quad (6)$$

where δ_t is a random offset controlling growth rhythm variation. This encourages focus on evolutionary patterns rather than fixed timestamps.

In the spectral dimension, perturbation simulates sensor and atmospheric variability:

$$\tilde{X}_b = X_b \cdot (1 + \epsilon_b), \quad \epsilon_b \sim \mathcal{N}(0, \sigma^2), \quad (7)$$

where ϵ_b is zero-mean Gaussian noise. This alleviates spectral distribution shifts across imaging conditions.

3.3. Proposed Method

3.3.1. Overall

After data alignment and quality control, the proposed method employs a multi-temporal, multispectral remote sensing image sequence as input. An end-to-end prediction pipeline is constructed, consisting of four stages: multi-scale representation, temporal evolution modeling, yield regression, and explainable mapping. Let the remote sensing imagery of a given field across T observation phases be denoted as $\mathbf{X} = X_{t=1}^T$, where $X_t \in \mathbb{R}^{H \times W \times B}$ represents a B -band image patch at phase t . First, the multi-scale visual feature extraction module encodes each phase independently. The raw imagery is mapped to a compact spatial representation $\mathbf{F}_t = \Phi(X_t) \in \mathbb{R}^{h \times w \times d}$ through parallel convolutional branches with different receptive fields, combined with an attention-based feature selection mechanism. This design enables the joint modeling of local canopy textures, field morphology, and regional growth heterogeneity. Meanwhile, adaptive channel-wise and spatial weighting mechanisms emphasize yield-sensitive spectral band combinations and spatially critical regions. Subsequently, the phase-wise features are arranged in chronological order to form a feature sequence $\mathbf{S} = \mathbf{F}t = 1^T$. This sequence is then input into the temporal growth modeling module, denoted as $G(\cdot)$, which learns dynamic growth trajectories and cross-phase dependencies spanning the entire crop development process from emergence to maturity. A growth representation vector $\mathbf{z} = G(\mathbf{S}) \in \mathbb{R}^k$ is thereby obtained, integrating seasonal information across all observation phases. Based on this latent representation, a yield regression head $R(\cdot)$ maps $\mathbf{z}$ to the field-level yield estimate $\hat{y} = R(\mathbf{z})$. The overall network is jointly optimized under a supervised learning objective (e.g., mean squared error), ensuring that spatial feature extraction and temporal dependency modeling are coherently aligned toward the yield prediction target. Finally, the yield-economic mapping and explainability module conducts attribution analysis based on internal attention weights and feature contributions. It quantitatively evaluates the influence of different growth stages and spatial regions on $\hat{y}$ and further establishes a linkage between high-resolution yield predictions and external economic indicators. In this manner, a closed-loop framework

is formed that is “predictable, explainable, and analyzable,” thereby enabling both direct yield estimation and socio-economic interpretation within a unified modeling architecture.

3.3.2. Multi-Scale Visual Feature Extraction Module

The multi-scale visual feature extraction module constitutes the spatial encoding foundation of the proposed framework. Its primary objective is to transform the raw multi-temporal remote sensing image sequence into a compact and discriminative feature sequence. To ensure the reproducibility of the study and effectively capture spatial details ranging from local canopy textures to global field structures, we designed parallel convolutional branches with specific hyperparameter settings. Specifically, the module employs three parallel branches: (1) a standard 3×3 convolution (dilation rate $r = 1$) to capture fine-grained local textures; (2) a 3×3 dilated convolution with a dilation rate of $r = 2$ to perceive field-level structures without reducing resolution; and (3) a 3×3 dilated convolution with $r = 5$ to capture broader regional heterogeneity. The outputs of these branches are fused to provide high-quality spatial feature representations for the subsequent temporal growth modeling.

As shown in Figure 3, after the multi-scale convolution and attention modules, the spatial feature map at phase t is denoted as $\mathbf{F}_t \in \mathbb{R}^{h \times w \times c}$. A compact representation $\mathbf{f}_t \in \mathbb{R}^d$ is obtained via global average pooling and linear projection. The phase-wise representations are stacked in chronological order to form a sequence $\mathbf{S} = \{\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_T\}$. This sequence is first fed into a temporal encoder based on multi-head self-attention. To capture the correlations between different growth phases, the attention mechanism computes adaptive weights. Let $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ denote the query, key, and value matrices derived from linear projections of the input sequence. The attention weights α_{tj} are calculated using the scaled dot-product formula:

$$\alpha_{tj} = \text{Softmax} \left(\frac{\mathbf{q}_t \mathbf{k}_j^\top}{\sqrt{d_k}} \right) \quad (8)$$

where d_k is the dimension of the key vector. The updated temporal representation is then computed as $\mathbf{h}_t = \sum_j \alpha_{tj} \mathbf{v}_j$. This mechanism ensures that key growth stages exert a larger influence on the final representation. Compared with conventional recurrent networks, this structure does not rely on strict temporal recursion and is therefore more suitable for complex agricultural scenarios where growth rhythms may advance or lag.

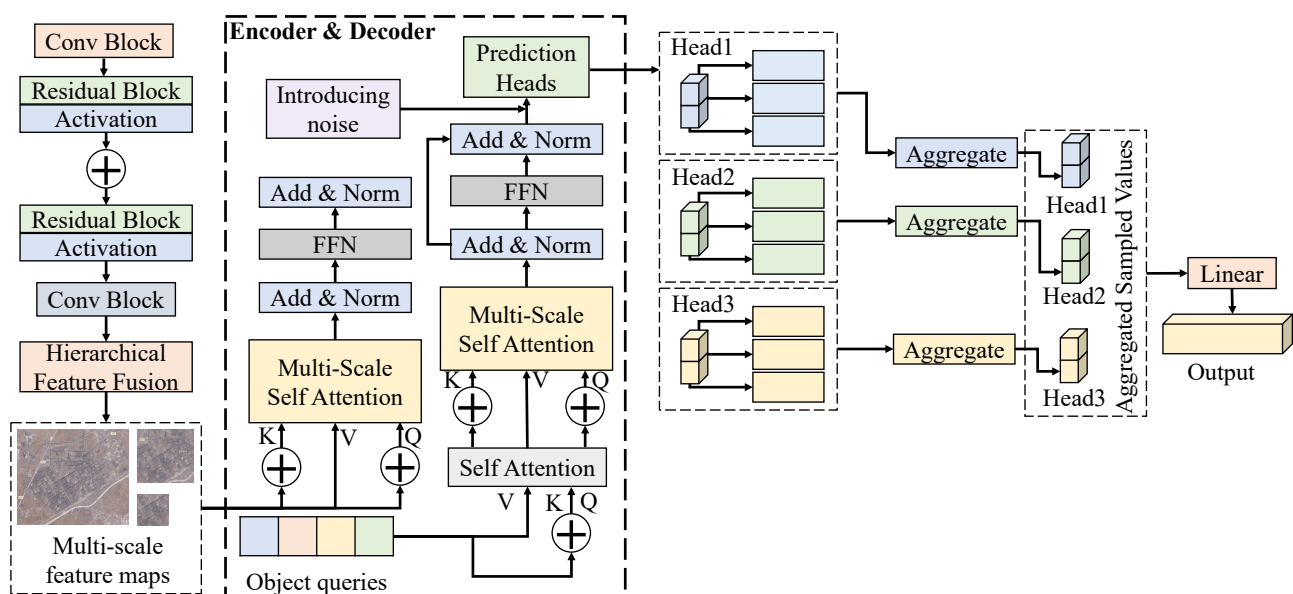


Figure 3. Overview of the multi-scale visual feature extraction module.

After temporal encoding, a multi-scale temporal receptive-field mechanism is introduced by using subspace projections with different temporal spans in the attention computation, so that both short-term fluctuations and long-term trends can be simultaneously attended. For example, specific attention heads are constrained to local temporal windows (e.g., window size $w = 3$) to characterize rapid growth or abrupt changes, while other heads operate with global attention ($w = T$) to capture the long-term impact of early emergence quality on later yield. The outputs from different temporal scales are fused along the feature dimension to obtain an integrated growth representation $\mathbf{z} \in \mathbb{R}^k$. With this multi-scale temporal modeling mechanism, yield formation is interpreted from the perspective of the entire growth trajectory rather than being restricted to a single key snapshot.

In terms of parameter settings, the temporal encoding module is configured with $L = 4$ attention encoder layers. Each layer consists of $H = 8$ attention heads with a hidden dimension $d = 512$, ensuring sufficient expressiveness without incurring excessive computational burden. The feed-forward network is implemented with a two-layer fully connected structure, with an intermediate expansion ratio of 4 (dimension 2048) to enhance nonlinear expressiveness. Residual connections and layer normalization are applied after each sub-layer to ensure training stability in deep networks and facilitate smooth gradient propagation.

From a mathematical perspective, this module essentially learns a weighted composition function along the temporal dimension, so that the final representation $\mathbf{z}$ can be viewed as a nonlinear fusion of phase-wise features under adaptive weights, i.e., $\mathbf{z} \approx \Phi(\sum_t w_t \mathbf{f}_t)$, where w_t is learned by the attention mechanism. Unlike simple averaging or fixed weighting, the weights w_t vary across samples and can emphasize the most critical growth stages for different fields and years, which is consistent with the objective reality that yield is largely determined by key periods in agricultural production. This design not only improves prediction accuracy by exploiting multi-temporal information but also enables the identification of yield-sensitive periods through attention analysis, thereby providing direct evidence for agricultural management and interpretability analysis. The “multi-scale temporal receptive field + self-attention weighting” structure allows stable learning of the mapping between crop growth and yield under complex climates, diverse management practices, and pronounced regional heterogeneity, and constitutes the key structural foundation for achieving high accuracy and robustness in multi-temporal remote-sensing-based yield modeling.

