Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Multisource Data Integration for Enhanced Wheat Yield Prediction Using Optimized Ensemble Machine Learning Approaches

Farbod Farhangi

farhangi.phd@outlook.com

Article

Keywords: Wheat yield prediction, Climate data, Remote sensing indices, Machine learning, Metaheuristic optimization

Posted Date: June 11th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9778014/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Multisource Data Integration for Enhanced Wheat Yield Prediction Using Optimized Ensemble Machine Learning Approaches

Farbod Farhangi [1, *]

[1] Independent Researcher, Isfahan, 81799-57971, Iran.

[*] Corresponding author: farhangi.phd@outlook.com

Multisource Data Integration for Enhanced Wheat Yield Prediction Using Optimized Ensemble Machine Learning Approaches

Abstract: Accurate wheat yield prediction is vital for food security in the face of increasing climate variability, yet traditional models often struggle to capture the complex interactions between crops and their environment. This study evaluates a multisource prediction framework across China's major wheat-growing regions, integrating climate and remote sensing data within stage-specific scenarios using advanced machine learning models. Random forest, extreme gradient boosting (XGBoost), and light gradient boosting machine models were optimized with improved grey wolf (IGWO) and whale (IWO) optimization algorithms and applied to twelve vegetation indices alongside key climatic variables. Results show that combining climate and remote sensing inputs significantly improves prediction accuracy over single-source data, with the best performance achieved when full-season remote sensing is paired with early-stage climate variables. Monte Carlo analysis identified water-related factors—wind speed, soil moisture, precipitation, and evaporation—as the most influential drivers of yield, while temperature variables and vegetation indices had weaker effects. Metaheuristic optimization consistently enhanced model performance, with XGBoost-IGWO achieving the highest accuracy (R^2 = 0.755, Willmott index = 0.833, RMSE = 0.495 t/ha). Model accuracy also varied across regions, performing better in smaller, more homogeneous areas. Overall, integrating stage-specific multisource data with optimized machine learning models improves both the reliability and interpretability of large-scale wheat yield predictions.

Keywords: Wheat yield prediction, Climate data, Remote sensing indices, Machine learning, Metaheuristic optimization

1. Introduction

Due to the increasing threat of climate variability and extreme weather events to global food supply stability, governments are placing greater emphasis on forecasting potential declines in agricultural production and formulating effective policy responses [1]. Consequently, accurate and timely crop yield predictions have become vital for shaping agricultural and food security strategies at both national and global levels [2]. An operational crop forecasting system can provide decision-makers with timely, reliable, and spatially detailed yield estimates, enabling them to anticipate critical needs and respond proactively.

Wheat is one of China's most important staple crops and a major source of dietary energy, particularly in the northern regions. It contributes approximately 17–20% of the nation's total caloric intake, underscoring its central role in ensuring food security [3]. Wheat yield formation in China is governed by a complex interplay of biophysical and management factors, including temperature, precipitation, solar radiation, soil fertility, irrigation scheduling, and fertilizer application [4]. Given the nonlinear interactions among these variables and the pronounced spatial heterogeneity across China's diverse agro-ecological zones, advanced modeling frameworks capable of integrating multi-source data and capturing dynamic crop–environment feedbacks are essential for accurately simulating yield responses and enhancing agricultural resilience.

Various modeling paradigms have been proposed to simulate crop growth processes and predict yields under diverse environmental conditions. Process-based models such as DSSAT [5], APSIM [6], World Food Study [7], and AquaCrop [8] are widely used to simulate crop growth and forecast yields by representing plant physiology across the full life cycle [9]. While these models provide valuable insights into crop-environment interactions, they necessitate numerous detailed input parameters, including soil characteristics, nutrient application, climate data, and vegetation attributes, making their calibration typically intricate and time-intensive. In contrast, regression equation-based models integrate satellite data with global contextual datasets to estimate crop yields [9]. Though effective for reconstructing historical trends, these approaches have limitations in capturing recent developments and fail to account for nonlinear relationships, thereby motivating the growing adoption of machine learning techniques [10]. Indeed, literature has demonstrated that machine learning consistently outperforms both regression and process-based models in this field [11].

Machine learning has undergone substantial advancements and has emerged as a prominent approach in crop yield forecasting, owing to its high precision and capacity to process vast datasets on powerful computing platforms [12]. By integrating diverse data sources and extracting salient features, machine learning algorithms have proven effective in addressing complex

nonlinear challenges [13,14]. These methods offer new opportunities to improve agricultural decision-making by providing a flexible and robust framework that incorporates expert knowledge into data-driven systems [15]. Furthermore, machine learning techniques can capture and model interactions among heterogeneous data sources while explicitly considering their spatial and temporal dependencies [16].

Various machine learning approaches have been utilized for crop yield forecasting with promising results. Section II provides an overview of recent research in the field. Despite these advances, machine learning–based crop yield forecasting faces performance challenges because it is strongly influenced by crop type [17], growth stage [9], and regional conditions [18]. Consequently, there is increasing research interest in utilizing machine learning to integrate multisource data for improved crop yield forecasting. However, when integrating multisource data, high-dimensional training data pose significant limitations, including increased risk of overfitting [19], potential information loss [20], and substantial computational demands [21]. Hyperparameter tuning addresses these challenges by optimizing model parameters and selecting the most informative features, leading to improved accuracy and generalizability. Recent literature in crop yield forecasting has addressed this issue utilizing grid search [2,9,18,22,23] and random search [24] methods. Nevertheless, traditional optimization methods such as deterministic, stochastic, and gradient-based approaches can be costly in high-dimensional spaces and prone to local optima, often yielding suboptimal hyperparameter configurations [25,26]. In contrast, metaheuristic approaches effectively steer the search process toward near-optimal solutions within complex, non-convex hyperparameter spaces, thereby lowering the risk of becoming trapped in local minima. These methods are intrinsically self-organizing and distributed, facilitating efficient exploration of high-dimensional search spaces through collective dynamics and enhancing the model's overall generalization performance [27,28].

This study introduces a novel framework to address existing limitations in wheat yield forecasting across China. By integrating stage-specific driver selection with metaheuristic optimization, this study is guided by two key research questions: how stage-specific driver selection improves model relevance, and whether metaheuristic optimization yields superior results over traditional machine learning tuning in this field. Accordingly, two enhanced nature-inspired algorithms—the improved grey wolf optimizer (IGWO) and the improved whale optimization algorithm (IWO)—are employed to optimize the hyperparameters of random forest (RF), extreme gradient boosting (XGBoost), and light gradient boosting machine (LightGBM), thereby capturing temporally varying crop–environment interactions, reducing overfitting, and improving predictive accuracy. The main contributions of this work are

threefold: (i) the construction of a stage-specific, multi-source forecasting framework integrating climate and remote sensing data for improved wheat prediction; (ii) the selection of key drivers and optimization of model hyperparameters to achieve performance gains over baseline models; and (iii) a systematic evaluation of model performance to ensure reproducibility and practical applicability. These contributions are intended to provide a clearer methodological basis for wheat yield forecasting and to support adaptive agricultural practices across heterogeneous cultivation regions

Overview of recent research on crop yield forecasting

An overview of recent studies on crop yield forecasting using remote sensing and machine learning is presented in Table 1. The table summarizes data sources, methodological approaches, and principal findings across the literature [2,9,17,18,22,24,29-33]. Most studies did not perform explicit hyperparameter tuning, and those that did predominantly employed search-based methods such as grid search or random search. Across the literature, non-linear models consistently outperformed linear approaches, while multi-source crop yield forecasting achieved higher accuracy than single-source models. Among the algorithms reviewed, RF, XGBoost, GBM, support vector machine (SVM), least absolute shrinkage and selection operator (LASSO), long short-term memory (LSTM), and k-nearest neighbors (KNN) were the most frequently used, with RF and XGBoost consistently achieving the highest accuracy.

Table 1. Summary of recent research on crop yield forecasting with remote sensing and machine learning

Paper	Objective	Data	Machine learning	Key findings
[9]	Winter wheat yield prediction in Pakistan under full-seasonal and peak seasonal scenarios.	Landsat 5/7/8 (NDVI, GNDVI, EVI, SAVI, ARVI); Climate data; Yield data	RF; SVM; LASSO; Ridge regression	RF outperformed the other models; Peak seasonal mean generally yielded higher accuracy than full-seasonal mean.
[2]	Wheat yield forecasting under environmental variability in Morocco during wet and dry seasons.	Field-level wheat yield data; Sentinel-2 (NDVI, MSI, MTCI); ERA5-Land temperature data; PERSIANN-CCS precipitation data	RF; XGBoost.	RF generally outperformed XGBoost; Weather data provided better early-season predictions; Spectral indices improved predictions closer to harvest; Accuracy was higher in wet season than in the dry season.
[29]	Early prediction model for olive crop and oil yields in Jaén, Spain.	Landsat 7/8 (ARVI, NDMI, GCI, NDVI, SAVI, SIPI); MODIS (surface temperature); ECMWF ERA5 (rainfall); Official	RF; SVM-linear kernel; SVM-Gaussian kernel; Neural Networks	The SVM–Gaussian model outperformed all others; Integrating vegetation indices improved accuracy compared to using meteorological

		municipal-level olive yield statistics; Land use data		variables alone and enabled reliable early forecasts eight weeks in advance.
30	Improving wheat yield prediction in arid regions.	Landsat 8; Sentinel-2; TerraClimate dataset; Ground truth from field surveys; Wheat yield data	RF; GBM; Decision tree; Bagging trees	RF with ARVI during the heading–grain filling stage achieved the best accuracy.
24	Predicting Summer and Kharif rice yields at 45, 60, 90 days post-transplanting.	Fresh biomass; Dry biomass; Plant moisture content; Leaf area index; Plant height: Yield; Sentinel-1; Sentinel-2 (NDVI, RENDVI, RVI, LSWI)	Elastic Net; RF; Multivariate adaptive regression splines; SVM: Linear Regression; KNN, XGBoost; Neural Network; Cubist	Predictions were more accurate closer to harvest; XGBoost consistently performed best; Accuracy was higher when combining SAR, optical, and biophysical parameters than using single data sources.
18	Enhancing rice yield prediction accuracy using a process-based crop model.	Rice yield data; MODIS (NDVI, EVI, LSWI); Shuttle Radar Topography Mission; ECMWF ERA5 daily; International rice research institute lowland rice extent map	Feed-forward Neural Network; LSTM; Gated recurrent units; Bidirectional LSTM	LSTM performed best; Performance declined when predicting for regions excluded from training, highlighting the need for diverse datasets.
31	Developing an accurate wheat yield prediction model for Multan District, Punjab, Pakistan.	Landsat 8 (NDVI, EVI, B2–B5, B7); Climate data from NASA POWER Data Access Viewer & Pakistan Meteorological Department; Soil data; Yield statistics	RF; SVM; LASSO	RF achieved the highest accuracy; Combining climate, soil, and satellite data outperformed using single-source inputs.
32	In-season estimation of winter wheat production at the county scale in China.	Sentinel-2 (Bands 7, 8, 8A, 12; 10–20 m resolution); Historical and real-time weather forecast data; County-level winter wheat yields; Winter wheat harvested area dataset	LSTM–Graph Neural Network	Combining weather forecasts with remote sensing improved accuracy compared to using remote sensing alone.
22	Predicting canola and soybean yields in Canada under climate scenarios and honeybee influence.	Precipitation; Temperature; Soil characteristics; Honeybee data; MODIS MYD13Q1 Aqua satellite product (NDVI and EVI); Land cover; Crop yields	Adaptive Boosting; GBM; XGBoost, Light GBM; RF	XGBoost achieved the best performance; Yield maps were generated at the CCS level for both crops.
33	High-resolution, long-term soybean yield mapping in Brazil's	Landsat 5/7/8; MODIS (MOD09GA, MOD09GQ, MODTBGA, MOD44C); SRTM topographic data; Climatic	RF	The integration of remote sensing, topography, climate, and weather data yielded the highest performance; Soil

