Machine learning-based winter wheat yield prediction using multisource data

Khosravani Shariati, Seyed Arash; Abbasi, Ali

DOI
10.1016/j.agwat.2025.109951

Publication date
2025

Document Version
Final published version

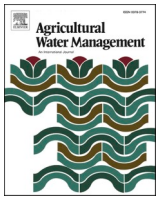
Published in
Agricultural Water Management

Agricultural Water Management 322 (2025) 109951

Contents lists available at ScienceDirect

Agricultural Water Management

journal homepage: www.elsevier.com/locate/agwat



Machine learning-based winter wheat yield prediction using multisource data

Seyed Arash Khosravani Shariati [a], Ali Abbasi [a,b,*]

[a] Department of Civil Engineering, Faculty of Engineering, Ferdowsi University of Mashhad, Mashhad 9177948974, Iran
[b] Faculty of Civil Engineering and Geosciences, Water Resources Section, Delft University of Technology, Delft 2628 CN, the Netherlands

ARTICLE INFO

Handling Editor - Ranvir Singh

Keywords:
Crop yield prediction
Evapotranspiration
Feature selection
Machine learning
XGBoost

ABSTRACT

Accurate crop yield prediction and understanding its underlying factors facilitate better food supply management and more informed decision-making. To forecast crop yield, the majority of previous studies have utilized vegetation indices and meteorological data. However, other important factors are often overlooked. Moreover, the temporal influence of input variables has been underexplored in prior research. To fill these gaps, we integrated a diverse range of satellite-based data, including vegetation indices and actual evapotranspiration (ET_a), with climate and soil information. Then, the input variables were narrowed down using a feature selection approach to provide the most relevant variables for predictive models. Three machine learning algorithms, Extreme Gradient Boosting (XGBoost), Random Forest (RF), and Linear Regression (LR), were trained to forecast winter wheat yield across Oklahoma and Kansas counties. The models were trained on 2014–2021 data and tested on 2022–2023 yields. According to the results, XGBoost emerged as the most accurate algorithm in both test years. It achieved an R^2 of 0.71 (RMSE = 0.46 t ha^{-1}) in 2022 and an R^2 of 0.63 (RMSE = 0.60 t ha^{-1}) in 2023 when using the selected feature set. For most models, particularly in 2022, using the selected features instead of the entire set improved the accuracy. We also found that ET_a is a promising factor in yield prediction, as it was selected multiple times across the growing season in the feature selection process. Additionally, correlation analysis showed that April and May, which are two to three months before harvest, were the most sensitive months in shaping the final yield.

1. Introduction

Food insecurity is a global problem that needs immediate attention and intervention. The severity of this issue is underscored by the fact that nearly 733 million people suffered from hunger in 2023. This large number represents one in eleven individuals across the globe and one in five in Africa (FAO, 2024). Climate change is making this situation even worse, especially since extreme events like droughts are happening more often. Such unfavorable conditions may reduce agricultural productivity or even destroy crops (Bokusheva et al., 2016). Meanwhile, the world's population has more than tripled since the middle of the 20th century, and it is predicted to increase by almost 2 billion people within the next 30 years (United Nations, 2025). Therefore, providing adequate food for this increasing population will be challenging and requires a high degree of agricultural planning and accurate production estimation.

Crop yield is critical for the development of rural areas and a key metric for assessing food security (Li et al., 2007). Accurate yield predictions aid authorities in making strategic decisions, managing storage, allocating resources, and setting prices (Balaghi et al., 2008; Chen et al., 2019). Moreover, it can help farmers adjust their agronomic management before harvest to meet crop demands by applying adequate irrigation and fertilizers.

Wheat is the most widely grown food crop in the world. It provides approximately 18 % of global dietary calorie intake and 19 % of protein consumption (Reynolds and Braun, 2022). Given wheat's cardinal role within the global food system, accurate wheat yield forecasting has become indispensable to the preservation of nutritional stability. As a result, this topic has drawn substantial attention from scholars and officials over the past decades.

In general, the process of agricultural yield prediction consists of two steps: data acquisition and establishing a predictive model. From the perspective of data sources, remote sensing images from space are

* Correspondence to: Department of Civil Engineering, Faculty of Engineering, Ferdowsi University of Mashhad, Azadi square, Mashhad, Razavi Khorasan Province 9177948974, Iran.
E-mail addresses: ar.khosravanishariati@alumni.um.ac.ir (S.A. Khosravani Shariati), aabbasi@um.ac.ir, a.abbasi@tudelft.nl (A. Abbasi).

https://doi.org/10.1016/j.agwat.2025.109951
Received 17 August 2025; Received in revised form 12 October 2025; Accepted 30 October 2025
Available online 6 November 2025

becoming a popular source of information for yield and biomass estimation due to their availability, temporal, and spatial resolution. Satellite imagery spectral bands used for agricultural monitoring include visible, thermal, and microwave spectra. Vegetation indices (VIs) such as the normalized difference vegetation index (NDVI) are extensively used to investigate crop conditions and yield. For example, Petersen (2018) used monthly anomalies of NDVI, enhanced vegetation index (EVI), and normalized difference water index (NDWI) to predict corn, soybean, and sorghum yields in Illinois and successfully applied this method to each country in Africa. In addition to VIs, land surface temperature (LST) has also been widely used as an indicator to assess canopy temperature and crop stress. Pede et al. (2019) computed the killing degree day (KDD) using LST images and compared it to KDD derived from air temperature (Tair) to predict corn yield. They found that the LST-based model outperformed the Tair model.