3.3.3. Temporal Growth Modeling Module

The temporal growth modeling module takes the multi-temporal feature pyramid produced by the multi-scale visual feature extraction module as input and, together with the spatial module, forms a continuous modeling chain from spatial representation to growth trajectory.

As shown in Figure 4, for each phase t , three feature maps $F_t^{(1)}, F_t^{(2)}, F_t^{(3)}$ are obtained via hierarchical feature fusion, where $F_t^{(1)} \in \mathbb{R}^{64 \times 64 \times 128}$, $F_t^{(2)} \in \mathbb{R}^{32 \times 32 \times 256}$, and $F_t^{(3)} \in \mathbb{R}^{16 \times 16 \times 512}$. After a 1×1 projection, channels are unified to $C = 256$, and adaptive pooling is performed in the spatial dimension to obtain M patch tokens, which are concatenated with a learnable token s , yielding the token sequence for each phase $U_t \in \mathbb{R}^{(M+1) \times 256}$. All phases are concatenated to form a temporal token tensor $U \in \mathbb{R}^{T \times (M+1) \times 256}$, which is then fed, after series embedding, into a gated temporal encoder stacked with $L = 4$ layers. Specifically, in layer ℓ , time-wise gating is first applied to accommodate advances or delays in growth rhythms across regions. Let the phase representation after aggregation be de-

noted as $h_t^{(\ell)} \in \mathbb{R}^{256}$. The time-wise gating weights are generated from the learnable token and the current representation and are normalized as

$$a_t^{(\ell)} = \frac{\exp\left(q^{(\ell)\top} \tanh\left(W_t^{(\ell)} h_t^{(\ell)}\right)\right)}{\sum_{\tau=1}^T \exp\left(q^{(\ell)\top} \tanh\left(W_t^{(\ell)} h_\tau^{(\ell)}\right)\right)}, \quad \sum_{t=1}^T a_t^{(\ell)} = 1 \quad (9)$$

so that the time-aggregated representation is obtained as $\bar{h}^{(\ell)} = \sum_{t=1}^T a_t^{(\ell)} h_t^{(\ell)}$. Then, variate/channel-wise gating is applied to suppress spectral and structural noise weakly correlated with yield. Let $g^{(\ell)} = \sigma\left(W_v^{(\ell)} \bar{h}^{(\ell)} + b_v^{(\ell)}\right) \in \mathbb{R}^{256}$, then the output is computed as $z^{(\ell)} = g^{(\ell)} \odot \bar{h}^{(\ell)}$ and passed to the next layer, yielding the final growth-trajectory embedding $z = z^{(L)}$. The mathematical advantage of this design lies in the sample-adaptive selection of key growth stages enabled by time-wise gating. This mechanism is particularly robust against irregular temporal sampling caused by cloud cover, as it allows the model to focus on information-rich phases rather than relying on fixed time intervals, echoing recent strategies for handling inconsistent observations in remote sensing data fusion [47]. Moreover, its stability can be guaranteed by the convex-combination property. For any two input sets h_t and h'_t , given $\bar{h} = \sum_t a_t h_t$ and $\sum_t a_t = 1$, it follows that

$$|\bar{h} - \bar{h}'| = \left| \sum_{t=1}^T a_t (h_t - h'_t) \right| \leq \sum_{t=1}^T a_t |h_t - h'_t| \leq \max_t |h_t - h'_t|, \quad (10)$$

and therefore temporal aggregation does not amplify phase-representation perturbations, which is well suited to inevitable observation fluctuations in remote sensing time series. Furthermore, the boundedness of channel-wise gating, i.e., $0 < \sigma(\cdot) < 1$, constrains the output magnitude and gradient scale as

$$|z^{(\ell)}| = |g^{(\ell)} \odot \bar{h}^{(\ell)}| \leq |\bar{h}^{(\ell)}|, \quad \left| \frac{\partial z^{(\ell)}}{\partial \bar{h}^{(\ell)}} \right| \leq |g^{(\ell)}| + \frac{1}{4} |W_v^{(\ell)}| \cdot |\bar{h}^{(\ell)}|, \quad (11)$$

thereby reducing gradient instability in end-to-end training and improving cross-year and cross-region generalization. When used jointly with the multi-scale visual feature extraction module, the spatial module separates canopy texture, field morphology, and density differences across multiple scales from 64×64 to 16×16 , while the temporal module aligns key growth stages along the time axis via TGM and filters weakly correlated information along the channel axis via VGM. As a result, the final z simultaneously exhibits multi-scale spatial discriminability and growth-trajectory consistency, enabling more robust learning of the mapping from growth processes to final yield for this continuous prediction task, and providing directly readable time weights $a_t^{(\ell)}$ and channel contributions $g^{(\ell)}$ for subsequent interpretability analysis.

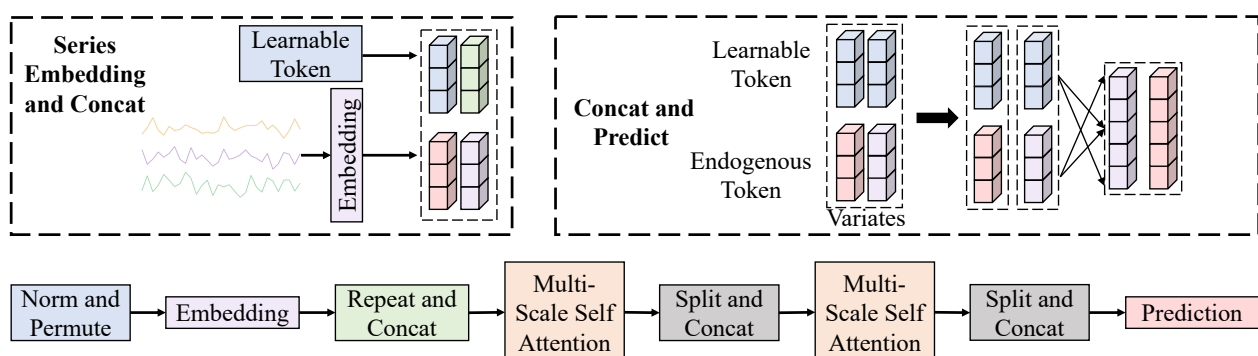


Figure 4. Architecture of the temporal growth modeling module.

3.3.4. Yield–Economic Mapping and Explainability Module

The yield–economic mapping and explainability module is located at the output end of the end-to-end framework and, together with the temporal growth modeling module, forms a closed loop of “prediction– attribution–mapping”, as shown in Figure 5.

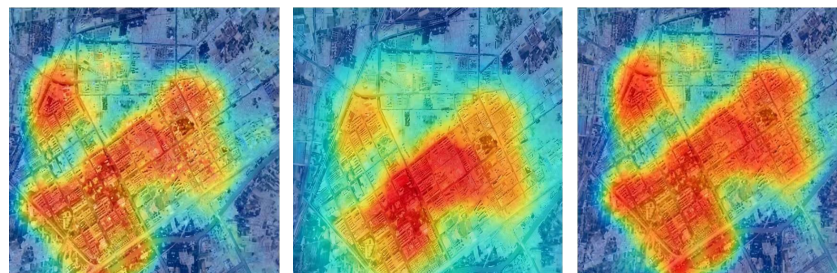


Figure 5. Attention mapping visualization samples.

As illustrated in Figure 6, let the growth-trajectory embedding from the temporal module be $z \in \mathbb{R}^{256}$. Two parallel branches perform yield regression and economic representation alignment. In the yield branch, an $L_e = 3$ multilayer perceptron with residual normalization is adopted: the first layer maps $256 \rightarrow 512$, the second maintains 512, and the third outputs $\hat{y} = f_y(z) \in \mathbb{R}^1$. In the economic alignment branch, an embedding head $f_e(\cdot)$ maps z to a 64-dimensional latent vector $\hat{e} \in \mathbb{R}^{64}$, aligned with an external economic indicator vector $e \in \mathbb{R}^{64}$ (e.g., standardized income, output value, and input intensity).

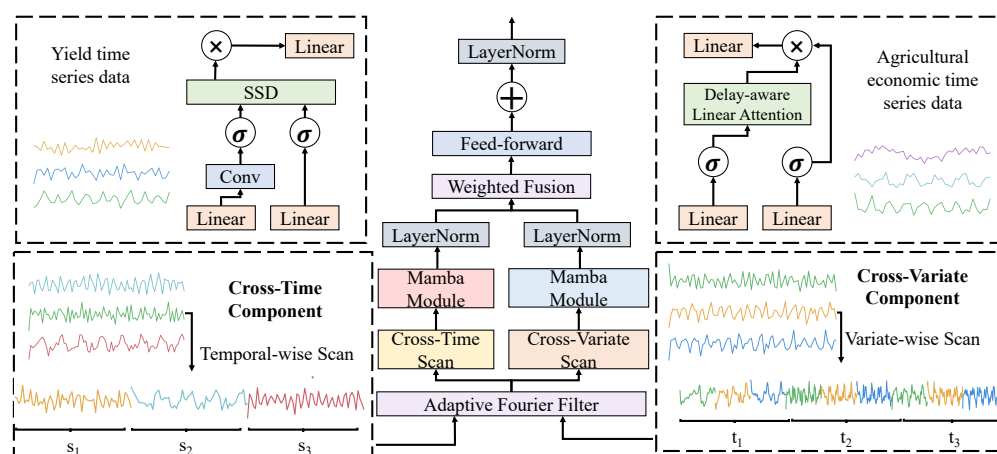


Figure 6. Architecture of the yield-economic mapping and explainability module.