	tropical agricultural systems.	research unit time-series v4.04 climate and weather data; Soil Data; Yield statistics		variables did not improve performance.
17	County-level wheat yield prediction across China's principal wheat cultivation zones.	Wheat yield statistics; Phenology data; Wheat-growing area map; ERA5 climate data; MODIS (NDVI, EVI); Solar-induced chlorophyll fluorescence	RF; SVM-Gaussian kernel; LASSO	SVM and RF generally outperformed LASSO; Combining climate and remote sensing data yielded the highest accuracy; Water-related variables outperformed temperature; Climate data dominated early and peak periods in winter wheat zones; Accuracy was lower in spring wheat zones.
23	Early forecasting of cereal yields in Morocco at the provincial scale.	Crop yield data; MODIS (NDVI, VCI, TCI); ESA CCI SM COMBINED soil moisture; ERA5 rainfall and air temperature; North Atlantic Oscillation, Scandinavian Pattern; Atlantic Tripole SST; Atlantic Niño SST	Multiple Linear Regression; SVM; RF, XGBoost	XGBoost outperformed the other models; Combining satellite indices, weather, and climate data improved accuracy over single-source inputs; Satellite indices peaked near harvest; Weather data led early forecasts; Climate indices dominated earliest lead times.
34	Predicting wheat yield at a large scale in Mexico before harvest.	Wheat yield data; MOD13Q1 v.6 product (NDVI); ERA5 climate data	Generalized linear model; Ridge regression; LASSO; Partial least squares; KNN; SVM-radial; XGBoost; RF	Non-linear models outperformed linear ones; RF achieved the best performance; Combining NDVI and climate with yield data improved predictions compared to using yield data alone.

CCS (census consolidated subdivision), GCI (green chlorophyll index), GBM (gradient boosting machine), GNDVI (green NDVI), GNN (graph neural network), KNN (k-nearest neighbors), LASSO (least absolute shrinkage and selection operator), LSTM (long short-term memory), LSWI (land surface water index), MODIS (moderate resolution imaging spectroradiometer), MSI (moisture stress index), MTCI (MERIS terrestrial chlorophyll index), NDMI (normalized difference moisture index), PERSIANN-CCS (precipitation estimation from remotely sensed information using artificial neural networks–cloud classification system), RENDVI (red edge NDVI), RVI (ratio vegetation index), SIPI (structure insensitive pigment index), SRTM (shuttle radar topography mission), SST (sea surface temperature), SVM (support vector machine), TCI (temperature condition index), TerraClimate (high-resolution global climate dataset), VCI (vegetation condition index).

2. Study area

China's agricultural landscape is characterized by strong climatic gradients and diverse soil–ecological conditions, which shape regional differences in cropping systems and wheat production (Figure 1). To account for this spatial variability, wheat cultivation across china is organized into three dominant planting zones based on the climate, soil attributes, and wheat quality: the northern

winter wheat region (Zone 1), the southern winter wheat region (Zone 2), and the northern spring wheat region (Zone 3) [35]. Zone 1 is the leading wheat-growing region in China, producing the highest yields and accounting for about 73% of the nation's total wheat output (including both winter and spring wheat). In comparison, Zones 2 and 3 contribute approximately 16% and 11% of national production, respectively [17].

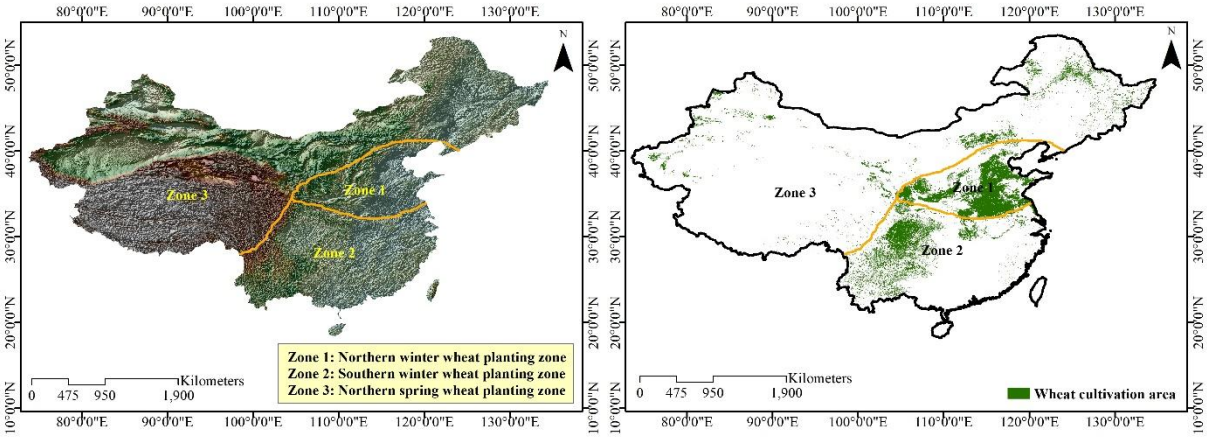


Figure 1. Location map of the study area [35]

The crop calendar for wheat differs across the three zones (Figure 2). In Zones 1 and 2, winter wheat cultivation begins in Oct, with the early growth period lasting until the end of the Mar in Zone 1 and until the end of the Feb in Zone 2. This is followed by a peak period from Apr to May and a late period extending into Jun for Zone 1, and a peak period from Mar to Apr with a late period extending into May for Zone 2. In Zone 3, spring wheat planting starts in Apr, with the early period lasting May, the peak period occurring from May to Jun, and the late period continuing into Jul [17].

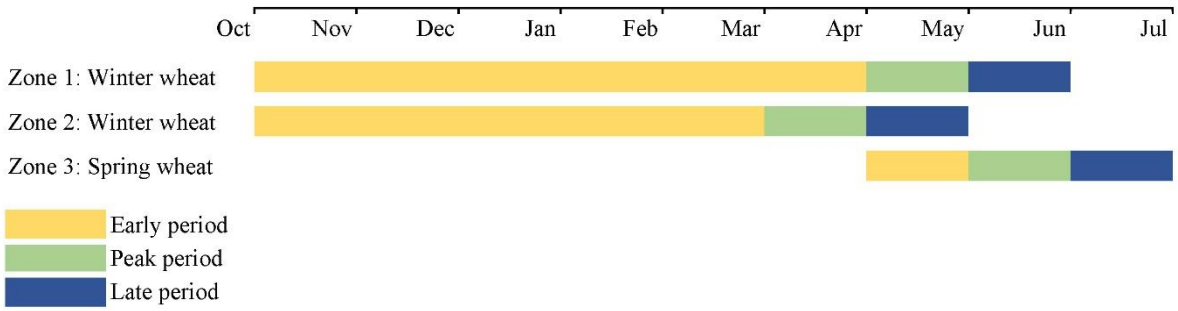


Figure 2. Crop calendar for wheat differs across the three zones [17]

3. Materials And Methods

3.1. Workflow

A six-step workflow was utilized for enhanced crop yield forecasting (Figure 3). Initially, a binary cultivation mask was generated from the wheat cultivation map (1 = cultivated, 0 =

uncultivated), ensuring that only wheat-producing regions were retained for subsequent analyses. Next, wheat yield data, climate variables, and remote sensing indices were obtained and resampled to a uniform spatial resolution. These datasets were then masked using the wheat-growing area binary map. Subsequently, to select the optimal integration of remote sensing and climate information in yield prediction, we designed six scenarios that test different combinations of seasonal and stage-specific inputs. These scenarios were tested using RF, XGBoost, and LightGBM models.

Afterward, a Monte Carlo sensitivity analysis was conducted with 1,000 iterations for each modeling feature independently using RF. The impact of perturbations in each feature on the model's performance was quantified to identify and select the most influential driving factors in modeling. Then, two metaheuristic optimization algorithms of IGWO and IWO were employed for machine learning optimization. Performance of optimized models was compared against baseline models to measure the improvements following the optimization process. Finally, the uncertainty of the predictions was measured using prediction interval coverage probability (PICP) and prediction interval width (PIW) across several nominal confidence levels.