In addition to remote sensing data, another frequently used data source for yield prediction is climate variables. Many studies have investigated the impact of variables such as precipitation, temperature, wind, and solar radiation, either as original variables or drought indices, on crop yield (Baffour-Ata et al., 2021; Sridhara et al., 2020). For instance, Bazrafshan et al. (2022) employed meteorological and fertilizer data to estimate rainfed wheat yields in Iran using multilayer perceptron (MLP). Some previous studies reported that when different sources (e.g., remote sensing and climate data) are combined, more accurate predictions are made (Cai et al., 2019; Bouras et al., 2021). Wang et al. (2020) combined four different sources, including climate data, satellite images, soil maps, and historical wheat yields across the CONUS. They stated the efficacy of using multi-source data, especially soil maps, to improve yield estimation. However, based on prior studies, little is known about the inclusion of moderate-resolution satellite-based actual evapotranspiration (ET_a) as an input variable in large-scale yield estimation. To the best of our knowledge, only Naghdyzadegan Jahromi et al., (2023) incorporated fine-resolution ET_a from the METRIC model for wheat prediction and reported that ET_a can provide critical and supplementary information. In the past, the science of ET estimation was limited only to field-scale studies. However, due to the advent of satellite imagery and advances in computing power, ET_a modelling and analysis have become possible on a large scale at moderate resolution.

Statistical and process-based models are two primary methods that have been employed to forecast crop yields. Process-based models are designed to mathematically quantify crop growth and yield by considering the interaction between the crop and its environment. These models utilize various data, including weather information, soil and cultivar characteristics, and crop management (Asseng et al., 2014). However, the difficulty of adopting process-based models is that they often require a large amount of input data for model parameterization and calibration, which is unavailable in many regions. In addition, these models are often regarded as "point-based," which can make them unsuitable for application at regional and national scales (Basso et al., 2013). On the other hand, statistical methods establish a relationship between historical crop yield records and various predictive variables to make future predictions. The advantage of these models is that they have limited, if any, reliance on field calibration. Their model error and uncertainty assessments are also more straightforward (Lobell and Burke, 2010).

Establishing a model is the second critical component of prediction. A wide range of machine learning algorithms has been successfully used in yield prediction. They are powerful, easy-to-use, and efficient tools for establishing a function between a target variable and its determinants. For example, Zhang et al. (2017) applied stepwise regression to forecast winter wheat yield in Oklahoma, the U.S., and their model could account for 70 % of the yield variation ($R^2 = 0.7$). Statistical models are only able to specify a linear function between yield and its predictors; however, recent machine learning approaches can even capture nonlinear relationships between variables. Recent studies have been implementing more than one machine learning model to compare their performances and identify the best model. For instance, Tian et al. (2021) reported that the long short-term memory (LSTM) had a higher accuracy than the backpropagation neural network (BPNN) and support vector machine (SVM) in wheat yield estimation in the Guanzhong Plain, PR China. Another study developed a multiple linear regression (MLR) model and three nonlinear machine learning models, including SVM, RF, and XGBoost, to make an early cereal yield forecast in Morocco, and they reported that the XGBoost algorithm outperformed the others (Bouras et al., 2021). However, a superior algorithm has not yet been identified; therefore, further extensive studies are needed to compare the performances of state-of-the-art algorithms for each case study.

In data-driven frameworks, identifying key variables is an important step in establishing a more robust and interpretable model. For this reason, many yield-prediction studies adopted feature selection approaches to find optimal feature sets and reduce computation time (Abdel-salam et al., 2024; Lischeid et al., 2022). Nevertheless, feature selection techniques need to be further applied and verified across different regions and environments to reach a consensus about their efficacy. Beyond the modelling aspect, finding influential factors benefits farmers and decision makers to better monitor and refine crop-management strategies. Many existing studies employed correlation analysis or mutual information to gauge the relative importance of variables (Fu et al., 2025; Wang et al., 2020). However, previous analyses have largely demonstrated the overall significance of variables rather than the within-growing-season variability in their influence. Given that crop growth and yield formation are time-dependent processes, examining temporal feature importance offers valuable insights.

In this study, we integrated satellite-based variables, including NDVI, LST, and ET_a, with climate, soil properties, and soil moisture data to develop models for predicting winter wheat yields at the county level. For this purpose, we adopted various machine learning algorithms such as LR, XGBoost, and RF, and compared their performances. In addition, the temporal importance of each variable has been discussed. The specific aims of this study are to: 1) develop a robust and accurate framework for wheat yield prediction; 2) examine the performances of different machine learning models and identify the most accurate model; 3) evaluate the importance of each variable during the crop growth period; 4) identify the best subset of features and investigate how feature selection affects model performance; 5) examine the adaptability and spatial errors of models. This study introduces two key novelties. We combined remote sensing-based actual evapotranspiration data at the spatial resolution of Landsat imagery with other relevant variables to predict wheat yields. Additionally, we applied a feature selection method, used for the first time in a crop yield forecasting application, to identify the most critical variables and improve the accuracy of models.