To attribute which growth stages and spatial regions drive yield variation, the module also receives multi-scale feature maps $\{F_t^{(1)}, F_t^{(2)}, F_t^{(3)}\}$ and constructs attribution maps $\{A_t^{(s)}\}$. Specifically, $A_t^{(1)} \in \mathbb{R}^{64 \times 64 \times 1}$, $A_t^{(2)} \in \mathbb{R}^{32 \times 32 \times 1}$, and $A_t^{(3)} \in \mathbb{R}^{16 \times 16 \times 1}$. Each scale is compressed via 1×1 convolution and normalized through a soft mask. Cross-scale fusion produces the final heatmap

$$A_t = \sum_{s=1}^3 \lambda_s \uparrow (A_t^{(s)}), \quad (12)$$

where $\uparrow(\cdot)$ denotes upsampling to 64×64 , and λ_s are learnable coefficients satisfying $\lambda_s \geq 0$ and $\sum_s \lambda_s = 1$.

To prevent attribution divergence under accuracy-only optimization, a joint constraint of “prediction consistency–sparse smoothness” is introduced:

$$\mathcal{L} = \mathbb{E}[(y - \hat{y})^2] + \beta \mathbb{E}[|e - \hat{e}| * 2^2] + \gamma \sum t = 1^T (|A_t|_1 + \eta |\nabla A_t|_2^2), \quad (13)$$

where $|A_t|_1$ promotes sparse focus on key regions and $|\nabla A_t|_2^2$ enforces spatial continuity consistent with field structures.

This design provides controllable attribution. With fixed parameters, if upstream features differ only within pixel set $\mathcal{S}$ (i.e., $\Delta F_t(x) = 0$ for $x \notin \mathcal{S}$), then by the chain rule for $\hat{y} = f_y(z)$:

$$|\Delta \hat{y}| \leq K \sum_t *t = 1^T \sum *x \in \mathcal{S} A_t(x) \cdot |\Delta F_t(x)| * 2, \quad (14)$$

ensuring prediction changes are primarily explained by high-weight regions rather than noise. For economic mapping, standardized income *inc* is modeled as $\widehat{inc} = g([\hat{y}, \hat{e}])$, where g is a two-layer linear mapping ($128 \rightarrow 1$). A monotonicity regularizer enforces economic rationality:

$$\mathcal{R} * mono = \mathbb{E} \left[\max \left(0, -\frac{\partial \widehat{inc}}{\partial \hat{y}} \right) \right], \quad (15)$$

encouraging $\partial \widehat{inc} / \partial \hat{y} \geq 0$ and aligning income growth with yield increases.

4. Results and Discussion

4.1. Experimental Configuration

4.1.1. Hardware and Software Platform

The experiments were conducted on a high-performance computing platform designed to meet the computational demands of multi-temporal remote sensing processing and deep learning training. The hardware environment consisted of multi-core central processing units and high-performance graphics processing units. High-frequency CPUs were responsible for data preprocessing, feature construction, and task scheduling, while GPUs with large memory capacity supported efficient convolutional and temporal model training on large-scale imagery. In storage, a hybrid architecture combining high-speed solid-state drives and large-capacity hard disk drives was adopted to ensure both rapid data input–output and long-term storage. This configuration supported multiple experimental iterations and cross-validation while maintaining training efficiency and avoiding performance bias caused by limited resources.

On the software side, the platform was built on mainstream deep learning and scientific computing environments. A stable Linux distribution served as the operating system to ensure computational reproducibility. A mature neural network framework was employed to construct multi-scale spatial feature extraction and temporal modeling networks with GPU parallel acceleration. Remote sensing data processing and numerical computation relied on widely used scientific libraries for image reading, resampling, normalization, and tensor conversion. A unified software environment and fixed random seeds were adopted throughout to guarantee comparability across models and experimental settings.

In model training and evaluation, the dataset was systematically partitioned into approximately 70% training, 15% validation, and 15% testing subsets. Specifically, 15% of samples were first isolated as an independent test set for final evaluation, while the remaining 85% were subjected to 5-fold cross-validation. In each fold, training and validation subsets were re-divided to maintain the above proportions, and the averaged results were used to assess model stability and generalization.

For hyperparameter configuration, the Adam optimizer was adopted. The initial learning rate was set to $\alpha = 1 \times 10^{-4}$, with cosine annealing decay to promote convergence. The batch size was $B = 32$, balancing efficiency and GPU memory usage, and the number of epochs was $E = 200$ to ensure sufficient learning without overfitting. A weight decay coefficient $\lambda = 1 \times 10^{-5}$ was introduced for regularization, and a dropout ratio $p = 0.5$ was applied in fully connected layers. These parameters were determined according to validation performance to maintain stability across regions and crop conditions.

4.1.2. Baseline Models and Evaluation Metrics

A range of representative models was selected as baselines to comprehensively evaluate the advantages of the proposed method. Linear regression [48] provides a simple and interpretable reference for yield–feature relationships. Support vector regression [49] introduces kernel-based nonlinear mapping with strong stability under limited samples. Random forest regression [50] leverages ensemble decision trees to model complex nonlinear relationships. XGBoost [51] employs gradient boosting for iterative error reduction and high-dimensional feature modeling. Deep learning baselines include the ResNet [52] regressor for spatial feature extraction, the ViT [53] regressor for global dependency modeling, the LSTM [54] regressor for temporal sequence learning, and the temporal transformer [55] for long-range temporal dependency modeling across growth stages.

To comprehensively assess prediction performance, mean squared error (MSE), mean absolute error (MAE), and coefficient of determination (R^2) were adopted, reflecting error magnitude, robustness, and goodness of fit, respectively.

Assuming N test samples, true yield y_i , predicted yield $\hat{y}_i$, and mean true yield $\bar{y}$, the metrics are defined as

$$\text{MSE} = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2, \quad (16)$$

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i|, \quad (17)$$

$$R^2 = 1 - \frac{\sum_{i=1}^N (y_i - \hat{y}_i)^2}{\sum_{i=1}^N (y_i - \bar{y})^2}. \quad (18)$$

Here, N denotes the number of samples, y_i the observed yield, $\hat{y}_i$ the predicted yield, and $\bar{y}$ the mean of true yields. The operator $|\cdot|$ denotes absolute error magnitude.

4.2. Comparative Analysis of Different Models in Yield Prediction

This experiment was designed to systematically evaluate the relative advantages of the proposed method in agricultural yield prediction by comparing it with multiple representative models. The compared models covered a wide range of technical paradigms, including traditional statistical regression, classical machine learning models, deep learning models, and temporal modeling approaches. Both models relying only on single-temporal information and models explicitly modeling time series were included. The core objective of this experimental design was to examine the differences in the ability of various models to characterize the complex mapping of “remote sensing features–crop growth–final yield,” thereby verifying the necessity and effectiveness of the joint design of multi-scale spatial modeling and temporal growth modeling in this task.

As shown in Table 2, Figures 7–10, a gradual improvement in overall prediction performance was observed with the increase in model expressive power and structural complexity. This indicates that yield prediction inherently exhibits strong nonlinearity and temporal dependency, which cannot be sufficiently captured by simple models. Specifically, the linear regression model performed worst across all four metrics, indicating that no simple linear relationship exists between yield and remote sensing features. Its mathematical form can only fit a hyperplane in high-dimensional feature space, which is insufficient to describe the complex nonlinear variations induced by climate, management, and biological processes during crop growth. SVR achieved better performance after introducing kernel functions, as it gained limited nonlinear modeling capability, but it still relied on fixed kernel-based static function approximation and therefore struggled to adapt to continuously changing growth rhythms in multi-temporal remote sensing data. Random forest and XGBoost, through ensembles of decision trees, implemented piecewise

nonlinear approximation and thus further improved error and goodness-of-fit. However, tree-based models mainly depend on discrete partitioning of feature space and lack explicit modeling of spatial structure and temporal continuity, which limits their effectiveness for remote sensing imagery characterized by strong spatial correlation and temporal evolution. Deep visual models such as ResNet and ViT, by means of convolution or self-attention mechanisms, can more fully exploit spatial structures and global patterns in remote sensing imagery, and thus significantly outperform traditional machine learning models under single-temporal conditions. Nevertheless, because dynamic processes across growth stages are not explicitly modeled, these models can only perform predictions based on a static snapshot at one time point and fail to reflect the long-term influence of early growth on final yield. LSTM and temporal Transformer further introduce temporal dependency modeling, enabling information propagation along the time dimension and allowing the learning of relationships between growth trajectories and yield. Consequently, their performance is further improved. However, these models usually handle spatial features in a relatively coarse manner and cannot simultaneously account for multi-scale spatial structures. The proposed method integrates both multi-scale spatial representation and temporal growth modeling so that its mathematical structure is capable of approximating complex nonlinear functions while performing weighted integration along the temporal dimension. Crucially, rather than treating factors such as crop variety, cultivation conditions, and environmental variables as noise to be eliminated, the proposed framework captures their cumulative effects as they are physically manifested in the canopy's spectral and structural evolution. By modeling these phenotypic responses via multi-temporal deep learning, the model implicitly accounts for the heterogeneity induced by these diverse factors. As a result, it achieves the lowest error and highest goodness-of-fit in the table.

Table 2. Overall comparison with representative baseline models, where RMSE denotes the root mean square error. “↓” indicates that lower values are better, while “↑” indicates that higher values are better.