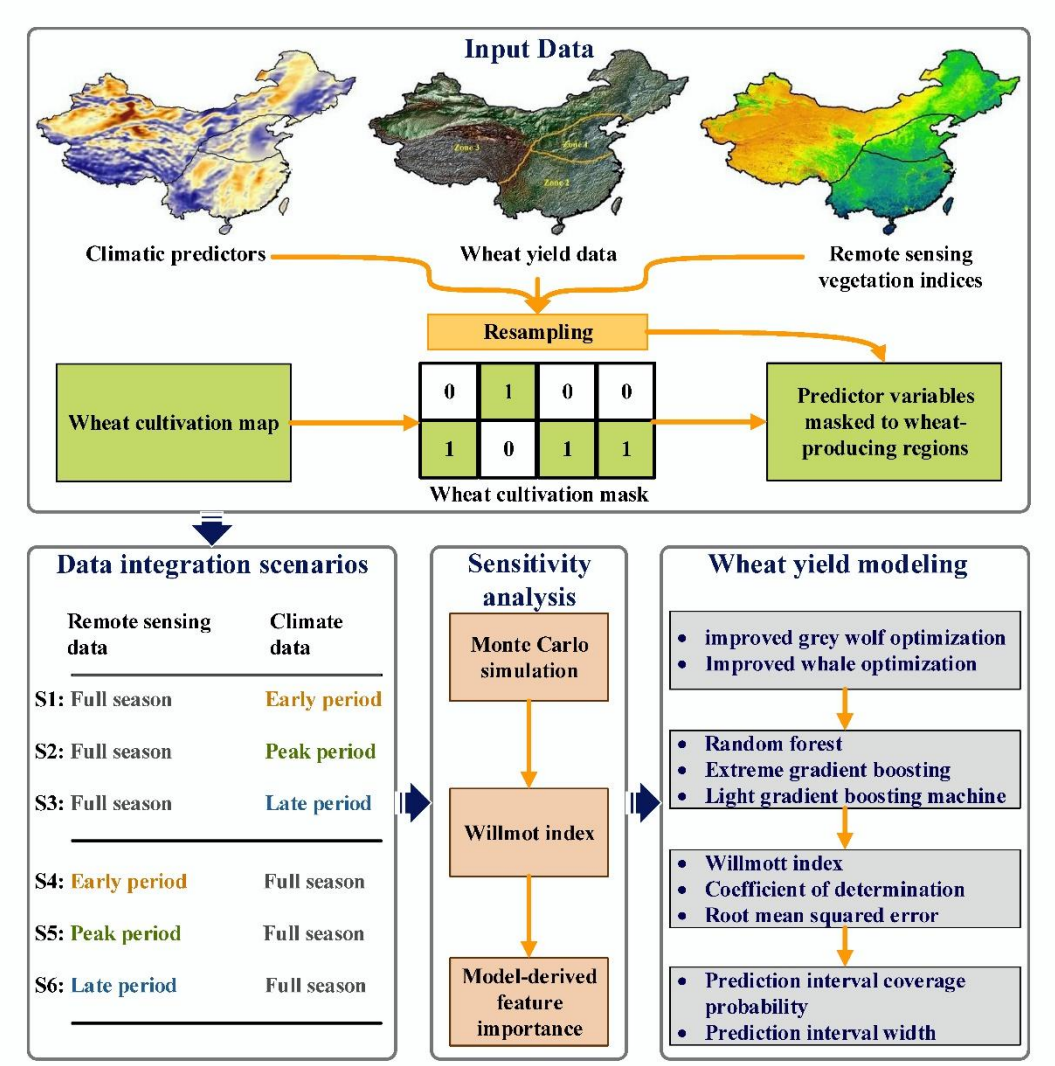


Figure 3. Research methodology

3.2. Datasets

The datasets applied in the reviewed literature (Table 1) were considered to select the most relevant variables for wheat yield forecasting. Table 2 presents the features incorporated into the modeling process. Considered features include agricultural parameters, climate conditions, and satellite-derived vegetation indices, providing a comprehensive representation of factors influencing wheat yield.

Table 2. Utilized features in modeling

Category	Feature	Resolution	Time	Source
Wheat data	Yield (kg/ha)	1000 meters	2002-2010	Agricultural Yearbook
	Phenology (date)	Station	2001-2010	CMA
	Wheat cultivation area	1000 meters		NLCD
Climate data	Air humidity	11132 meters	2001-2010	

	Maximum air temperature Minimum air temperature Total precipitation Actual evaporation Potential evaporation Surface pressure Solar radiation Northward wind speed Eastward wind speed Soil moisture (7 - 28 cm) Soil moisture (28 - 100 cm)			ERA5-Land Monthly Aggregated - ECMWF Climate Reanalysis [36]
Remote sensing indices	NDVI EVI SAVI ARVI	500 meters	2001-2010	MOD09A1.061 Terra Surface Reflectance 8-Day Global 500m [37]

3.3. Wheat yield data

County-level wheat production records for the years 2002–2010 were obtained from the Agricultural Yearbook (https://www.stats.gov.cn). Before analysis, we performed a basic quality screening to identify anomalous entries, removing values that exceeded plausible biophysical limits for wheat productivity as well as outliers. Because the dataset spans only a relatively short time frame, we retained the raw yield values rather than applying any detrending procedure.

Phenological information on wheat development (sowing, jointing, heading, maturity) was sourced from the China Central Meteorological Agency (CMA, https://www.cma.gov.cn). Early-, mid-, and late-growth periods were defined as sowing–jointing, jointing–heading, and heading–maturity, respectively. The timing of each phenological stage in each zone was determined by averaging the monthly records from all stations located within that zone. Figure 1 presents the wheat cultivation map, derived from the 1-km National Land Cover Dataset (NLCD).

3.4. Climate data

Twelve monthly climatic predictors were included in the modeling (Table 2). These consisted of air humidity, maximum and minimum air temperature, total precipitation, actual and potential evaporation, surface pressure, solar radiation, northward and eastward wind speed, and soil moisture at depths of 7–28 cm and 28–100 cm. All variables were derived from the ERA5-Land Monthly Aggregated dataset (ECMWF Climate Reanalysis) at a spatial resolution of 10 km. Figure 4 shows the spatial distribution of monthly climatic predictors across the study area.

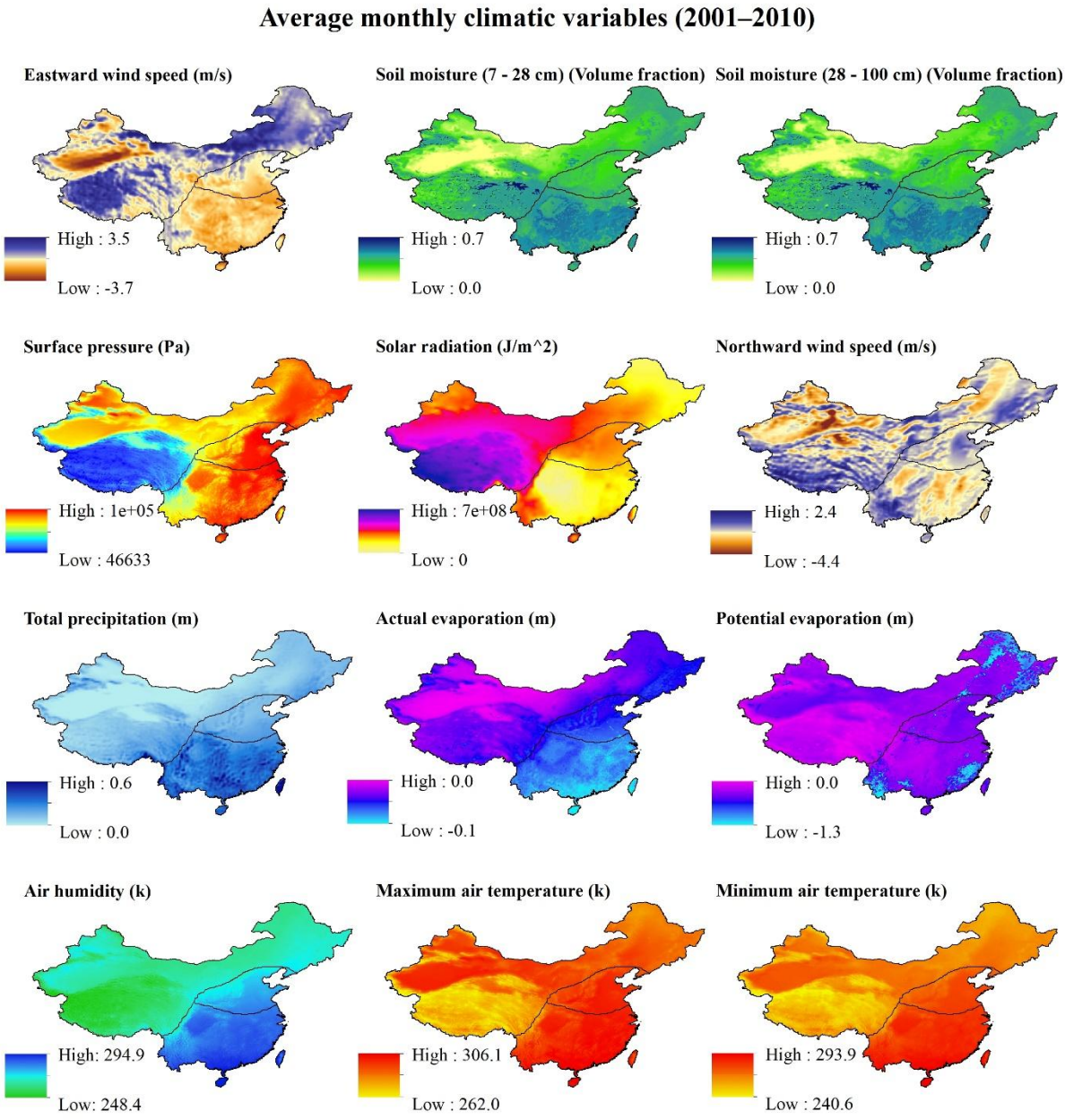


Figure 4. Spatial distribution of average monthly climatic variables (2001–2010)

Table 3 summarizes the descriptive statistics of the monthly climatic variables across the three cultivation zones. The statistics show differences among the zones. Temperature-related variables exhibit lower standard deviations in Zone 1, higher values in Zone 3, and intermediate values in Zone 2. Precipitation-related variables present higher values in Zone 1, lower values in Zone 3, and intermediate values in Zone 2. Overall, measures of dispersion are smallest in Zone 1 and largest in Zone 3.

Table 3. Summary statistics of monthly climatic variables and vegetation indices across the three cultivation zones

Input variables	Unit	Zone 1		Zone 2		Zone 3	
		Average	Standard deviation	Average	Standard deviation	Average	Standard deviation
Air humidity	k	277.154	3.698	284.862	4.045	265.072	4.605
Maximum air temperature	k	297.184	2.872	299.188	4.075	288.254	7.560
Minimum air temperature	k	274.458	3.232	280.770	4.619	262.972	6.858
Total precipitation	m	0.062	0.015	0.130	0.031	0.038	0.028
Actual evaporation	m	-0.0551	0.0091	-0.0752	0.0108	-0.0283	0.0148
Potential evaporation	m	-0.156	0.041	-0.172	0.098	-0.135	0.087
Surface pressure	Pa	95311	6491	92377	8862	78682	16047
Solar radiation	J/m^2	473060750	41905579	428061418	53782068	545368707	79918654
Northward wind speed	m/s	0.090	0.279	-0.052	0.465	0.012	0.667
Eastward wind speed	m/s	0.016	0.432	-0.318	0.427	0.660	0.897
Soil moisture (7–28 cm)	Volume fraction	0.275	0.043	0.389	0.050	0.240	0.125
Soil moisture (28–100 cm)	Volume fraction	0.266	0.049	0.391	0.051	0.240	0.125
EVI	-	0.244	0.067	0.325	0.073	0.118	0.096
NDVI	-	0.399	0.107	0.564	0.124	0.197	0.166
ARVI	-	0.252	0.123	0.466	0.143	0.033	0.180
SAVI	-	0.240	0.061	0.313	0.068	0.125	0.089

3.5. Satellite remote sensing data

The normalized difference vegetation index (NDVI), enhanced vegetation index (EVI), soil-adjusted vegetation index (SAVI), and atmospherically resistant vegetation index (ARVI) were derived from the Moderate Resolution Imaging Spectroradiometer (MODIS) surface reflectance product using the Google Earth Engine platform (Figure 5). Prior to index calculation, all scenes were pre-processed to exclude pixels contaminated by clouds and

shadows based on the StateQA quality assurance band, and to mask invalid reflectance values outside the acceptable range. Following this pre-processing, monthly composite images were generated for each month from 2001 to 2010, and the corresponding vegetation indices were subsequently calculated and extracted.