2. Materials and methods

2.1. Study area

Kansas and Oklahoma, two dominant wheat-producing states in the U.S., were chosen as the study area for crop yield modeling. The National Agricultural Statistics Service (NASS) reported that in 2023, Kansas and Oklahoma produced about 5.5 and 1.9 million tons of winter wheat, respectively. Thus, they accounted for 22 % of the total U.S. production in that year (USDA NASS, 2024). In terms of administrative divisions, Kansas has 105 counties, whereas Oklahoma contains 77 counties. However, only 161 counties were selected as the focus of this study, 102 from the former and 59 from the latter state, as these selected counties constituted nearly 98 % of this region's total winter wheat production from 2014 to 2023. In this study period, the sample size for county-level yield data is 1268.

To understand the monthly variation in weather conditions, the average temperature and precipitation in different months from 2014 to

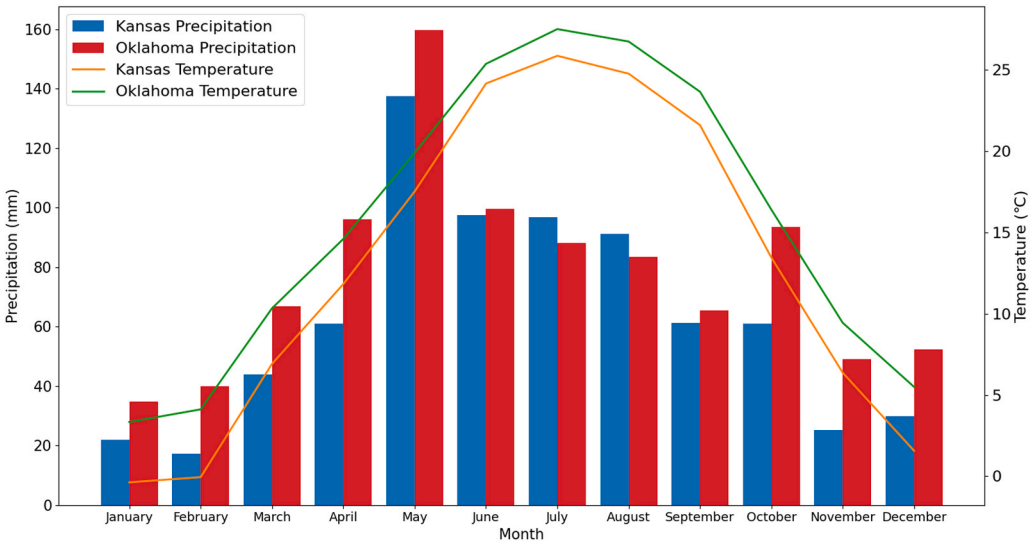


Fig. 1. Average precipitation and temperature of Kansas and Oklahoma from 2014 to 2023.

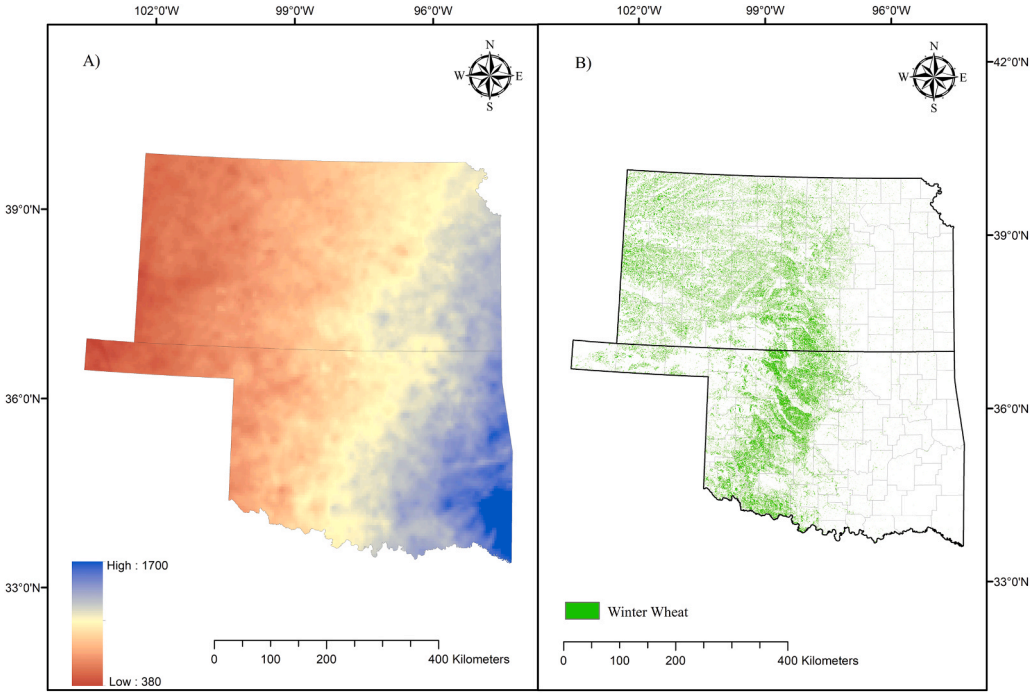


Fig. 2. Kansas and Oklahoma: (a) Average annual precipitation (mm/year) during 2014–2023, and (b) winter wheat distribution in 2023 derived from Cropland Data Layer (CDL).