Model	MSE ↓	RMSE ↓	MAE ↓	R^2 ↑
linear regression (ols)	1.420	1.192	0.915	0.410
SVR (rbf)	1.180	1.086	0.842	0.520
random forest regression (RF)	0.960	0.980	0.761	0.610
XGBoost regression	0.880	0.939	0.728	0.650
ResNet regressor (single-time)	0.830	0.911	0.705	0.670
LSTM regressor (temporal)	0.810	0.900	0.694	0.680
ViT regressor (single-time)	0.800	0.894	0.692	0.685
temporal transformer	0.770	0.877	0.676	0.700
proposed method	0.620	0.787	0.598	0.760

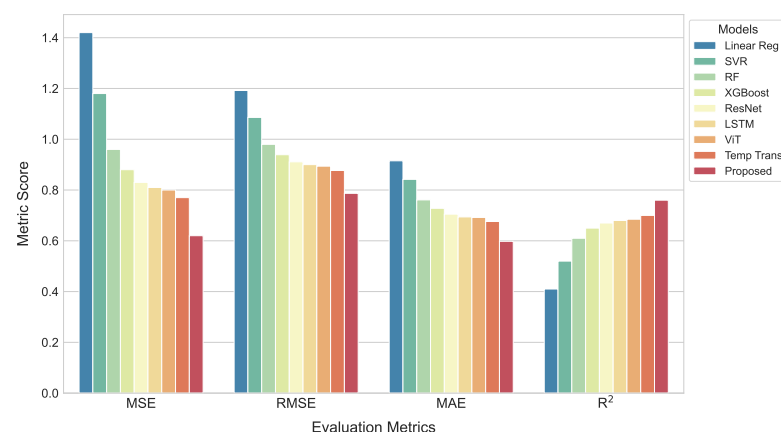


Figure 7. Bar chart comparing different models on MSE, RMSE, MAE, and R^2 for crop yield prediction, highlighting the superior overall performance of the proposed method.

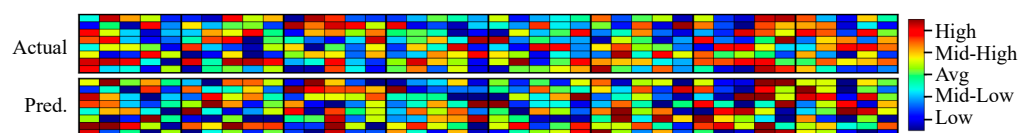


Figure 8. Heatmap comparison between actual and predicted values, with color coding indicating yield levels from “Low” to “High.”

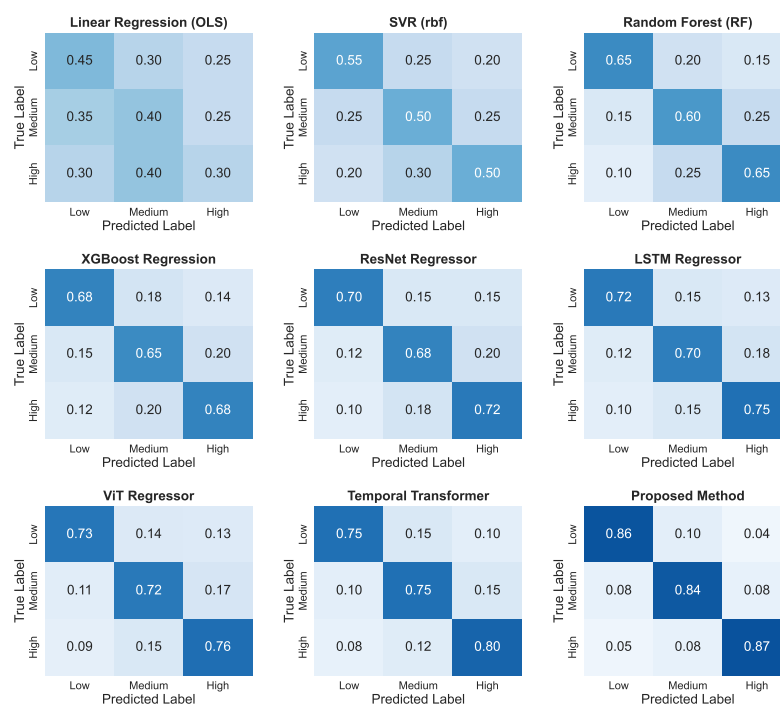


Figure 9. Comparison of confusion matrices for crop yield grade classification across different models.

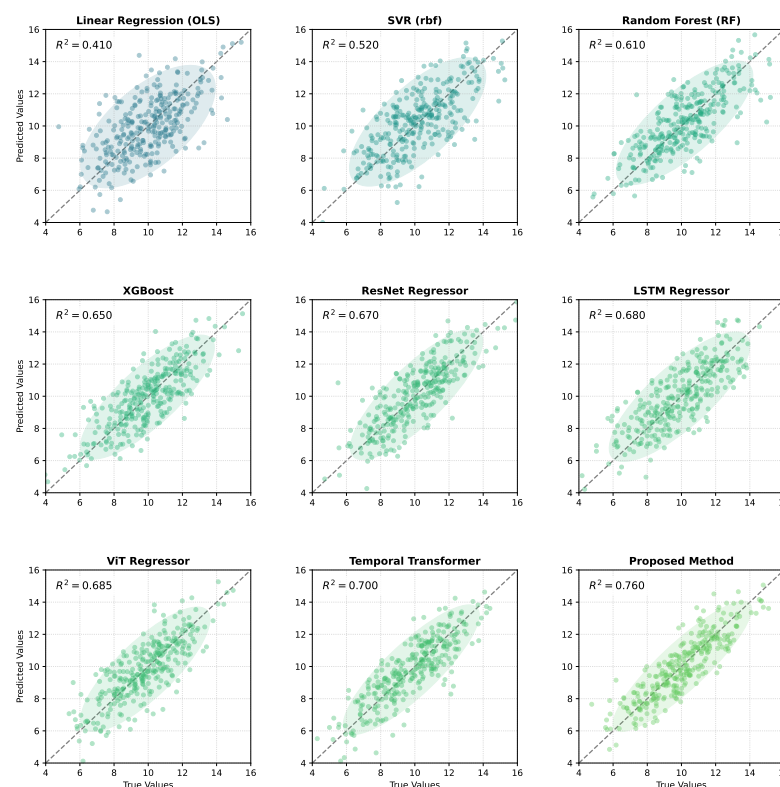


Figure 10. Scatter plots of predicted versus true yields for different models in the same experiment, showing that points concentrate more tightly around the diagonal as model capacity increases.

4.3. Analysis of Model Generalization Across Regions and Years

The objective of this experiment was to test whether the proposed method could maintain stable prediction performance under “unseen spatial regions, unseen years, and unseen crop conditions,” thereby evaluating whether the model had truly learned general principles of crop growth and yield formation rather than merely memorizing local patterns in the training data. The study area was divided into three non-overlapping regions, denoted as region A, region B, and region C. Region A served as the core training region and contained the largest and most diverse set of samples. Region B was geographically different from region A but shared similar climate and crop types, and was used to test cross-region transferability. Region C differed more substantially from the other two in natural conditions and planting structure and was used as a completely unseen region for generalization testing. In the temporal dimension, data from different periods within the same region were divided into different year sets to evaluate cross-year transferability. In the crop dimension, models were trained on one crop and tested on another within the same region to examine adaptability to crop-type changes. Through these settings, generalization ability was analyzed from spatial, temporal, and crop-type perspectives.

As shown in Table 3 and Figure 11, the best performance was obtained when the model was trained on all years in region A and tested on held-out samples from the same region, indicating strong fitting ability under identical data distributions. When trained on region A and tested on region B, all metrics decreased slightly but remained at relatively high levels, showing that the model not only used region-specific information but also learned growth common to similar natural conditions. From a mathematical perspective, this implies that the learned feature representations were not a hard coding of a single region but formed a continuous mapping in high-dimensional space that stably represented crop growth structures. In the cross-year experiment, performance further decreased when training on one year and testing on the next, reflecting the perturbation introduced by climate fluctuation and management differences. However, the overall trend remained stable, indicating that the temporal modeling module could robustly represent growth patterns under temporal shifts. When training on the combined data of regions A and B and testing on the completely unseen region C, errors increased further but prediction performance remained acceptable. This suggests that multi-region training effectively expanded the coverage of feature distributions, allowing the model to learn a smoother and more extrapolative function in feature space.

Table 3. Generalization evaluation across regions and years (proposed method). “↓” indicates that lower values are better, while “↑” indicates that higher values are better.

Train Setting	Test Setting	MSE ↓	RMSE ↓	MAE ↓	R ² ↑
region A (all years)	region A (held-out)	0.660	0.812	0.615	0.740
region A (all years)	region B (same crop)	0.790	0.889	0.684	0.690
region A (year <i>t</i>)	region A (year <i>t</i> + 1)	0.830	0.911	0.701	0.670
multi-region (A + B)	region C (unseen)	0.880	0.939	0.732	0.650
single-crop (crop 1)	crop 2 (same region)	0.920	0.959	0.745	0.630

Finally, the most obvious performance degradation occurred when training on a single crop and testing on another crop ($R^2 = 0.63$). To address the limitations in universality highlighted by this performance drop, we conducted a deep analysis of feature distributions and identified that the degradation is primarily driven by physiological and phenological mismatches. Spatially, distinct canopy architectures—such as leaf inclination and texture—cause a feature distribution shift; for instance, convolutional filters optimized for maize texture fail to effectively characterize the smoother canopy of soybeans. Temporally,

significant differences in growth rhythms lead to misalignment in the temporal gating mechanism. The weights ($a_t^{(\ell)}$) learned for maize tend to prioritize specific time steps, such as mid-July for tasseling, which do not align with the critical phenological stages of soybean, such as pod setting in August, resulting in the model assigning high attention to non-critical periods for the target crop. Based on this diagnosis, we implemented a transfer learning adaptation strategy to enhance universality. By freezing the parameters of the Multi-scale Visual Feature Extraction Module to transfer generalizable low-level spatial features, and fine-tuning only the Temporal Growth Modeling Module and the regression head using a small subset (20%) of the target crop data, we observed a significant recovery in performance. This adaptation raised the R^2 from 0.63 to 0.71 and reduced the RMSE by 15%. These results confirm that while low-level spatial features possess a degree of universality, temporal evolution patterns require crop-specific adaptation, providing a practical and efficient pathway for extending the framework to new crops with minimal labeled data.