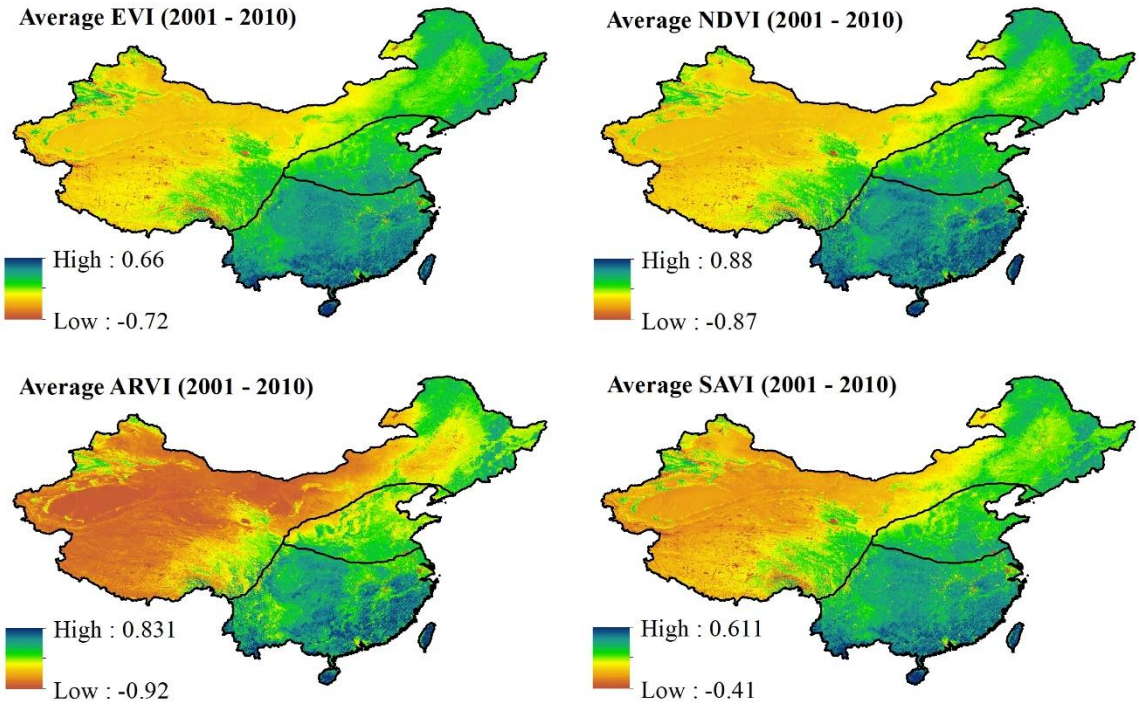


Figure 5. Spatial distribution of average monthly vegetation indices (2001–2010)

The NDVI defined as $(NIR - Red)/(NIR + Red)$, depends on spectral absorption properties in the red and near-infrared (NIR) bands, enhancing its sensitivity to chlorophyll content and usefulness in monitoring vegetation vigor [38]. The EVI formulated as $2.5\times(NIR - Red)/((NIR + 6\times Red - 7.5\times Blue)+1)$, incorporates NIR, red, and blue bands, providing improved sensitivity in high-biomass regions and enhancing responsiveness to variations in canopy structure while reducing atmospheric effects and separating vegetation signals from background influences [17]. The SAVI expressed as $(1+L)\times(NIR - Red)/(NIR + Red + L)$, uses the NIR and red bands with a soil-brightness correction factor (L = 0.5), enabling more accurate vegetation monitoring in areas where sparse canopy cover causes substantial soil background interference [39]. The ARVI calculated as $(NIR-(2\times Red - Blue))/(NIR+(2\times Red - Blue))$, employs NIR, red, and a blue-band correction to mitigate atmospheric scattering, ensuring vegetation signal accuracy under haze, dust, or aerosol-rich conditions [30].

Table 3 presents the summary statistics of these monthly vegetation indices across the three cultivation zones. The statistics show clear differences in the distribution of NDVI, ARVI,

SAVI, and EVI among the zones. Mean values of all four indices are highest in Zone 1, intermediate in Zone 2, and lowest in Zone 3. Additionally, these indices show the least variation (standard deviation) in Zone 1, more variation in Zone 2, and the most variation in Zone 3.

3.6. Machine learning methods

The RF, XGBoost, and LightGBM models were implemented in Python, using the scikit-learn library (version 1.3.2) for RF, the xgboost library (version 3.0.3) for XGBoost, and the lightgbm (version 4.6.0) library for LightGBM.

RF employs bootstrapped samples to generate multiple decision trees and random feature selection at each split, aggregating predictions through majority voting for classification tasks or averaging for regression tasks. This approach reduces variance, mitigates overfitting, and yields stable, interpretable results, particularly on tabular and structured data. RF is relatively simple to tune, robust to noisy data, and performs well across diverse feature types, making it a reliable baseline model in many applications [40,41].

XGBoost is a powerful gradient boosting framework that builds decision trees sequentially, with each new tree correcting the residual errors of the existing ensemble. It incorporates strong regularization to improve generalization, handles missing data natively, and leverages techniques such as shrinkage (learning rate), column subsampling, and tree pruning to enhance performance and prevent overfitting. XGBoost is widely recognized for its high predictive accuracy and efficiency, making it a popular choice in competitive machine learning, though it often requires careful hyperparameter tuning to achieve optimal results [42].

LightGBM is a gradient boosting framework optimized for both computational speed and memory efficiency. It employs a histogram-based algorithm and a leaf-wise tree growth strategy with depth constraints, enabling significantly faster training times and reduced memory usage compared to traditional gradient boosting methods. LightGBM excels at handling large-scale, high-dimensional, or sparse datasets and supports categorical features directly without one-hot encoding. LightGBM offers excellent scalability and accuracy; however, it can be more prone to overfitting on small datasets if not properly tuned, and its leaf-wise growth strategy can produce deeper trees that require careful regularization [42].

3.7. Improved grey wolf optimization (IGWO)

GWO is a nature-inspired technique that simulates the pack hierarchy and predatory strategies of grey wolves. It assigns leadership ranks (α, β, δ) to steer the search process, while the rest of the wolves (ω) follow their lead. The algorithm operates through three main stages: surrounding the prey, pursuing it, and launching the final attack to identify the global best solution [43].

Taking *t* as the current iteration, X as the position vector of a grey wolf, and X_p as the prey's position, the encircling behavior of grey wolves in the GWO algorithm is represented mathematically as follows [43]:

$$D = \left| C \,.X_p(t) - X(t) \right| \tag{1}$$

$$X(t+1) = X_p(t) - A.D \tag{2}$$

The coefficient vectors *C* and *A*, which are used to model the encircling behavior of grey wolves, are defined as follows [43]:

$$A = 2\,.a.r_1 - a(t) \tag{3}$$

$$C = 2.r_2 \tag{4}$$

The random vectors r_1 and r_2 are distributed uniformly within the interval [0,1]. The control parameter *a* decreases linearly from 2 to 0 throughout the iterations, aiming to balance the exploration and exploitation phases. During the first half of the iterations, the algorithm emphasizes exploration, while the second half gradually shifts towards exploitation [43]:

$$a(t) = 2 - (2t) / MaxIter \tag{5}$$

Taking the positions of the of α, β, δ wolves as the top three solutions discovered up to that point, the other wolves adjust their positions by following the direction provided by these leaders [43]:

$$X(t+1) = \frac{1}{3}\left(X_\alpha + X_\beta + X_\delta \right) \tag{6}$$

where X_i represents the position vector of the best *i-th* solution.

The attack phase commences when the prey ceases movement, guided by the parameter a (Eq. 7). During this phase, wolves adjust their positions randomly between their current location and that of the prey, facilitating local exploitation [43].

The IGWO introduces an innovative search approach based on iterative step selection and refinement. It operates through three primary phases: initialization, movement, and selection and updating, as outlined below [43].

The initialization phase consists of randomly distributing *N* wolves throughout the search space, within the defined range of $\left[l_i, u_j \right]$:

$$X_{ij}(t) = l_j + rand_j[0,1] \times \left(u_j - l_j \right), i\grave{o}[1,N], j\grave{o}[1,D] \tag{7}$$

The position of the wolf *i* at iteration *t* is represented as $X_i(t) = [X_{i1}, X_{i2}, \ldots, X_{iD}]$, refers to the number of dimensions in the problem. The entire population of wolves is stored in a matrix of

size ($N \times D$) matrix named population. The fitness score for each wolf, $X_i(t)$, is determined using the fitness function $f(X_i(t))$ [43].

The IGWO introduces an additional movement strategy called the dimension learning-based hunting (DLH) approach. In this strategy, each wolf adjusts its behavior by learning from its neighbors, providing an alternative method for updating its position vector, $X_i(t)$.

The DLH search strategy addresses several key issues of the standard GWO, such as premature convergence, early population homogenization, and vulnerability to local optima. It does this by promoting individualized hunting behaviors shaped by the influence of neighboring wolves. This process starts with calculating a radius $R_i(t)$, which represents the Euclidean distance between the current position $X_i(t)$ and the alternative candidate position $X_{i-GWO}(t+1)$ [43]:

$$R_i(t) = \left| X_i(t) - X_{i-GWO}(t+1) \right| \tag{8}$$

The set of neighbors corresponding to $X_i(t)$, denoted as $N_i(t)$, is defined as follows:

$$N_i(t) = [X_i(t) \mid D_i(X_i(t), X_j(t)) \le R_i(t), X_j(t) \hat{o} Pop] \tag{9}$$

where D_i represents the Euclidean distance function. Once the neighbors are defined, a multi-neighbor learning approach is applied as follows [43]:

$$X_{i-DLH,d}(t+1) = X_{i,d}(t) + rand \times (X_{n,d}(t) - X_{r,d}(t)) \tag{10}$$

In this equation, the *d-th* dimension of $X_{i-DLH,d}$ is calculated using the *d-th* dimension of a randomly chosen neighbor $X_{n,d}(t)$ from the set $N_i(t)$, as well as the *d-th* dimension of a randomly selected wolf $X_{r,d}(t)$ from population.

In the selection and updating phase, the optimal solution is determined by assessing and comparing the fitness scores of two candidate solutions [43]:

$$X_i(t+1) = \begin{cases} X_{i-DLH}(t+1) \; if \; f(X_{i-GWO}) > f(X_{i-DLH}) \\ X_{i-GWO}(t+1) \qquad\qquad otherwise \end{cases} \tag{11}$$

Then, $X_i(t)$ is updated using the chosen solution if the fitness score of the selected candidate is lower than that of $X_i(t)$. If the score is not lower, $X_i(t)$ remains unchanged. This process is applied to each individual, and the iteration counter is incremented. The search continues until the maximum number of iterations is reached [43].

Next, to update the new position of $X_i(t+1)$, if the fitness score of the selected candidate is lower than that of $X_i(t)$, $X_i(t)$. is replaced by the selected candidate. Otherwise, $X_i(t)$ remains unchanged within the population.