2023 are shown in Fig. 1. In the study area, monthly precipitation ranged from 17 to 160 cm, peaking in May in both states. From a temperature standpoint, Oklahoma consistently had higher average temperatures than Kansas throughout every month.

The geographical distribution of precipitation clearly shows a west-to-east gradient (Fig. 2(a)). Western regions of both states receive lower precipitation, with annual amounts as low as 380 mm. Moving eastward, precipitation increases steadily and reaches up to 1700 mm in the eastern edge of the study area. This distinct spatial pattern portrays the transition from semi-arid in the west to a more humid condition in the east.

According to NASS, between 2000 and 2019, Kansas's irrigated winter wheat production represented 7.3–12.5 % of the total output. Similarly, in Oklahoma, from 2000 to 2009, it ranged between 3.5 and 9 % of the total production. These numbers show that irrigation has a very small effect on the total output.

In Kansas and Oklahoma, winter wheat is usually planted from mid-September to early October. It emerges in late fall and then undergoes vernalization in the winter. As the soil warms in spring, the wheat resumes growth, continuing until it is harvested between June and July. Planting, growth stages, and harvesting schedules vary from county to county and year to year in these states. However, for consistency in research, the growing season is defined from the previous October to the end of July. Winter wheat is mostly planted in the western two-thirds of these states, as shown in Fig. 2(b).

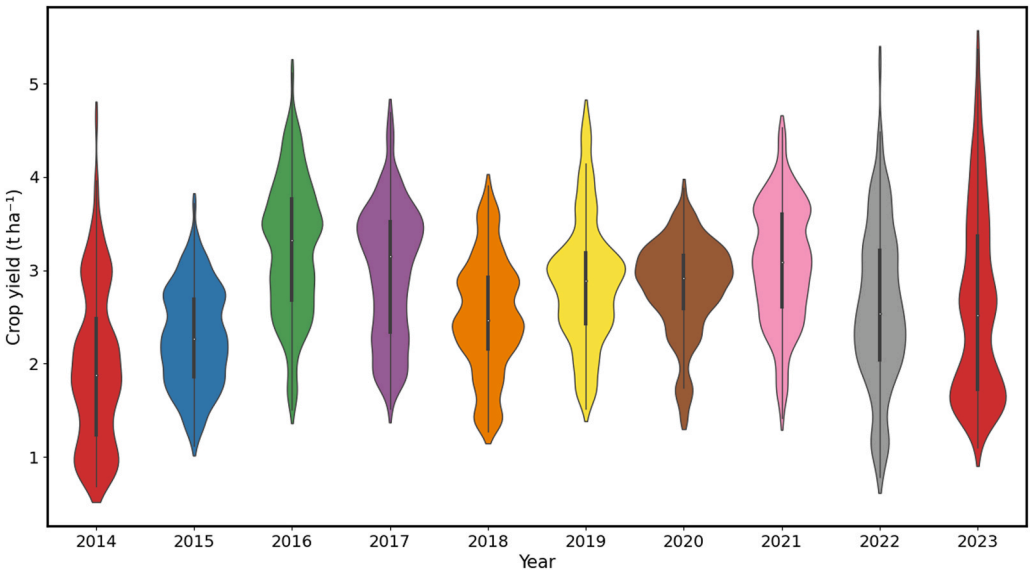


Fig. 3. The distribution of winter wheat yield from 2014 to 2023.

2.2. Data

2.2.1. Crop data

The annual county-level winter wheat yield records were collected from the USDA NASS database (USDA NASS, 2023). NASS provides comprehensive and reliable yield data from the county to the national level through surveys and censuses. Our study period was a 10-year span from 2014 to 2023. During this timeframe, yields widely ranged from approximately 0.7–5.4 t ha^{-1}.

The yearly yield distribution is shown in Fig. 3, indicating a high variability among different years. For example, in 2014, winter wheat yields dropped to exceptionally low levels, the lowest in Kansas since 1995, and in Oklahoma since 1967, mainly because of drought and freeze damage. In contrast, due to favorable weather conditions, Kansas experienced an all-time high yield in 2016, averaging over 3.8 t ha^{-1} across this state. The high temporal and spatial yield variability in these years poses a challenge in yield modeling and prediction.

Some studies have detrended crop yield in order to remove the effect of technological advances on the yield trend and only considered the climate effects. However, because the span of 10 years is not long enough for any major changes in agronomical technology and practices, detrending has not been investigated in this study.