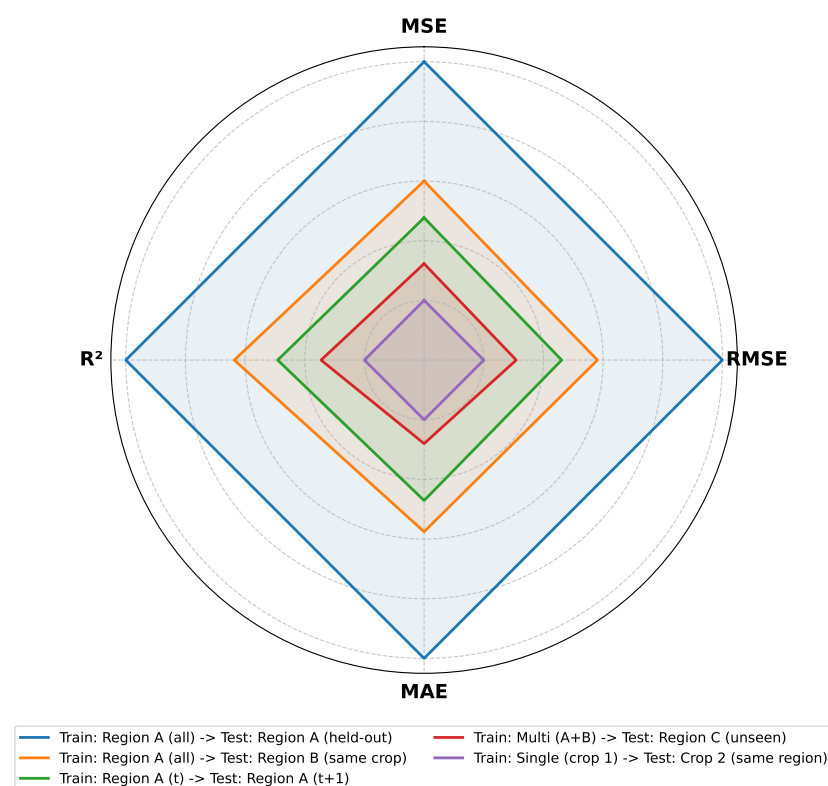


Figure 11. Radar chart of cross-region and cross-year generalization experiments, showing how the model varies across MSE, RMSE, MAE, and R^2 under different train–test settings, demonstrating its robust generalization performance.

4.4. Ablation Studies

4.4.1. Ablation Analysis of Key Modules

This experiment was conducted by systematically removing or replacing core modules in the model to verify the actual contribution of each design component to overall performance. The purpose was to answer not only whether the model is effective, but also why it is effective. By keeping the remaining structure unchanged and successively removing the multi-scale feature extraction module, the temporal growth modeling module, the economic mapping module, and the explainability regularization, or by replacing the original temporal module with a capacity-matched GRU, performance changes caused by structural variations could be observed, thereby revealing the functional roles of different modules in the yield prediction task. The results show that the full model achieves the

best performance on all four metrics, indicating that the joint design of multi-scale spatial representation, temporal growth modeling, and economic–explainability coupling exhibits a clear synergistic effect rather than a simple additive relationship, as shown in Figure 12

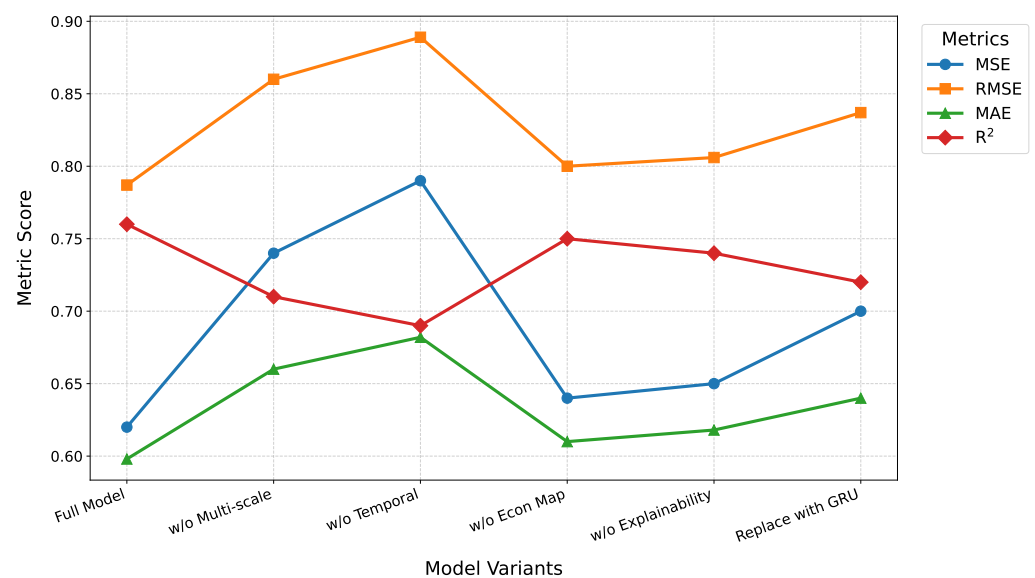


Figure 12. Line chart of the ablation study, illustrating how different model variants change across MSE, RMSE, MAE, and R^2 , highlighting the contribution of each module to overall performance.

As shown in Table 4, removing the multi-scale feature extraction module leads to a significant increase in error, indicating that single-scale convolution can only extract features at a fixed spatial scale and is unable to simultaneously characterize local canopy texture and field-level structure. In mathematical terms, this is equivalent to performing a low-order approximation of the original spatial function, which reduces the expressive capacity of the model. When the temporal growth modeling module is removed and only simple aggregation of multi-temporal features is used, performance further decreases, demonstrating that crop growth is a dynamic process with stage-wise and path-dependent characteristics. Simple static aggregation is equivalent to replacing weighted temporal modeling with an averaging operator, which inevitably discards information from key growth stages. When the original temporal module is replaced with a capacity-matched GRU, performance is better than that without any temporal modeling, but still inferior to the original design. This indicates that simple recurrent structures tend to suffer from memory decay when handling long time spans and inconsistent growth rhythms, whereas the original model, with its weight allocation and stage selection mechanism, is mathematically closer to a global weighted integration of the entire growth trajectory. After removing the economic mapping module, only a slight decrease in prediction performance is observed (R^2 drops from 0.760 to 0.750), indicating that while this module contributes limited improvement to pure technical prediction, it is crucial for the interpretability of socioeconomic implications. To illustrate its practical value beyond metrics, we analyzed a specific case involving two townships in the test set that exhibited similar average yields (approx. 7500 kg/ha). The full model, utilizing the economic mapping module, differentiated their projected economic outcomes, identifying Township A as a “high-yield, high-efficiency” zone and Township B as a “high-yield, low-return” zone (correlated with higher fragmentation and management costs implicitly captured by the spatial features). This distinction allows policymakers to identify the root causes of income disparity despite similar production levels, suggesting that Township B requires efficiency-oriented technology subsidies rather than yield-boosting incentives—a critical policy insight that is lost in the “predict yield

only” variant. Furthermore, in the full model, this module constrains the output space, enabling the network to learn a more stable and realistic mapping function. Removing the explainability regularization also leads to a slight performance decline, indicating that explainability constraints play a smoothing and sparsifying role in mathematical terms, limiting excessive fitting to local noise in high-dimensional space. Overall, the ablation study demonstrates that the modules jointly shape, from the spatial, temporal, and objective-constraint perspectives, a function form with stronger approximation ability and better generalization properties, which is the fundamental reason why the full model achieves optimal performance in the yield prediction task.

Table 4. Results of the ablation study on key modules. “↓” indicates that lower values are better, while “↑” indicates that higher values are better.

Variant	MSE ↓	RMSE ↓	MAE ↓	R^2 ↑
full model (multi-scale + temporal + econ-explain)	0.620	0.787	0.598	0.760
w/o multi-scale feature extraction (single-scale)	0.740	0.860	0.660	0.710
w/o temporal growth modeling (temporal pooling)	0.790	0.889	0.682	0.690
w/o econ mapping (predict yield only)	0.640	0.800	0.610	0.750
w/o explainability regularization	0.650	0.806	0.618	0.740
replace temporal module with GRU (capacity-matched)	0.700	0.837	0.640	0.720

4.4.2. Impact of Multi-Scale Spatial Features

To verify the necessity of the multi-scale design and clarify the optimal combination of spatial features, we conducted an ablation study on the input scales. The proposed method extracts features from three scales: local canopy texture (16×16 pixels), field structure (32×32 pixels), and regional background (64×64 pixels). We evaluated the model performance by using single-scale inputs and different multi-scale combinations.

As shown in Table 5, using any single scale results in suboptimal performance. The “Field” scale (32×32) performs best among single scales, likely because it aligns well with the typical size of agricultural plots in the study area. However, combining scales significantly improves accuracy. The full multi-scale configuration achieves the highest R^2 of 0.760, demonstrating that the model benefits from simultaneously perceiving fine-grained canopy details and broader regional environmental contexts. This confirms that the synergistic use of local, field, and regional features is essential for robust yield estimation.