3.8. Improved whale optimization algorithm (IWO)

The WOA is an innovative optimization method introduced by Mirjalili and Lewis in 2016. It is a population-based algorithm that utilizes a set of random candidate solutions and simulates the social behavior patterns of humpback whales. The algorithm follows three key principles to adjust the candidate positions iteratively: encircling prey, spiral updating of positions, and prey searching [44].

When encircling prey ($with -1 < A < 1 and\ p < 0.5$), the position of a candidate is updated as follows [45]:

$$\vec{D} = \left| \vec{C}.\vec{X}^* - \vec{X}(t) \right| \tag{12}$$

$$\vec{X}(t+1) = \left| \vec{X}^*(t) - \vec{A}.\vec{D} \right| \tag{13}$$

where p is a random value between 0 and 1, and $\vec{X}^*$ represents the best solution found in the current generation. The values of $\vec{A}$ and $\vec{C}$ are calculated as follows:

$$\vec{A} = \left| 2\vec{a}.\vec{r} - \vec{a} \right| \tag{14}$$

$$\vec{C} = 2.\vec{r} \tag{15}$$

where $\vec{a}$ decreases linearly from 2 to 0, and $\vec{r}$ is a random vector within the range [0, 1].

Prey searching ($with -1 < A < 1\ \&\ p < 0.5$) is employed during the exploration phase to enable the WOA to perform a global search. This process is similar to encircling prey but uses a random candidate solution ($\vec{X}_{rand}$) instead of the optimal solution $\vec{X}^*$ [45]. The update is calculated as follows:

$$\vec{D} = \left| \vec{C}.\vec{X}_{rand} - \vec{X}(t) \right| \tag{16}$$

$$\vec{X}(t+1) = \left| \vec{X}_{rand}(t) - \vec{A}.\vec{D} \right| \tag{17}$$

Spiral position updating ($with\ p \geq 0.5$) involves both encircling and spiral movement strategies during the exploitation phase of the WOA. The positions of individuals are modified using the spiral updating mechanism as follows [45]:

$$\vec{X}(t+1) = \vec{D}'.e^{bl}.\cos(2\pi l) + \vec{X}^*(t) \tag{18}$$

Here, $\vec{D} = \left| \vec{X}^*(t) - \vec{X}(t) \right|$ represents the distance between the current candidate solution and the best solution found so far, b is a constant parameter, and $l \grave{o} [-1,1]$.

The IWO algorithm is built on the idea of combining the WOA's operators with the mutation operator from differential evolution (DE) to improve the balance between exploration and exploitation. It is structured into two primary phases: exploration and exploitation [45]. During the exploration phase, individuals' positions are adjusted when $rand < \lambda$:

$$\lambda = \frac{-t}{t_{max}} + 1 \tag{19}$$

Here, t represents the current generation, and t_{max} is the maximum number of generations. The parameter λ controls the exploration-exploitation balance and decreases from 1 to 0 as generation progress. This allows individuals to explore more in the early generations, gradually shifting toward exploitation as the process continues [45].

The exploration phase of IWO involves using the mutation operation from DE along with the prey-locating behavior of WOA. The inclusion of DE's mutation operator enhances the ability to search through the solution space effectively [46]. In the exploitation phase, IWO follows a process similar to WOA but introduces an elitism strategy. This strategy involves selecting the better solution between the parent (X_i .) and the offspring (U_i) to determine the new position for the *i*-th solution. Additionally, it is important to ensure that solutions remain within the defined boundary limits. If the solutions violate these boundaries, a correction rule is applied as follows [45]:

$$X_i(j) = \begin{cases} \delta_j + rndreal(0,1)(\mu_j - \delta_j) if\ X_i(j) < \delta_j \\ \mu_j + rndreal(0,1)(\mu_j - \delta_j) if\ X_i(j) < \mu_j \end{cases}. \tag{20}$$

Here, μ_j represents the upper bound, and δ_j denotes the lower bound of the *j*-th dimension. $X_i(j)$ refers to the value of the *j*-th dimension for the *i*-th candidate. $rndreal(0,1)$ is a random value selected from the interval [0, 1].

4. Results

4.1. Evaluation of period-specific modeling performance

To guide the optimal integration of climatic and remote sensing variables in wheat yield prediction, six scenarios were designed. Table 4 summarizes the input variables within each period-based scenario. In scenarios 1–3, remote sensing data averaged over the entire growing season were combined with climate data averaged over early, peak, or late growth periods. In scenarios 4–6, climate data averaged over the entire growing season were combined with remote sensing data averaged over early, peak, or late periods.

Wheat yield modeling was evaluated across the six input scenarios using the Willmott index (WI). For scenarios 1–3 (i), performance was compared against a baseline model constructed solely from remote sensing inputs. In scenarios 4–6 (ii), performance was compared against a baseline model constructed solely from climate inputs (Figure 6).

Table 4. Remote sensing and climatic input variables used across data input scenarios for LightGBM, XGBoost, and RF models

Scenario	Primary focus	Remote sensing inputs at pixel level	Climatic inputs at pixel level
1	Early climate	Full growing season (average)	Early growth (average)
2	Peak climate		Peak growth (average)
3	Late climate		Late growth (average)
4	Early remote sensing	Early growth (average)	Full growing season (average)
5	Peak remote sensing	Peak growth (average)	
6	Late remote sensing	Late growth (average)	

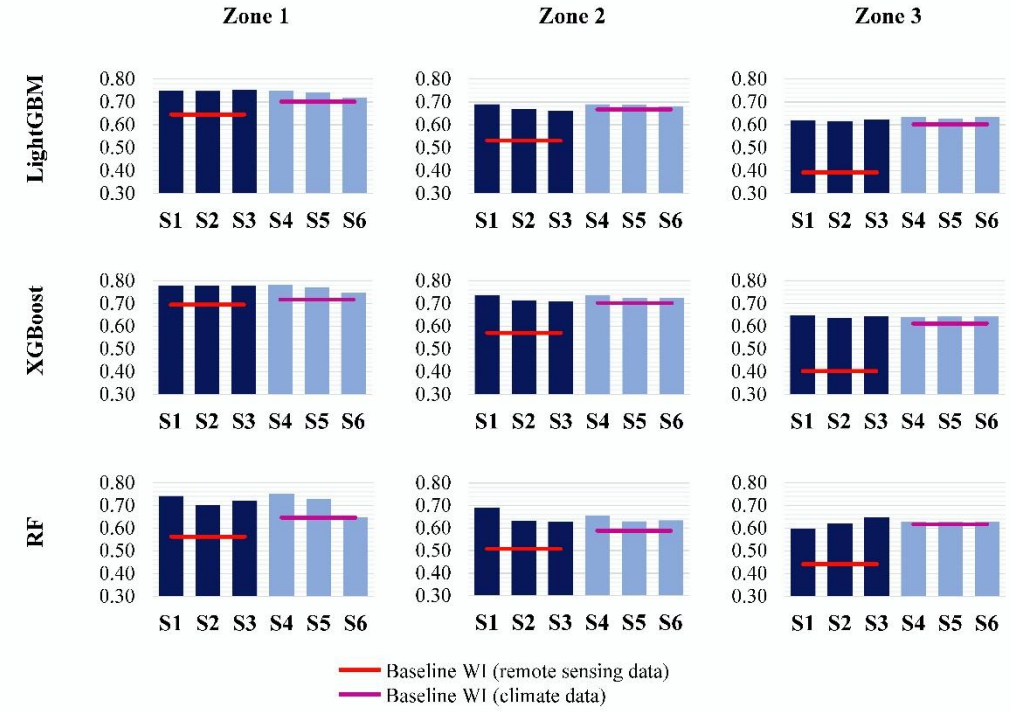


Figure 6. Evaluation of wheat yield prediction using the WI across six data input scenarios

On average, the highest WI improvements were observed for remote sensing + early climate (0.167), remote sensing + late climate (0.157), remote sensing + peak climate (0.152), climate + early remote sensing (0.046), climate + peak remote sensing (0.037), and climate + late remote sensing (0.022), respectively. Overall, the greatest gains were achieved when full-season remote sensing was combined with climate periods, yielding WI improvements ranging from 0.084 to 0.246 (average: 0.159). In contrast, pairing full-season climate data with remote

sensing periods produced smaller improvements, ranging from 0.001 to 0.105 (average: 0.035). WI improvement patterns also varied across zones.

In addition to differed WI gains in data input scenarios, improvement patterns also varied across geographical zones. On average, Zone 3 exhibited the highest WI gains (0.119), followed by Zone 2 (0.088) and Zone 1 (0.083).

4.2. Monte Carlo sensitivity analysis

After identifying remote sensing + early climate as the best-performing scenario, a Monte Carlo sensitivity analysis was conducted in Python (version 3.8.8) to quantify the relative importance of input variables and to evaluate model robustness under varying input conditions. RF was selected for the analysis, given its frequent application in crop yield prediction (Table 2).

In each iteration, a single feature was randomly selected and perturbed by shuffling its values across all records. The dataset was then divided into training (80%) and testing (20%) subsets, and the model was trained and validated using the perturbed data. This procedure was repeated 1,000 times, and the WI was averaged for each feature. Higher ΔWI (Unperturbed WI – Perturbed WI) values indicated a greater influence of the corresponding feature on crop yield prediction.

The sensitivity analysis results are presented in Figure 7. Accordingly, climatic variables (average ΔWI of 0.235) had the greatest impact on crop yield prediction compared to remote sensing predictors (average ΔWI of 0.077). Among the climatic variables, northward wind speed (0.335), soil moisture at 7–28 cm (0.332), total precipitation (0.327), and actual evaporation (0.319) exhibited the highest influence. Among remote sensing indices, ARVI (0.092), SAVI (0.076), EVI (0.071), and NDVI (0.067) were the most influential features.

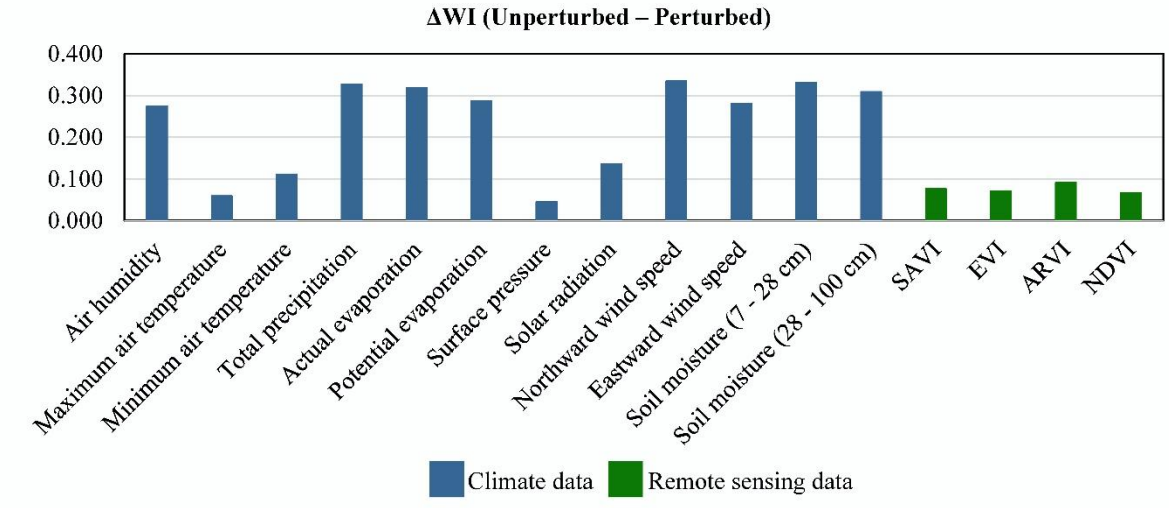


Figure 7. Monte Carlo sensitivity analysis of modeling features using ΔWI values

4.3. Enhanced crop yield modeling

The pyMetaheuristic library (version 1.9.4) was used to implement the IGWO and IWO algorithms for optimizing wheat yield models. The optimization procedure included tuning hyperparameters trained on the training set and aimed to minimize the WI of predictions evaluated on the testing set. This process ran for 500 iterations (no other stopping criteria were applied) with a hunting party/pack size of 25, searching the predefined value range for each hyperparameter. Table 5 presents the selected hyperparameters and their corresponding value ranges used in the optimization process.