The NASS Cropland Data Layer (CDL) was obtained to distinguish winter wheat planting pixels. CDL is an annual crop-specific classification map of the entire continental United States in raster format. This product, which has been extensively used in agricultural studies, is generated by satellite imagery and ground truth data that are collected during the growing season, and is available at 30-m spatial resolution since 2008 (Boryan et al., 2011). In our study, this crop layer was used to create annual winter wheat masks, which were applied to spatial datasets to isolate winter wheat pixels when averaging variables at the county level.

2.2.2. Climate data

The meteorological data within the growing seasons were retrieved from the Parameter-elevation Regressions on Independent Slopes Model (PRISM) dataset, which is freely available on the Oregon State University PRISM Climate Group website (http://www.prism.oregonstate.edu). This daily gridded dataset is available at 4 km spatial resolution since 1981. The meteorological data extracted from this dataset include minimum temperature, mean temperature, maximum temperature, minimum vapor pressure deficit and maximum vapor pressure deficit (hereafter referred to as T_{min}, T_{mean}, T_{max}, VPD_{min}, and VPD_{max}, respectively). These variables were aggregated to a monthly time scale before being fed to the models. Furthermore, to reflect both the short-term and mid-term effects of precipitation throughout the growing season on wheat yield, the standardized precipitation index (SPI) (McKee et al., 1993) was calculated at three different time scales: the one, two and six-month intervals (SPI1, SPI2 and SPI6). Accordingly, to calculate SPI, a commonly used drought monitoring index, 43 years of monthly precipitation records (1981–2023) were collected from the PRISM dataset and then fitted to a gamma distribution function, which was subsequently transformed into a normal distribution with a mean of zero.

We collected soil moisture data from the Global Land Data Assimilation System Version 2 (GLDAS-2.1), which provides soil moisture data at various depths with a spatial resolution of 0.25 ° × 0.25 ° from 2000 to the present. The GLDAS soil moisture was evaluated in many studies and demonstrated to be in good agreement when compared with in situ measurements (Bi et al., 2016; Yang et al., 2022). Soil moisture has been found to exhibit a high correlation with crop yields in previous studies and serves as an indicator for monitoring agricultural drought. In this study, the root-zone soil moisture was used and aggregated to a monthly scale.

2.2.3. Satellite and modelled data

In this study, satellite images were used to approximate plant biomass and monitor vegetation health. We derived the conventional normalized difference vegetation index (NDVI) from the near-infrared and red bands of the MODIS MCD43A4 daily product. This index is a crop monitoring tool that ranges between −1 and 1. Healthy vegetation reflects near-infrared radiation more than red radiation; therefore, higher NDVI values indicate healthier and greener plants. NDVI proved to be highly efficient in estimating vegetation density and crop yields in many studies (Becker-Reshef et al., 2010; Schwalbert et al., 2020). From the same MCD43A4 dataset, we also extracted the enhanced vegetation index (EVI), which has higher sensitivity in areas with dense vegetation, and the normalized difference water index (NDWI) as a metric of vegetation water status. Along with these indices, leaf area index (LAI), a measure of green leaf area per unit ground surface, was obtained from MCD15A3H.

Additionally, daytime land surface temperature (LST) has been retrieved from the MODIS MOD11A1, which is a daily product with a 1 km spatial resolution. LST can provide information on canopy

Table 1
A summary of input variables.

Category	Variable name	Data source	Spatial resolution	Temporal resolution
Crop data	Winter Wheat yield	USDA NASS (USDA NASS, 2023)	County-level	Yearly
	Crop map	USDA NASS (Boryan et al., 2011)	30 m	Yearly
Climate	Tmin, Tmean, Tmax	Prism (Daly et al., 2008)	4 km	Daily
	VPDmin, VPDmax			Daily
	Precipitation (SPI1, SP2, SPI6)			Monthly
	Root zone Soil Moisture (SM)	GLDAS (Beaudoing and Rodell, 2020)	0.25°	3 h
Satellite-based	NDVI EVI NDWI	MCD43A4 (Schaaf and Wang, 2021)	500 m	Daily
	LST	MOD11A1 (Wan et al., 2021)	1 km	Daily
	LAI	MCD15A3H (Myneni et al., 2021)	500 m	4-day
	ET_a (SSEBop)	OpenET (Melton et al., 2022)	30 m	Monthly
Soil properties	Clay Content (CC)	OpenLand Map (Hengl, 2018a)	250 m	Static
	Sand Content (SC)	OpenLand Map (Hengl, 2018b)		
	Soil Organic Carbon Content (SOC)	OpenLand Map (Hengl and Wheeler, 2018)		
	PH	OpenLand Map (Hengl, 2018c)		

temperature and surface-energy balance for assessing crop water stress. For all these MODIS products, a quality filter was applied using the QA or QC bands to ensure that only high-quality pixels were included in the analysis.