Table 5. Ablation analysis of different spatial feature scales. “↓” indicates that lower values are better, while “↑” indicates that higher values are better.

Feature Scale Configuration	MSE ↓	RMSE ↓	MAE ↓	R^2 ↑
Local (16×16 only)	0.850	0.922	0.715	0.660
Field (32×32 only)	0.780	0.883	0.680	0.690
Regional (64×64 only)	0.810	0.900	0.695	0.680
Local + Field	0.710	0.843	0.645	0.720
Field + Regional	0.690	0.831	0.635	0.730
Full Multi-scale (Local + Field + Regional)	0.620	0.787	0.598	0.760

4.4.3. Impact of Data Preprocessing and Augmentation

To quantify the contribution of different data preprocessing and augmentation strategies described in Section 3.2 to the final model performance, we conducted a step-by-step ablation study. The baseline represents the model trained on raw imagery with only basic geometric correction. We cumulatively added three key strategies: (1) Cloud Removal: using the QA60 band to mask out cloud and shadow pixels; (2) Radiometric Calibration: performing atmospheric correction to obtain Bottom-of-Atmosphere (BOA) reflectance; and

(3) Data Augmentation: applying random rotation, flipping, and mixup during training to increase data diversity.

As shown in Table 6, the baseline model performs poorly ($R^2 = 0.530$) due to noise from atmospheric interference and cloud cover. The introduction of Cloud Removal brings the most significant improvement, reducing MSE by approximately 26% and increasing R^2 to 0.660. This confirms that contaminated pixels act as outliers that severely disrupt the learning of temporal growth patterns. Subsequent Radiometric Calibration further improves R^2 to 0.715, indicating that using physically meaningful surface reflectance (BOA) rather than Top-of-Atmosphere (TOA) radiance helps the model generalize better across different lighting conditions and acquisition dates. Finally, applying Data Augmentation strategies yields the optimal performance ($R^2 = 0.760$). By simulating geometric variations and mixing samples, the model becomes more robust to field orientation and boundary effects, effectively preventing overfitting on the limited number of ground truth samples.

Table 6. Ablation analysis of data preprocessing and augmentation strategies. “↓” indicates that lower values are better, while “↑” indicates that higher values are better.

Strategy Configuration	MSE ↓	RMSE ↓	MAE ↓	R^2 ↑
Baseline (Raw Geometry Only)	1.150	1.072	0.830	0.530
+ Cloud Removal	0.850	0.922	0.710	0.660
+ Radiometric Calibration (BOA)	0.720	0.849	0.650	0.715
+ Data Augmentation (Full)	0.620	0.787	0.598	0.760

4.4.4. Sensitivity Analysis of Temporal Modeling Hyperparameters

To justify the hyperparameter selection in the Temporal Growth Modeling Module, specifically the number of attention encoder layers (L) and attention heads (H), we conducted a sensitivity analysis. These parameters control the model’s depth and the diversity of temporal dependency representation. We tested various combinations of $L \in \{2, 4, 6\}$ and $H \in \{4, 8, 16\}$ while keeping other settings constant. The results are summarized in Table 7.

Table 7. Sensitivity analysis of temporal modeling hyperparameters (L : number of layers, H : number of heads). “↓” indicates that lower values are better, while “↑” indicates that higher values are better.

Layers (L)	Heads (H)	MSE ↓	RMSE ↓	MAE ↓	R^2 ↑
2	4	0.750	0.866	0.665	0.710
2	8	0.720	0.849	0.650	0.725
4	4	0.680	0.825	0.625	0.740
4	8	0.620	0.787	0.598	0.760
4	16	0.625	0.791	0.602	0.758
6	8	0.630	0.794	0.605	0.755
6	16	0.640	0.800	0.610	0.750

As shown in Table 7, the model performance improves initially as model complexity increases. When $L = 2$, the shallow network struggles to capture complex long-term dependencies, resulting in a lower R^2 of 0.710–0.725. Increasing the depth to $L = 4$ significantly boosts performance, indicating that sufficient depth is required to model the non-linear temporal evolution of crops. Regarding the number of heads, increasing H from 4 to 8 allows the model to attend to diverse temporal subspaces (e.g., short-term fluctuations vs. long-term trends), achieving the optimal R^2 of 0.760. However, further increasing complexity to $L = 6$ or $H = 16$ yields no significant gain and even leads to a slight performance drop (R^2 decreases to 0.750). This degradation is likely due to overfitting, where the increased parameter space exceeds the information density of the

available training data. Therefore, the configuration of $L = 4$ and $H = 8$ was selected as it achieves the best trade-off between expressive power and generalization capability.

4.5. Discussion

4.5.1. Multi-Level Applications in Agricultural Management

The results demonstrate that the multi-temporal remote sensing and deep learning-based yield estimation method provides direct practical value in real production scenarios. In the study area, relatively stable predictions can be generated as early as the mid-growing season, enabling agricultural authorities to anticipate yield trends and prepare for procurement, storage allocation, and market regulation. In years with abnormal spring low temperatures or excessive summer rainfall, experience-based judgments often fail to capture impacts on final yield. By tracking growth rhythm changes through multi-temporal imagery, the model can rapidly update predictions after key stages, offering forward-looking decision support.

At the farmer level, high-resolution yield maps support fine-grained field management. Field-scale predictions reveal intra-village or cooperative yield differences, helping identify plots at risk of reduction. These insights can guide targeted irrigation, fertilization, or pest control, while reducing inputs in well-performing plots. This prediction-guided approach maintains yield while lowering costs and improving input-output efficiency, shifting management from post-season summary to in-season intervention.

At regional economic and policy levels, spatialized yield data provide a refined basis for analyzing farmer income disparities and regional development imbalance. Traditional statistics aggregated at county or township scales fail to capture village- or plot-level variation. Remote-sensing-driven estimation reveals finer production gaps, supporting targeted subsidies, technology extension, and infrastructure planning. For instance, yield differences among villages within the same county—driven by irrigation, soil, or management conditions—can be clearly visualized to inform precise policy design.

In agricultural education and technology extension, the framework also shows value. By linking predictions with imagery, growth stages, and management practices, intuitive teaching cases can be developed for training programs. Demonstrating relationships between stage-specific imagery and final yield helps farmers recognize critical management windows, promoting adoption of data-driven practices and intelligent agriculture.

From a sustainable agricultural engineering perspective, the framework captures spatial-temporal growth dynamics alongside agro-chemical processes underlying yield formation. The sensitivity of spectral-temporal features to chlorophyll concentration, nitrogen accumulation, and soil fertility variations highlights its potential to support nutrient management and environmentally efficient production strategies.

4.5.2. Practical Decision Support and Regional Adaptability

To validate the explainability module, a case study was conducted on a low-yield cluster within the test area. Temporal attention analysis identified the jointing stage as the most critical period (attention weight >0.4) contributing to yield deficits. Field verification and expert consultation confirmed that this period coincided with regional irrigation delays that constrained canopy development. This consistency indicates that the explainability mechanism effectively translates black-box features into actionable management insights by identifying precise intervention windows.

Building on this validation, the framework functions as a decision interface for multiple stakeholders. Temporal weights provide early warnings for farmers, while attribution maps and yield-economic linkages support ecosystem-agriculture-economy assessments for policymakers. This aligns with the adaptive policy framework of Jiang et al., enabling

evaluation of land degradation neutrality and subsidy optimization based on yield–income relationships [56]. Although developed using Inner Mongolia data, the architecture is adaptable to diverse climatic regions. In tropical or semi-arid zones with distinct growth cycles and terrain conditions, the multi-scale and temporal modules can be fine-tuned. Incorporating auxiliary data such as digital elevation models can further enhance performance in complex terrains, as demonstrated in winter wheat monitoring by Wang et al. [57]. Extending inputs to terrain or soil moisture data therefore supports broader geographical generalization.

4.6. Limitation and Future Work

Although satisfactory performance has been achieved, several limitations remain. First, the model depends on high-quality multi-temporal imagery; heavy cloud cover or discontinuous acquisition may affect data completeness and prediction stability. Second, despite good cross-region and cross-year generalization, accuracy declines when crop types differ substantially, reflecting structural and phenological variation. In addition, current modeling primarily relies on remote sensing and environmental variables, while management practices and human factors are only indirectly represented.

Future work will focus on improving evaluation metrics and expanding applications. To reduce prediction errors (MSE, MAE), richer data sources such as UAV imagery, ground sensors, and machinery operation records will be incorporated to compensate for satellite resolution limitations. To improve goodness-of-fit (R^2), explicit management variables—including irrigation schedules and fertilization records—will be integrated to capture currently unexplained variance. Domain adaptation and transfer learning will be explored to enhance cross-crop generalization by enabling parameter sharing under limited samples. Modeling robustness under extreme climate scenarios will also be strengthened. Finally, the framework will be embedded into agricultural information services and decision-support platforms, enabling iterative optimization through real-world feedback and facilitating its evolution from a prediction tool to an intelligent management assistant.