Table 5. Selected hyperparameters for optimization [47]

Model	Hyperparameter	Value range
RF	Number of trees	150 – 800
	Maximum depth of each tree	2 – 30
	Minimum samples to split a node	2 – 10
	Minimum samples required in a leaf node	2 – 10
XGBoost	Number of trees	150 – 800
	Step size shrinkage to prevent overfitting	0.01 – 0.2
	Maximum depth of each tree	2 – 30
	Minimum weight of instances in a child node	1 – 5
LightGBM	Number of trees	150 – 800
	Step size shrinkage to prevent overfitting	0.01 – 0.2
	Maximum depth of each tree	2 – 30
	Minimum weight of instances in a child node	1 – 5

Prior to the optimization, feature selection was conducted within the remote sensing + early-season climate scenario using sensitivity analysis (Figure 7). Accordingly, surface temperature, maximum and minimum air temperature, and solar radiation were excluded in modeling.

The performance of IGWO and IWO was compared with the baseline (random search-optimized) versions (Figure 8). All three models demonstrated improvements across the cultivation zones when combined with IGWO and IWO. Specifically, IGWO achieved gains ranging from 0.021 to 0.038 (average: 0.028), while IWO produced improvements ranging from 0.016 to 0.031 (average: 0.021).

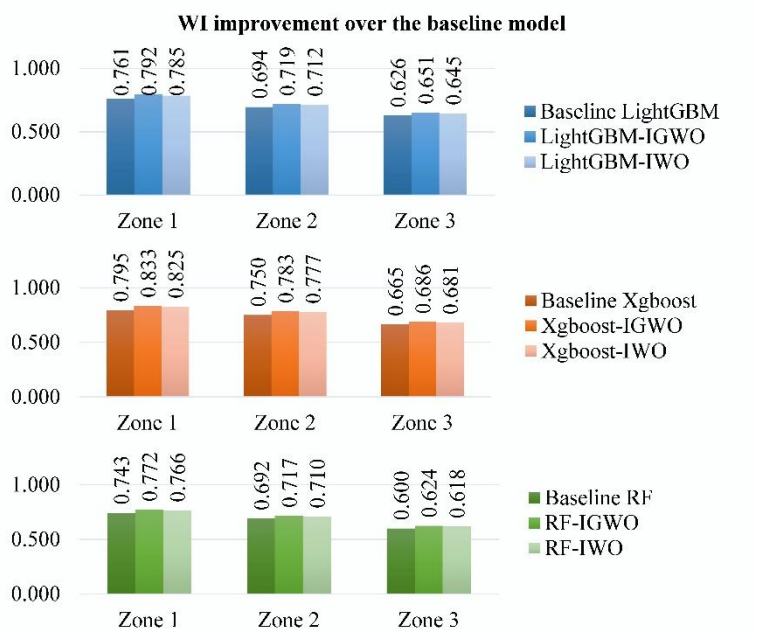


Figure 8. Performance comparison between baseline and optimized models across three cultivation zones

As IGWO achieved greater WI improvements, it was selected for further analysis. The predictive performance of the wheat yield models optimized with IGWO was then validated using R2, WI, and RMSE (Figure 9). XGBoost-IGWO demonstrated the highest performance, with its highest accuracy observed in Zone 1 (R2 = 0.755, WI = 0.833, RMSE = 0.495 t/ha). LightGBM-IGWO achieved comparable predictive accuracy, with R2 values between 0.611 and 0.717. RF-IGWO maintained the lowest performance among the three optimized models, with R2 values ranging from 0.584 to 0.698.

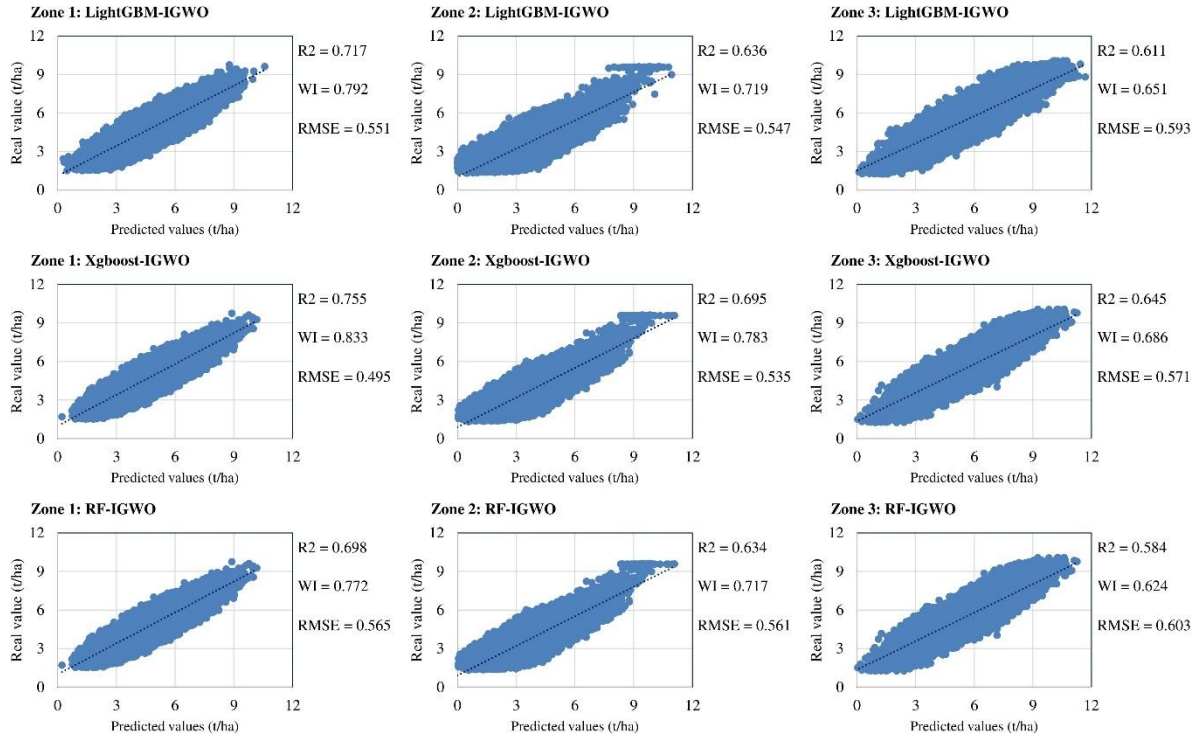


Figure 9. Validation of crop yield predictions using machine-learning models enhanced through optimization techniques

Beyond differing model performance, a spatial trend was evident across the study area. On average, Zone 1 exhibited the highest validation accuracy (R2 = 0.723, WI = 0.799, RMSE = 0.537 t/ha), whereas Zones 2 (R2 = 0.655, WI = 0.740, RMSE = 0.548 t/ha) and Zone 3 (R2 = 0.613, WI = 0.654, RMSE = 0.589 t/ha) showed comparatively lower values. Nevertheless, RMSE values remained below approximately 0.603 t/ha for all optimized models, indicating acceptable prediction errors for regional cropping systems.

4.4. Uncertainty analysis

Effective uncertainty analysis aims to balance coverage with precision by seeking high PICP (reliability against the nominal confidence level) and low PIW (precision of predictions). PICP measures how often true values fall within predicted intervals and reflects reliability against the nominal confidence level. PIW quantifies the average interval width and indicates the precision of predictions. Table 6 summarizes the computed PIW and PICP values for the predictive models.

Overall, all models achieved PICP values close to their nominal levels, confirming good calibration of their prediction intervals. XGBoost-IGWO showed superior performance in interval width across all confidence levels. In Zone 1, it achieved the smallest PIWs, namely 0.073 (70%), 0.120 (80%), 0.148 (90%), and 0.193 (95%). In contrast, the weakest performance

was observed for RF-IGWO in Zone 3, with PIWs of 0.218 (70%), 0.291 (80%), 0.331 (90%), and 0.395 (95%).

Table 6. Uncertainty analysis of optimized wheat yield models

	Nominal	RF-IGWO		XGBoost-IGWO		LightGBM-IGWO	
		PICP	PIW	PICP	PIW	PICP	PIW
Zone 1	70%	0.668	0.119	0.666	0.042	0.666	0.106
	80%	0.762	0.192	0.76	0.089	0.76	0.179
	90%	0.838	0.232	0.26	0.117	0.27	0.219
	95%	0.864	0.296	0.862	0.162	0.863	0.283
Zone 2	70%	0.668	0.156	0.667	0.075	0.668	0.155
	80%	0.762	0.229	0.761	0.122	0.762	0.228
	90%	0.838	0.269	0.27	0.15	0.838	0.268
	95%	0.868	0.333	0.864	0.195	0.865	0.332
Zone 3	70%	0.671	0.218	0.669	0.139	0.669	0.2
	80%	0.765	0.291	0.763	0.186	0.763	0.273
	90%	0.844	0.331	0.843	0.214	0.844	0.313
	95%	0.873	0.395	0.869	0.259	0.871	0.377

5. Discussion

Consistent with previous studies [2,23,29], the present results confirm that integrating climatic predictors with satellite-derived vegetation indices substantially enhances wheat yield prediction compared to reliance on a single data source. Combining full-season remote sensing data with climate variables aggregated over key phenological periods yielded markedly higher WI gains, indicating that climatic information provides complementary explanatory power beyond spectral indicators alone. This outcome reflects that satellite observations primarily capture the integrated canopy response to environmental forcing, whereas climatic variables directly represent the biophysical drivers of crop growth and yield formation, particularly during sensitive developmental stages such as grain filling [48]. The results therefore highlight the synergistic value of fusing Earth observation data with ground-based meteorological information in regional crop yield modeling frameworks.