We also incorporated actual evapotranspiration (ET_a) as a satellite-based input in our prediction models. Many studies demonstrated the role of ET_a in agricultural water planning and management applications (Ji et al., 2021), yet it has not been thoroughly investigated as an explanatory variable in yield prediction. To fill this gap, we chose the Operational Simplified Surface Energy Balance (SSEBop) model to obtain gridded ET_a. SSEBop is a simplified thermal-based surface energy balance approach to estimate actual evapotranspiration (ET_a) without solving the full surface energy balance equations. Its primary inputs include satellite-derived land surface temperature, daily maximum air temperature, and reference evapotranspiration. A detailed description of this model can be found in Senay et al. (2022), (2013).

The ET_a data for this study were obtained from the OpenET monthly product, accessed via the Google Earth Engine platform. OpenET implements satellite-based ET models, including SSEBop, on the GEE cloud computing platform to provide historical and near real-time ET products at fine spatial resolution (30 m). This offers unprecedented and consistent access to 30 m resolution ET_a that is applicable to agricultural and water resources management purposes.

Regarding the OpenET calculation of ET_a, the primary satellite input is Landsat imagery, which typically acquires data every 8 days, depending on cloud cover. On each valid overpass date, OpenET computes the fraction of reference evapotranspiration by dividing the satellite-derived ET by reference ET sourced from gridded weather datasets. These per-pixel ET fractions are then linearly interpolated across the days between clear-sky overpasses. The interpolated fractions are multiplied by daily reference ET values, creating a continuous daily time series of ET_a at 30 m resolution per pixel. Finally, these daily pixel-level ET_a values are aggregated into monthly and annual totals (Melton et al., 2022).

2.2.4. Soil properties

The variation of soil characteristics across fields affects crop growth and development, which ultimately impacts the final yield even under uniform environmental and management conditions (Adhikari et al., 2023). In this study, four soil variables—clay content (CC), sand content (SC), soil organic carbon content (SOC), and pH—were collected from global soil layers in OpenLandMap. OpenLandMap provided 250 m raster maps at different depths, which were built using machine-learning models trained on a global compilation of soil profiles. Wheat roots can grow beyond 1 m deep, with 95 % of them found in the top 104 cm of the soil (Fan et al., 2016). Therefore, we considered these soil properties at six depths (0, 10, 30, 60, 100, and 200 cm) to ensure that both surface and deep soil characteristics are included.

A summary of all input variables is shown in Table 1.

2.3. Wheat yield prediction framework

The overall workflow of this study is shown in Fig. 4. The first step of this study was data acquisition, which is described in detail in Section 2.2. In the next step, all spatial data were filtered through crop masks to identify wheat areas. Then, sequential data were upscaled to a monthly timescale, as the monthly interval was chosen to simplify our data structure and to unify it. These inputs were then spatially aggregated at the county level to ensure consistency with the scale of yield statistics. The Google Earth Engine (GEE) platform was used for data extraction and preprocessing.

Since the growing season lasts ten months, starting from early October to the end of July, each sequential variable consists of 10 temporal records. In this study, the independent variables included 150 sequential features (fifteen variables over ten months) alongside 24 static variables (four soil measurements at six depths). We also added the average wheat yield in each county from 2000 to 2021, which brought the total number of predictive features to 175. These explanatory variables were concatenated with 1269 wheat yield samples to compile the final dataset for model development.

This final dataset was divided into two periods: one from 2014 to 2021 for training and model optimization, and the other from 2022 to 2023 for performance. Meanwhile, the Pearson correlation was used to investigate the relationship between crop yield and predictive variables. Prior to the training and testing stages of machine learning models, we applied a feature selection method to eliminate irrelevant features and identify the effective ones. This step was performed to address the overfitting and multicollinearity issues.

The hyperparameters of three widely known machine learning algorithms—XGBoost, RF, and LR—were tuned using a grid search and five-fold cross-validation on the training dataset.

Following hyperparameter tuning, we conducted two evaluation experiments. In the first one, each model was trained using all explanatory features to establish a baseline. In the second experiment, the models were only fed the selected features from the feature selection process. Next, we used three metrics to assess the accuracy of models in both test years: coefficient of determination (R^2), root mean squared error (RMSE), and mean absolute error (MAE). This two-year evaluation method enabled us to test the robustness and stability of models across two different years and experimental scenarios. Therefore, the performance of the three machine learning models was compared across both test years and the two experiments.

2.4. Machine learning algorithms

We used both linear and nonlinear machine learning algorithms for wheat yield prediction in this study. Linear Regression served as a