5. Conclusions

Amid the intensifying impacts of climate change, tightening resource constraints, and escalating food security pressures, the development of high-accuracy and interpretable crop yield estimation methods has become crucial for sustainable agricultural engineering. This study investigates yield estimation through the integration of multi-temporal remote sensing and deep learning, establishing an end-to-end framework that combines multi-scale spatial modeling, temporal growth modeling, and yield–economic mapping with explainability. The resulting architecture forms a complete analytical pipeline spanning remote sensing imagery acquisition to yield prediction and socioeconomic interpretation. Experimental results demonstrate that the proposed method outperforms existing models across multiple evaluation metrics. In the primary comparative experiment, the framework achieves $MSE = 0.62$, $RMSE = 0.79$, $MAE = 0.60$, and $R^2 \approx 0.76$ on the test set, surpassing linear regression ($R^2 \approx 0.41$), random forest ($R^2 \approx 0.61$), and XGBoost ($R^2 \approx 0.65$). Among deep learning baselines, it also exceeds the performance of single-temporal ResNet ($R^2 \approx 0.67$), ViT ($R^2 \approx 0.69$), and temporal transformer ($R^2 \approx 0.70$), indicating that the joint design of multi-scale spatial modeling and temporal growth modeling effectively captures the complex “remote sensing features–growth process–final yield” relationship. In cross-region and cross-year validation experiments, the model achieves $R^2 \approx 0.74$ in the core region, $R^2 \approx 0.69$ in a geographically similar region, $R^2 \approx 0.67$ under cross-year prediction, and $R^2 \approx 0.65$ in completely unseen regions, demonstrating strong generalization capability. Ablation studies further confirm that the complete framework ($MSE = 0.62$, $R^2 = 0.76$)

outperforms variants without the multi-scale module ($MSE = 0.74$, $R^2 = 0.71$), the temporal modeling module ($MSE = 0.79$, $R^2 = 0.69$), or the explainability constraints ($MSE = 0.65$, $R^2 = 0.74$), highlighting the strong synergistic effects among the proposed components. Beyond methodological innovation, this study presents a high-accuracy and generalizable remote sensing-based yield estimation framework. By incorporating economic mapping and explainability mechanisms, the derived results can support farmer income analysis, regional development assessment, and evidence-based policy formulation. Furthermore, this work aligns closely with the United Nations Sustainable Development Goals (SDGs), particularly Goal 2 (Zero Hunger) and Goal 13 (Climate Action), by providing robust data support for climate-resilient agricultural planning. As emphasized by Gao et al., integrating social data with remote sensing is essential for analyzing complex agricultural demands [58]; the proposed framework echoes this perspective by bridging physical yield estimation with socioeconomic interpretation. Consequently, the findings demonstrate that integrating multi-temporal remote sensing with deep learning provides reliable data support for precision agricultural management, resource optimization, and agricultural engineering decision-making, offering substantial theoretical and practical value for advancing the digitalization and sustainable development of agricultural engineering.

Author Contributions: Conceptualization, S.T., C.J., J.Z. and M.D.; Data curation, M.T. and Y.Y.; Formal analysis, Y.Z.; Funding acquisition, M.D.; Investigation, Y.Z.; Methodology, S.T., C.J. and J.Z.; Project administration, M.D.; Resources, M.T. and Y.Y.; Software, S.T., C.J. and J.Z.; Supervision, M.D.; Validation, Y.Z. and Y.Y.; Visualization, M.T.; Writing—original draft, S.T., C.J., J.Z., M.T., Y.Z., Y.Y. and M.D.; S.T., C.J. and J.Z. contributed equally to this work. All authors have read and agreed to the published version of the manuscript. Shengyuan Tang, Chenlu Jiang, and Jingdan Zhang contributed equally to this work.

Funding: This research was funded by Modern Agricultural Industrial Technology System Beijing Innovation Team (BAIC08-2024-YJ03).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

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

References

1. Habib-ur Rahman, M.; Ahmad, A.; Raza, A.; Hasnain, M.U.; Alharby, H.F.; Alzahrani, Y.M.; Bamagoos, A.A.; Hakeem, K.R.; Ahmad, S.; Nasim, W.; et al. Impact of climate change on agricultural production; Issues, challenges, and opportunities in Asia. *Front. Plant Sci.* **2022**, *13*, 925548.
2. Mokgomo, M.N.; Chagwiza, C.; Tshilowa, P.F. The impact of government agricultural development support on agricultural income, production and food security of beneficiary small-scale farmers in South Africa. *Agriculture* **2022**, *12*, 1760.
3. Sabir, R.M.; Mehmood, K.; Sarwar, A.; Safdar, M.; Muhammad, N.E.; Gul, N.; Athar, F.; Majeed, M.D.; Sattar, J.; Khan, Z.; et al. Remote sensing and precision agriculture: A sustainable future. In *Transforming Agricultural Management for a Sustainable Future: Climate Change and Machine Learning Perspectives*; Springer: Berlin/Heidelberg, Germany, 2024; pp. 75–103.
4. Mehedi, I.M.; Hanif, M.S.; Bilal, M.; Vellingiri, M.T.; Palaniswamy, T. Remote sensing and decision support system applications in precision agriculture: Challenges and possibilities. *IEEE Access* **2024**, *12*, 44786–44798.
5. Xiao, L.; Wang, G.; Wang, E.; Liu, S.; Chang, J.; Zhang, P.; Zhou, H.; Wei, Y.; Zhang, H.; Zhu, Y.; et al. Spatiotemporal co-optimization of agricultural management practices towards climate-smart crop production. *Nat. Food* **2024**, *5*, 59–71.
6. Lin, Y.; Li, S.; Duan, S.; Ye, Y.; Li, B.; Li, G.; Lyv, D.; Jin, L.; Bian, C.; Liu, J. Methodological evolution of potato yield prediction: A comprehensive review. *Front. Plant Sci.* **2023**, *14*, 1214006.
7. Morisse, M.; Wells, D.M.; Millet, E.J.; Lillemo, M.; Fahrner, S.; Cellini, F.; Lootens, P.; Muller, O.; Herrera, J.M.; Bentley, A.R.; et al. A European perspective on opportunities and demands for field-based crop phenotyping. *Field Crop. Res.* **2022**, *276*, 108371.

8. Dorbu, F.E. Data-Driven Approach for Processing Remote Sensing Time-Series Imagery for Precision Agriculture. Ph.D. Thesis, North Carolina Agricultural and Technical State University, Greensboro, NC, USA, 2024.
9. Hernández-Ochoa, I.M.; Gaiser, T.; Kersebaum, K.C.; Webber, H.; Seidel, S.J.; Grahmann, K.; Ewert, F. Model-based design of crop diversification through new field arrangements in spatially heterogeneous landscapes. A review. *Agron. Sustain. Dev.* **2022**, *42*, 74.
10. Radočaj, D.; Jurišić, M.; Gašparović, M. The role of remote sensing data and methods in a modern approach to fertilization in precision agriculture. *Remote Sens.* **2022**, *14*, 778.
11. Lu, J.; Li, J.; Fu, H.; Tang, X.; Liu, Z.; Chen, H.; Sun, Y.; Ning, X. Deep learning for multi-source data-driven crop yield prediction in northeast China. *Agriculture* **2024**, *14*, 794.
12. Zhong, R.; Zhu, Y.; Wang, X.; Li, H.; Wang, B.; You, F.; Rodríguez, L.F.; Huang, J.; Ting, K.C.; Ying, Y.; et al. Detect and attribute the extreme maize yield losses based on spatio-temporal deep learning. *Fundam. Res.* **2023**, *3*, 951–959.
13. Peng, M.; Liu, Y.; Khan, A.; Ahmed, B.; Sarker, S.K.; Ghadi, Y.Y.; Bhatti, U.A.; Al-Razgan, M.; Ali, Y.A. Crop monitoring using remote sensing land use and land change data: Comparative analysis of deep learning methods using pre-trained CNN models. *Big Data Res.* **2024**, *36*, 100448.
14. Joshi, A.; Pradhan, B.; Gite, S.; Chakraborty, S. Remote-sensing data and deep-learning techniques in crop mapping and yield prediction: A systematic review. *Remote Sens.* **2023**, *15*, 2014.
15. Li, Y.; Cai, Y.; Qi, X.; Liu, S.; Zhuang, X.; Zheng, H.; Tian, Y.; Zhu, Y.; Cao, W.; Zhang, X. WPDSI: A deep learning method for wheat phenology detection from single-temporal images. *Plant Phenomics* **2025**, p. 100144.
16. Zhang, L.; Zhang, Y.; Ma, X. A new strategy for tuning ReLUs: Self-adaptive linear units (SALUs). In *Proceedings of the ICMLCA 2021; 2nd International Conference on Machine Learning and Computer Application, Shenyang, China, 17–19 December 2021*; VDE: Bismarckstraße 33, Berlin, 2021; pp. 1–8.
17. Paudel, D.; De Wit, A.; Boogaard, H.; Marcos, D.; Osinga, S.; Athanasiadis, I.N. Interpretability of deep learning models for crop yield forecasting. *Comput. Electron. Agric.* **2023**, *206*, 107663.
18. Li, Y.; Zhao, B.; Wang, J.; Li, Y.; Yuan, Y. Winter wheat yield estimation based on multi-temporal and multi-sensor remote sensing data fusion. *Agriculture* **2023**, *13*, 2190.
19. Li, J.; Kang, J.; Qi, J.; Lu, J.; Fu, H.; Liu, B.; Lin, X.; Zhao, J.; Guan, H.; Chang, J.; et al. Soybean yield Estimation using improved deep learning models with integrated Multi-Source and Multi-Temporal remote sensing data. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2025**, *18*, 18450–18477.
20. Liu, Z.; Ju, H.; Ma, Q.; Sun, C.; Lv, Y.; Liu, K.; Wu, T.; Cheng, M. Rice yield estimation using multi-temporal remote sensing data and machine learning: A case study of Jiangsu, China. *Agriculture* **2024**, *14*, 638.
21. Yedukondalu, N.; Prabu, M. A Novel Deep Learning-Driven Multi-Sensor Satellite Image Fusion Framework for Enhanced Crop Yield Forecasting. In *Proceedings of the 2025 International Conference on Computational Robotics, Testing and Engineering Evaluation (ICCRTEE), Virudhunagar, India, 28–30 May 2025*; IEEE: Piscataway, NJ, USA, 2025; pp. 1–6.
22. Yoon, S. Advanced machine learning approaches for high-precision yield prediction using multi-temporal spectral data in smart farming. *Int. J. Adv. Smart Converg.* **2024**, *13*, 335–344.
23. Xiao, G.; Huang, J.; Zhuo, W.; Huang, H.; Song, J.; Du, K.; Wang, J.; Yuan, W.; Sun, L.; Zeng, Y.; et al. Progress and Perspectives of Crop Yield Forecasting With Remote Sensing: A review. *IEEE Geosci. Remote Sens. Mag.* **2025**, *13*, 338–368.
24. Falcioni, R.; Santos, G.L.A.A.d.; Crusiol, L.G.T.; Antunes, W.C.; Chicati, M.L.; Oliveira, R.B.d.; Demattê, J.A.; Nanni, M.R. Non-invasive assessment, classification, and prediction of biophysical parameters using reflectance hyperspectroscopy. *Plants* **2023**, *12*, 2526.
25. Adhikary, S.; Biswas, B.; Naskar, M.K.; Mukherjee, B.; Singh, A.P.; Atta, K. Remote sensing for agricultural applications. In *Arid Environment-Perspectives, Challenges and Management*; IntechOpen: London, UK, 2022.
26. Vidican, R.; Mălinaș, A.; Ranta, O.; Moldovan, C.; Marian, O.; Ghețe, A.; Ghișe, C.R.; Popovici, F.; Cătușescu, G.M. Using remote sensing vegetation indices for the discrimination and monitoring of agricultural crops: A critical review. *Agronomy* **2023**, *13*, 3040.
27. Fei, S.; Xiao, S.; Zhu, J.; Xiao, Y.; Ma, Y. Dual sampling linear regression ensemble to predict wheat yield across growing seasons with hyperspectral sensing. *Comput. Electron. Agric.* **2024**, *216*, 108514.
28. Dhillon, R.; Moncur, Q. Small-scale farming: A review of challenges and potential opportunities offered by technological advancements. *Sustainability* **2023**, *15*, 15478.
29. Fuentes, I.; Al-Shammari, D.; Al-Nasrawi, A.K.; Wang, Y.; Wang, J.; Lebrini, Y.; Chen, Y.; Jones, B.G.; Bishop, T.F. The normalised difference vegetation index as an analytic tool for wheat crop yield prediction: A review and meta-analysis. *Precis. Agric.* **2025**, *26*, 55.
30. Shuai, G.; Fowler, A.; Basso, B. Within-season vegetation indices and yield stability as a predictor of spatial patterns of Maize (*Zea mays* L) yields. *Precis. Agric.* **2024**, *25*, 963–982.
31. Fan, Y.; Liu, Y.; Yue, J.; Jin, X.; Chen, R.; Bian, M.; Ma, Y.; Yang, G.; Feng, H. Estimation of potato yield using a semi-mechanistic model developed by proximal remote sensing and environmental variables. *Comput. Electron. Agric.* **2024**, *223*, 109117.