The period-specific modeling results further emphasize the temporal sensitivity of satellite remote-sensing inputs in yield prediction. Stronger performance gains associated with early-season remote-sensing observations suggest that satellite signals acquired during canopy establishment and early vegetative growth contribute more effectively to yield prediction than those obtained later in the season. Early-season spectral responses are better suited to capturing

initial crop vigor, stand establishment, and the early manifestation of climatic or biotic stressors that ultimately shape yield trajectories [49]. In contrast, mid- to late-season satellite signals are increasingly influenced by canopy closure, senescence processes, chlorophyll degradation, and post-anthesis physiological dynamics, which can obscure stress signatures and reduce their incremental predictive value [17]. Moreover, early-season vegetation indices more directly represent critical phenological transitions such as tillering and stem elongation, which are closely linked to final grain number and yield potential. These findings underscore the importance of phenology-aware integration strategies that prioritize early-season satellite indicators when designing yield prediction models.Complementing the period-specific modeling results, the Monte Carlo sensitivity analysis demonstrates that climatic predictors exert a stronger overall influence on wheat yield prediction than satellite-derived variables. The dominance of precipitation-related and moisture-linked climatic variables aligns with extensive evidence highlighting the central role of water availability in regulating wheat productivity under both irrigated [50] and rainfed [51] conditions. The relatively limited influence of temperature-based predictors, including maximum and minimum air temperature and solar radiation, is consistent with earlier findings suggesting that their effects may be indirect or moderated by water availability and phenological timing [17]. The high importance of wind speed can be attributed to its documented effects on evapotranspiration demand, canopy microclimate, biomass accumulation, and lodging risk during stem elongation [17,52]. Similarly, soil moisture at 7–28 cm depth emerged as a key predictor, reflecting its role in governing root-zone water availability and sustaining vegetative growth [53]. Actual evaporation provided an integrative measure of crop water status, capturing the combined influence of atmospheric demand and soil moisture dynamics more effectively than individual climatic variables [54].

Although satellite-derived vegetation indices contributed less overall than climatic predictors, their relative importance remains notable, particularly for indices designed to mitigate atmospheric and background effects. The higher contribution of the ARVI relative to NDVI and EVI is consistent with its documented robustness under variable atmospheric conditions [30,55]. The relatively similar influence magnitudes among indices suggest inherent limitations in their ability to capture fine-scale structural and physiological variability when used individually, reinforcing the value of combining multiple indices to better represent crop canopy dynamics across growth stages.

The performance differences among XGBoost, LightGBM, and RF were minimal, with XGBoost generally exhibiting the highest predictive accuracy and lower PIW values. This finding is broadly consistent with prior comparative studies reporting the frequent superiority

of XGBoost over RF [22-24]; however, instances of RF outperforming XGBoost have also been documented [2,34]. Such discrepancies may reflect the influence of regional characteristics, suggesting that the specific attributes of the study area can significantly affect the relative performance of these machine learning models.

While both metaheuristics enhanced model accuracy, IGWO consistently yielded greater improvements than IWO. IGWO's superior performance stems from its advanced optimization strategy, which balances exploration and exploitation more effectively. By dynamically adjusting search intensity and employing multiple leader wolves (alpha, beta, and delta) rather than a single best solution, IGWO maintains diversity in the search space while providing directional focus. This reduces the likelihood of convergence to local optima and enables more reliable feature selection and hyperparameter optimization, producing consistently high-accuracy outcomes across datasets [41].

The RMSE values obtained using the XGBoost-IGWO framework (zone 1: 0.495 t/ha; zone 2: 0.535 t/ha; zone 3: 0.571 t/ha) indicate a competitive level of predictive accuracy relative to recent studies. For instance, Fu, et al. [56] reported an RMSE of 0.593 t/ha when applying an IGWO-CNN model to MODIS vegetation indices in the Huang–Huai–Hai wheat production region of China, whereas Huang, et al. [57] achieved an RMSE of 0.595 t/ha through the assimilation of MODIS and Landsat surface reflectance into the WOFOST-PROSAIL crop growth model in southern Hebei and northern Shandong. Similarly, Cao, et al. [58] obtained an RMSE of 0.78 t/ha using machine-learning methods informed by global dynamical atmospheric predictions in northern China. In contrast, notably lower error values have been documented in studies employing hyperspectral data. Cheng, et al. [59], for example, achieved an RMSE of 0.201 t/ha using ZY-1 02D satellite hyperspectral imagery combined with an LSTM model in Beijing's Changping District.

Model performance differed across the three cultivation zones. Prediction accuracy was highest in Zone 1 and lowest in Zone 3. These patterns are consistent with the descriptive statistics presented in Table 4. Zone 1 is characterized by lower dispersion in both climatic variables and vegetation indices, whereas Zone 3 exhibits higher standard deviations and wider ranges across these variables, indicating greater environmental heterogeneity. Previous studies have also shown that increased climatic variability can reduce crop yield prediction accuracy, as demonstrated for maize yield estimation in Southwest China [60]. In addition, the larger spatial extent and shorter growing season of Zone 3 may further increase within-zone variability, complicating the characterization of crop conditions using climatic and remote sensing inputs alone. Consistent with the findings of Jeong, et al. [18], these results highlight the importance of

diverse and representative datasets for improving wheat yield prediction across heterogeneous cultivation zones.

Overall, these results collectively demonstrate that the proposed framework enables accurate and scalable yield prediction for operational decision-making. Notwithstanding these contributions, the limited temporal coverage and the focus on a single crop within one country may restrict the broader applicability of the results. Future research should therefore extend this framework across diverse crops, geographic regions, and longer time series to further evaluate its robustness and transferability under varying agricultural and environmental conditions.

5. Conclusion

This study evaluated the integration of climatic and remote sensing variables for wheat yield forecasting across three agro-climatic zones in China using machine learning models optimized with metaheuristic algorithms. Multi-source data fusion improved prediction accuracy over single-source approaches, with the highest improvements achieved by combining full-season vegetation indices and early-season climatic variables. Sensitivity analysis highlighted the dominant role of water-related predictors, while individual vegetation indices showed lower influence.

All models benefited from optimization, with XGBoost-IGWO yielding the highest accuracy. Spatial analysis showed the strongest performance in Zone 1 and the weakest in Zone 3, reflecting greater variability in climatic and vegetation conditions. These findings emphasize the importance of representative data and phenology-aware variable selection when forecasting wheat yield across heterogeneous cultivation zones.

Declarations

Conflicts of Interest

The authors have no competing interests to declare.

Funding Statement

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

AI Use Statement: AI tools were utilized solely for language editing and grammar checking during the preparation of this manuscript. All text, ideas, and interpretations were reviewed, verified, and approved by the authors.

References

1 Arshad, S., Kazmi, J. H., Javed, M. G. & Mohammed, S. Applicability of machine learning techniques in predicting wheat yield based on remote sensing and climate data in Pakistan, South Asia. *European Journal of Agronomy* **147**, 126837 (2023). https://doi.org/10.1016/j.eja.2023.126837

2 Khechba, K. *et al.* The impact of spatiotemporal variability of environmental conditions on wheat yield forecasting using remote sensing data and machine learning. *International Journal of Applied Earth Observation and Geoinformation* **136**, 104367 (2025). https://doi.org/10.1016/j.jag.2025.104367

3 Mottaleb, K., Sonder, K., Lopez-Ridaura, S. & Frija, A. Wheat consumption dynamics in selected countries in Asia and Africa: Implications for wheat supply by 2030 and 2050. (2021).

4 Grosse-Heilmann, M., Cristiano, E., Deidda, R. & Viola, F. Durum wheat productivity today and tomorrow: A review of influencing factors and climate change effects. *Resources, Environment and Sustainability* **17**, 100170 (2024). https://doi.org/10.1016/j.resenv.2024.100170

5 Chen, S. *et al.* Dynamic within-season irrigation scheduling for maize production in Northwest China: A Method Based on Weather Data Fusion and yield prediction by DSSAT. *Agricultural and Forest Meteorology* **285**, 107928 (2020). https://doi.org/10.1016/j.agrformet.2020.107928

6 Jégo, G., Pattey, E. & Liu, J. Using Leaf Area Index, retrieved from optical imagery, in the STICS crop model for predicting yield and biomass of field crops. *Field Crops Research* **131**, 63–74 (2012). https://doi.org/10.1016/j.fcr.2012.02.012

7 Tariq, A., Yan, J., Gagnon, A. S., Riaz Khan, M. & Mumtaz, F. Mapping of cropland, cropping patterns and crop types by combining optical remote sensing images with decision tree classifier and random forest. *Geo-Spatial Information Science* **26**, 302–320 (2023).

8 Xu, H. *et al.* Machine learning approaches can reduce environmental data requirements for regional yield potential simulation. *European Journal of Agronomy* **129**, 126335 (2021). https://doi.org/10.1016/j.eja.2021.126335

9 Haseeb, M., Tahir, Z., Mahmood, S. A. & Tariq, A. Winter wheat yield prediction using linear and nonlinear machine learning algorithms based on climatological and remote sensing data. *Information Processing in Agriculture* (2025). https://doi.org/10.1016/j.inpa.2025.02.004

10 Burdett, H. & Wellen, C. Statistical and machine learning methods for crop yield prediction in the context of precision agriculture. *Precision agriculture* **23**, 1553–1574 (2022). https://doi.org/10.1007/s11119-022-09897-0

11 Leng, G. & Hall, J. W. Predicting spatial and temporal variability in crop yields: an inter-comparison of machine learning, regression and process-based models. *Environmental Research Letters* **15**, 044027 (2020). https://doi.org/10.1088/1748-9326/ab7b24

12 Raeisi, Z., Roshanzamir, A., Lomer, F. A. & Lashaki, R. A. YOLOv8 with innovative dilated residual and attention modules for mammographic tumor detection. *Computers and Electrical Engineering* **130**, 110903 (2026).