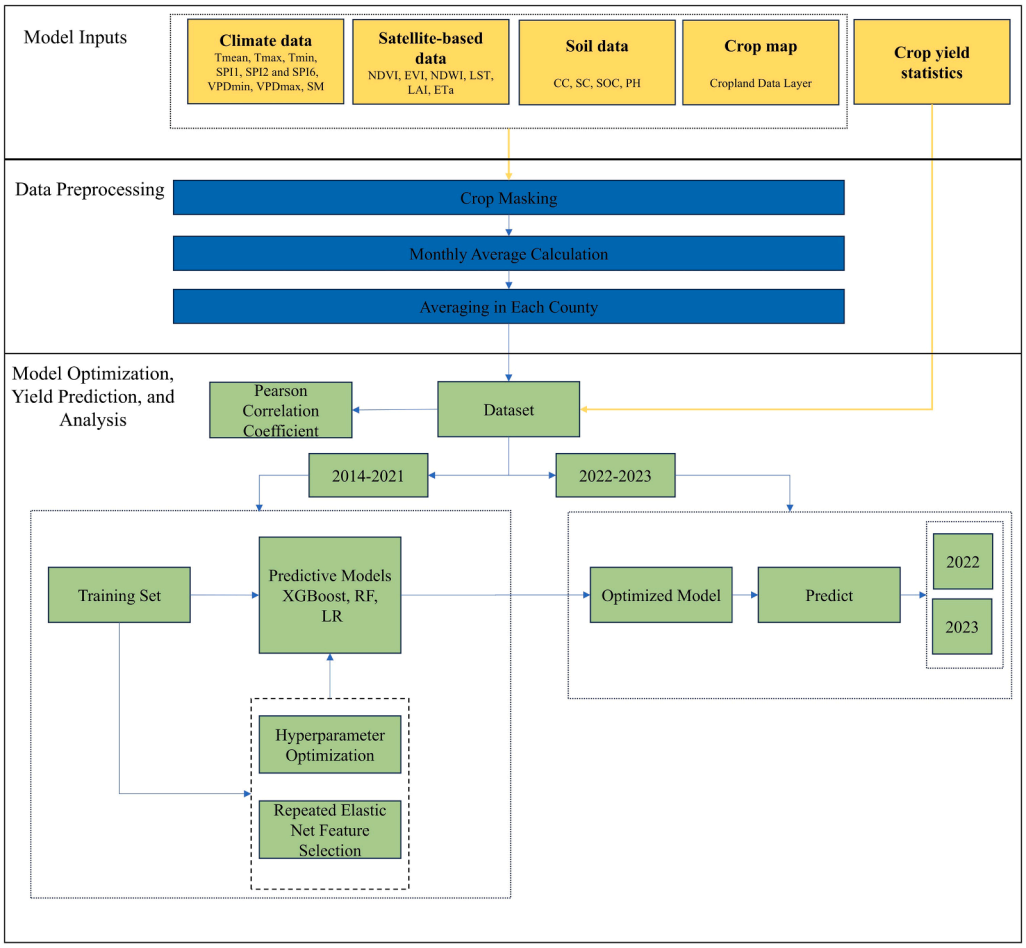


Fig. 4. Flowchart of wheat yield prediction framework.

benchmark linear model, whereas RF and XGBoost, two tree-based ensemble algorithms, were applied to model nonlinear relationships in the data.

2.4.1. Random forest (RF)

RF is an ensemble machine learning technique that combines multiple decision tree predictors. It can be used for both classification and regression tasks (Breiman, 2001). In the RF regression, each tree is constructed by picking a random subset of variables and samples from the dataset. Subsequently, each decision tree, representing a sub-regression model, develops a regression model and makes a prediction. Lastly, the final prediction is made by averaging all of the predictions from the individual decision trees.

In our study, we tuned four hyperparameters in the RF algorithm, which were the number of trees in the forest, the maximum depth of each tree, the minimum number of samples required at a leaf node, and the minimum number of samples required to split a node. These parameters were set to 200, 14, 2, and 2, respectively.

2.4.2. Extreme gradient boosting (XGBoost)

XGBoost is a decision-tree ensemble algorithm that was developed based on the gradient boosting framework. In XGBoost, decision trees are trained at each step to reduce the errors of the previous ensemble. Additionally, the objective function includes a regularization component to manage model complexity and overfitting. This algorithm was introduced by Tianqi Chen at the University of Washington (Chen and Guestrin, 2016). It is necessary to adjust the hyperparameters of this algorithm in order to build an optimal XGBoost model. We set up the XGBoost model for this study with a learning rate of 0.1, maximum tree depth of 4, and 200 boosting rounds (n_estimators = 200) to optimize performance while maintaining generalizability.

2.5. Feature importance

2.5.1. Pearson correlation

The Pearson correlation coefficient quantifies the linear association between two variables. It ranges from -1 to –1. When it equals 1, one variable changes in exact proportion to the other, whereas a value of -1 shows that they vary proportionally but in opposite directions. If it is zero, there is no linear correlation. The formula for the Pearson correlation coefficient between two variables, X and Y, denoted as $\rho_{X,Y}$, is:

$$\rho_{X,Y} = \frac{\mathrm{cov}(X, Y)}{\sigma_X \sigma_Y} \tag{1}$$

where *cov (X, Y)* signifies the covariance between *X* and *Y*, and σ_X and σ_Y are their respective standard deviations.