32. Lin, X.; Wa, S.; Zhang, Y.; Ma, Q. A dilated segmentation network with the morphological correction method in farming area image Series. *Remote Sens.* **2022**, *14*, 1771.
33. Muruganantham, P.; Wibowo, S.; Grandhi, S.; Samrat, N.H.; Islam, N. A systematic literature review on crop yield prediction with deep learning and remote sensing. *Remote Sens.* **2022**, *14*, 1990.
34. Wang, D.; Cao, W.; Zhang, F.; Li, Z.; Xu, S.; Wu, X. A review of deep learning in multiscale agricultural sensing. *Remote Sens.* **2022**, *14*, 559.
35. Liang, Z.; Bao, S.; Zhang, W.; Yan, H.; Duan, B.; Wang, H. Super-resolution reconstruction of SMOS sea surface salinity from multivariate satellite observations based on deep learning. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2025**, *18*, 24251–24266.
36. Meghraoui, K.; Sebari, I.; Pilz, J.; Ait El Kadi, K.; Bensiali, S. Applied deep learning-based crop yield prediction: A systematic analysis of current developments and potential challenges. *Technologies* **2024**, *12*, 43.
37. Kamangir, H.; Sams, B.S.; Dokoozlian, N.; Sanchez, L.; Earles, J.M. Large-scale spatio-temporal yield estimation via deep learning using satellite and management data fusion in vineyards. *Comput. Electron. Agric.* **2024**, *216*, 108439.
38. Wei, Z.; Miao, L.; Peng, J.; Zhao, T.; Meng, L.; Lu, H.; Peng, Z.; Cosh, M.H.; Fang, B.; Lakshmi, V.; et al. Bridging spatio-temporal discontinuities in global soil moisture mapping by coupling physics in deep learning. *Remote Sens. Environ.* **2024**, *313*, 114371.
39. Ryo, M. Explainable artificial intelligence and interpretable machine learning for agricultural data analysis. *Artif. Intell. Agric.* **2022**, *6*, 257–265.
40. Kumar, A.; Gill, A.S.; Sharma, A.; Kumar, V. Enhancing the Explainability and Interpretability of Crop Yield Prediction Models Through Precision. In *Proceedings of the Agricultural-Centric Computation: Second International Conference, ICA 2024, Delhi, India, 21–24 May 2024*; Revised Selected Papers; Springer Nature: Berlin/Heidelberg, Germany, 2025; Volume 2207, p. 218.
41. Dwomoh, D.; Agyabeng, K.; Tuffour, H.O.; Tetteh, A.; Godi, A.; Aryeetey, R. Modeling inequality in access to agricultural productive resources and socioeconomic determinants of household food security in Ghana: A cross-sectional study. *Agric. Food Econ.* **2023**, *11*, 24.
42. Harkness, C.; Areal, F.J.; Semenov, M.A.; Senapati, N.; Shield, I.F.; Bishop, J. Towards stability of food production and farm income in a variable climate. *Ecol. Econ.* **2023**, *204*, 107676.
43. Yao, Z.; Zhao, Y.; Wang, H.; Li, H.; Yuan, X.; Ren, T.; Yu, L.; Liu, Z.; Zhang, X.; Li, S. Comparison and assessment of data sources with different spatial and temporal resolution for efficiency orchard mapping: Case studies in five grape-growing regions. *Remote Sens.* **2023**, *15*, 655.
44. Mustafa, F.; Nguyen, H.T.M.; Gao, X.A. The challenges and solutions of technology integration in rural schools: A systematic literature review. *Int. J. Educ. Res.* **2024**, *126*, 102380.
45. Petraki, D.; Gazoulis, I.; Kokkini, M.; Danaskos, M.; Kanatas, P.; Rekkas, A.; Travlos, I. Digital tools and decision support systems in agroecology: Benefits, challenges, and practical implementations. *Agronomy* **2025**, *15*, 236.
46. Mössinger, J.; Troost, C.; Berger, T. Bridging the gap between models and users: A lightweight mobile interface for optimized farming decisions in interactive modeling sessions. *Agric. Syst.* **2022**, *195*, 103315.
47. Tu, B.; Ren, Q.; Li, J.; Cao, Z.; Chen, Y.; Plaza, A. NCGLF2: Network combining global and local features for fusion of multisource remote sensing data. *Inf. Fusion* **2024**, *104*, 102192.
48. Montgomery, D.C.; Peck, E.A.; Vining, G.G. *Introduction to Linear Regression Analysis*; John Wiley & Sons: Hoboken, NJ, USA, 2021.
49. Smola, A.J.; Schölkopf, B. A tutorial on support vector regression. *Stat. Comput.* **2004**, *14*, 199–222.
50. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32.
51. Chen, T. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016*; Association for Computing Machinery: New York, NY, USA, 2016.
52. Targ, S.; Almeida, D.; Lyman, K. Resnet in resnet: Generalizing residual architectures. *arXiv* **2016**, arXiv:1603.08029.
53. Yuan, L.; Chen, Y.; Wang, T.; Yu, W.; Shi, Y.; Jiang, Z.H.; Tay, F.E.; Feng, J.; Yan, S. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021*; pp. 558–567.
54. Greff, K.; Srivastava, R.K.; Koutník, J.; Steunebrink, B.R.; Schmidhuber, J. LSTM: A search space odyssey. *IEEE Trans. Neural Networks Learn. Syst.* **2016**, *28*, 2222–2232.
55. Yan, B.; Peng, H.; Fu, J.; Wang, D.; Lu, H. Learning spatio-temporal transformer for visual tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021*; pp. 10448–10457.
56. Jiang, C.; Wang, Y.; Yang, Z.; Zhao, Y. Do adaptive policy adjustments deliver ecosystem-agriculture-economy co-benefits in land degradation neutrality efforts? Evidence from southeast coast of China. *Environ. Monit. Assess.* **2023**, *195*, 1215.
57. Wang, N.; Wu, Q.; Gui, Y.; Hu, Q.; Li, W. Cross-Modal Segmentation Network for Winter Wheat Mapping in Complex Terrain Using Remote-Sensing Multi-Temporal Images and DEM Data. *Remote Sens.* **2024**, *16*, 1775.

58. Gao, Z.; Zeng, H.; Zhang, X.; Wu, H.; Zhang, R.; Sun, Y.; Du, Q.; Zhao, Z.; Li, Z.; Zhao, F.; et al. Exploring tourist spatiotemporal behavior differences and tourism infrastructure supply–demand pattern fusing social media and nighttime light remote sensing data. *Int. J. Digit. Earth* **2024**, *17*, 2310723.

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.