13 Li, L., Liu, L., Cheng, F., He, Y. & Zhong, Z. CN-UNet: ConvNeXt UNet With Slicing-Aided Hyper Segmentation for Infrared Small Target Detection. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **19**, 84–98 (2026). https://doi.org/10.1109/JSTARS.2025.3631384

14 Li, G. *et al.* Thawing permafrost under Qinghai-Xizang Highway and its impacts on road performance based on multi-source observed data. *Cold Regions Science and Technology* **244**, 104828 (2026). https://doi.org/10.1016/j.coldregions.2026.104828
15 Van Klompenburg, T., Kassahun, A. & Catal, C. Crop yield prediction using machine learning: A systematic literature review. *Computers and electronics in agriculture* **177**, 105709 (2020). https://doi.org/10.1016/j.compag.2020.105709
16 Shen, Z. *et al.* YCANet: Target Detection for Complex Traffic Scenes Based on Camera-LiDAR Fusion. *IEEE Sensors Journal* **24**, 8379–8389 (2024). https://doi.org/10.1109/JSEN.2024.3357826
17 Zhou, W. *et al.* Integrating climate and satellite remote sensing data for predicting county-level wheat yield in China using machine learning methods. *International Journal of Applied Earth Observation and Geoinformation* **111**, 102861 (2022). https://doi.org/10.1016/j.jag.2022.102861
18 Jeong, S., Ko, J., Ban, J.-o., Shin, T. & Yeom, J.-m. Deep learning-enhanced remote sensing-integrated crop modeling for rice yield prediction. *Ecological informatics* **84**, 102886 (2024). https://doi.org/10.1016/j.ecoinf.2024.102886
19 Chen, B. H. Minimum standards for evaluating machine-learned models of high-dimensional data. *Frontiers in Aging* **3**, 901841 (2022). https://doi.org/10.3389/fragi.2022.901841
20 Chen, W. *et al.* Hybrid dimensionality reduction forest with pruning for high-dimensional data classification. *IEEE Access* **8**, 40138–40150 (2020). https://doi.org/10.1109/ACCESS.2020.2975905
21 Tran, C. T. Ensemble learning approaches for classification with high-dimensional data. *Journal of Science and Technique-Section on Information and Communication Technology* **12** (2023). https://doi.org/10.1080/10095020.2022.2100287
22 Gharakhanlou, N. M. & Perez, L. From data to harvest: Leveraging ensemble machine learning for enhanced crop yield predictions across Canada amidst climate change. *Science of The Total Environment* **951**, 175764 (2024). https://doi.org/10.1016/j.scitotenv.2024.175764
23 Bouras, E. H. *et al.* Cereal yield forecasting with satellite drought-based indices, weather data and regional climate indices using machine learning in Morocco. *Remote Sensing* **13**, 3101 (2021). https://doi.org/10.3390/rs13163101
24 Sah, S., Haldar, D., Singh, R., Das, B. & Nain, A. S. Rice yield prediction through integration of biophysical parameters with SAR and optical remote sensing data using machine learning models. *Scientific reports* **14**, 21674 (2024). https://doi.org/10.1038/s41598-024-72624-4
25 Razavi-Termeh, S. V. *et al.* Solving water scarcity challenges in arid regions: A novel approach employing human-based meta-heuristics and machine learning algorithm for groundwater potential mapping. *Chemosphere* **363**, 142859 (2024). https://doi.org/10.1016/j.chemosphere.2024.142859
26 Yang, Z. *et al.* A Learning-Aided Unsupervised Method for Sparse Aperture ISAR Imaging and Autofocusing. *IEEE Transactions on Aerospace and Electronic Systems* **62**, 350–367 (2026). https://doi.org/10.1109/TAES.2025.3620790
27 Behnam, R., Baghaee, H. R., Gharehpetian, G. B., Ahmadiahangar, R. & Rosin, A. Resilient Reliability/Loss-Based Distribution Network Reconfiguration: A Strategy Against FDI Attacks During State Estimation Procedure. *IEEE Transactions on Network Science and Engineering* **12**, 1994–2006 (2025). https://doi.org/10.1109/TNSE.2025.3542632
28 Farahani, M. & Khodaygan, S. Minimization of non-repeatable runout (NRRO) in high-speed spindle bearings. Report No. 0148-7191, (SAE Technical Paper, 2021).

29 Ramos, M. I., Cubillas, J. J., Córdoba, R. M. & Ortega, L. M. Improving early prediction of crop yield in Spanish olive groves using satellite imagery and machine learning. *PloS One* **20**, e0311530 (2025). https://doi.org/10.1371/journal.pone.0311530

30 Raza, A. *et al.* Improving Wheat Yield Prediction with Multi-Source Remote Sensing Data and Machine Learning in Arid Regions. *Remote Sensing* **17**, 774 (2025). https://doi.org/10.3390/rs17050774

31 Ashfaq, M. *et al.* Accurate wheat yield prediction using machine learning and climate-NDVI data fusion. *IEEE Access* **12**, 40947–40961 (2024). https://doi.org/10.1109/ACCESS.2024.3376735

32 Peng, D. *et al.* A deep–learning network for wheat yield prediction combining weather forecasts and remote sensing data. *Remote Sensing* **16**, 3613 (2024). https://doi.org/10.3390/rs16193613

33 Song, X.-P., Li, H., Potapov, P. & Hansen, M. C. Annual 30 m soybean yield mapping in Brazil using long-term satellite observations, climate data and machine learning. *Agricultural and forest meteorology* **326**, 109186 (2022). https://doi.org/10.1016/j.agrformet.2022.109186

34 Gómez, D., Salvador, P., Sanz, J. & Casanova, J. L. Modelling wheat yield with antecedent information, satellite and climate data using machine learning methods in Mexico. *Agricultural and Forest Meteorology* **300**, 108317 (2021). https://doi.org/10.1016/j.agrformet.2020.108317

35 Zhonghu, H., Zuoji, L. & Loongjun, W. Classification of Chinese wheat regions based on quality. (2002).

36 Muñoz Sabater, J. (2019).

37 Vermote, E. MODIS/Aqua Surface Reflectance 8-Day L3 Global 500m SIN Grid V061. *NASA EOSDIS Land Processes Distributed Active Archive Center (DAAC) data set*, MYD09A01. 061 (2021). https://doi.org/10.5067/MODIS/MOD09A1.061

38 Chang, Y., Winkler, A. J., Noori, A., Knyazikhin, Y. & Myneni, R. B. Precipitation leads the long-term vegetation increase in the conterminous United States drylands. *Environmental Research Letters* **20**, 044006 (2025). https://doi.org/10.1088/1748-9326/adb985

39 Fadl, M. E., AbdelRahman, M. A., El-Desoky, A. I. & Sayed, Y. A. Assessing soil productivity potential in arid region using remote sensing vegetation indices. *Journal of Arid Environments* **222**, 105166 (2024). https://doi.org/10.1016/j.jaridenv.2024.105166

40 Farhangi, F., Sadegh-Niaraki, A., Vahid Razavi-Termeh, S. & Nahvi, A. Driver drowsiness modeling based on spatial factors and electroencephalography using machine learning methods: A simulator study. *Transportation Research Part F: Traffic Psychology and Behaviour* **98**, 123–140 (2023). https://doi.org/10.1016/j.trf.2023.08.007

41 Farhangi, F., Sadeghi-Niaraki, A., Razavi-Termeh, S. V., Farhangi, F. & Choi, S.-M. Cognitive modeling based on geotagged pictures of urban landscapes using mobile electroencephalogram signals and machine learning models. *Cognitive Systems Research* **90**, 101324 (2025). https://doi.org/10.1016/j.cogsys.2025.101324

42 Bentéjac, C., Csörgő, A. & Martínez-Muñoz, G. A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review* **54**, 1937–1967 (2021). https://doi.org/10.1007/s10462-020-09896-5

43 Nadimi-Shahraki, M. H., Taghian, S. & Mirjalili, S. An improved grey wolf optimizer for solving engineering problems. *Expert Systems with Applications* **166**, 113917 (2021). https://doi.org/10.1016/j.eswa.2020.113917

44 Mirjalili, S. & Lewis, A. The whale optimization algorithm. *Advances in engineering software* **95**, 51–67 (2016). https://doi.org/10.1016/j.advengsoft.2016.01.008

45 Mostafa Bozorgi, S. & Yazdani, S. IWOA: An improved whale optimization algorithm for optimization problems. *Journal of Computational Design and Engineering* **6**, 243–259 (2019). https://doi.org/10.1016/j.jcde.2019.02.002

46 Gong, W., Cai, Z. & Ling, C. X. DE/BBO: a hybrid differential evolution with biogeography-based optimization for global numerical optimization. *Soft Computing* **15**, 645–665 (2010). https://doi.org/10.1007/s00500-010-0591-1

47 Nagunwa, T. Comparative Analysis of Nature-Inspired Metaheuristic Techniques for Optimizing Phishing Website Detection. *Analytics* **3**, 344–367 (2024). https://doi.org/10.3390/analytics3030019

48 Barlow, K. M., Christy, B. P., O'Leary, G. J., Riffkin, P. A. & Nuttall, J. G. Simulating the impact of extreme heat and frost events on wheat crop production: A review. *Field crops research* **171**, 109–119 (2015). https://doi.org/10.1016/j.fcr.2014.11.010

49 Behmann, J., Mahlein, A.-K., Rumpf, T., Römer, C. & Plümer, L. A review of advanced machine learning methods for the detection of biotic stress in precision crop protection. *Precision agriculture* **16**, 239–260 (2015). https://doi.org/10.1007/s11119-014-9372-7

50 Cao, J. *et al.* Identifying the contributions of multi-source data for winter wheat yield prediction in China. *Remote Sensing* **12**, 750 (2020). https://doi.org/10.3390/rs12050750

51 Jeong, J. H. *et al.* Random forests for global and regional crop yield predictions. *PloS one* **11**, e0156571 (2016). https://doi.org/10.1371/journal.pone.0156571

52 Araghi, A., Maghrebi, M. & Olesen, J. E. Effect of wind speed variation on rainfed wheat production evaluated by the CERES-Wheat model. *International Journal of Biometeorology* **66**, 225–233 (2022). https://doi.org/10.1007/s00484-021-02209-7

53 Ma, J. *et al.* Winter Wheat Yield Prediction and Influencing Factors Analysis Based on FourierGNN–Random Forest Combined Modeling. *Agronomy* **15**, 641 (2025). https://doi.org/10.3390/agronomy15030641

54 Suliman, M., Scaini, A., Manzoni, S. & Vico, G. Soil properties modulate actual evapotranspiration and precipitation impacts on crop yields in the USA. *Science of the Total Environment* **949**, 175172 (2024). https://doi.org/10.1016/j.scitotenv.2024.175172

55 Somvanshi, S. S. & Kumari, M. Comparative analysis of different vegetation indices with respect to atmospheric particulate pollution using sentinel data. *Applied Computing and Geosciences* **7**, 100032 (2020). https://doi.org/10.1016/j.acags.2020.100032

56 Fu, H. *et al.* Winter Wheat Yield Prediction Using Satellite Remote Sensing Data and Deep Learning Models. *Agronomy* **15**, 205 (2025). https://doi.org/10.3390/agronomy15010205

57 Huang, J. *et al.* Evaluation of regional estimates of winter wheat yield by assimilating three remotely sensed reflectance datasets into the coupled WOFOST–PROSAIL model. *European Journal of Agronomy* **102**, 1–13 (2019). https://doi.org/10.1016/j.eja.2018.10.008

58 Cao, J., Wang, H., Li, J., Tian, Q. & Niyogi, D. Improving the Forecasting of Winter Wheat Yields in Northern China with Machine Learning–Dynamical Hybrid Subseasonal-to-Seasonal Ensemble Prediction. *Remote Sensing* **14**, 1707 (2022). https://doi.org/10.3390/rs14071707

59 Cheng, E. *et al.* Wheat yield estimation using remote sensing data based on machine learning approaches. *Frontiers in Plant Science* **Volume 13 - 2022** (2022). https://doi.org/10.3389/fpls.2022.1090970

60 Chen, X. *et al.* Prediction of maize yield at the city level in China using multi-source data. *Remote Sensing* **13**, 146 (2021). https://doi.org/10.3390/rs13010146