2.6. Feature selection

In this study, the Repeated Elastic Net Technique (RENT) was employed to find the best subset of input variables. This approach is introduced by Jenul et al. (2021). Contrary to the majority of feature techniques that only focus on optimizing model performance, this approach also considers the stability and robustness of the feature selection process. This technique begins by randomly sampling *K* independent subsets from the training dataset. Then, a linear model with elastic net regularization is fitted to each subset to obtain feature

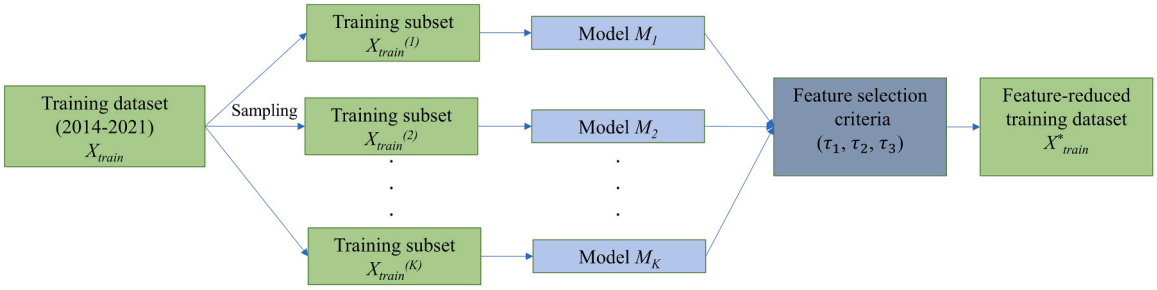


Fig. 5. RENT feature selection pipeline adopted from Jenul et al. (2021).

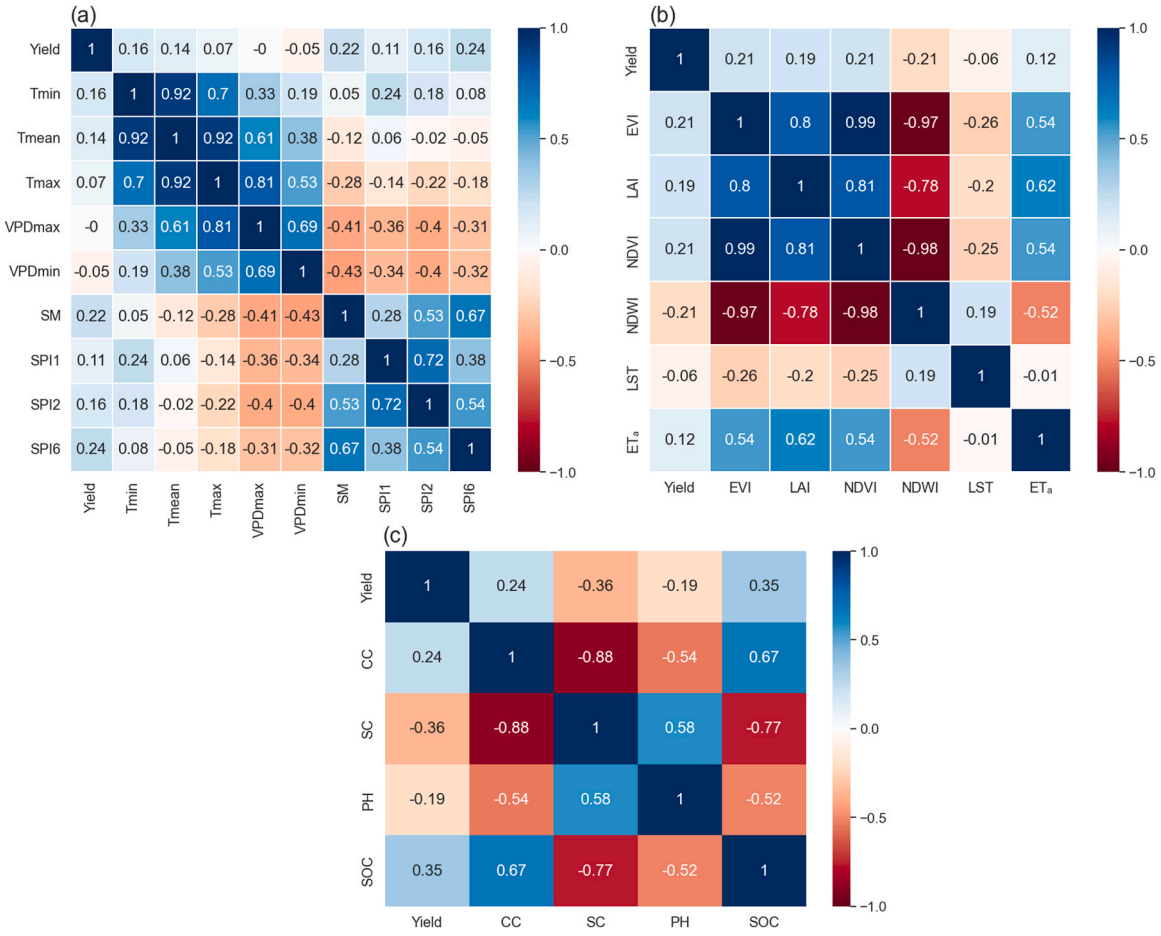


Fig. 6. The correlation heatmap of variables: (a) the average correlation value of climate data during the growing season, (b) the average correlation value of satellite-based data during the growing season, (c) the average correlation value of soil property data across different depths.

weights ($\beta_{k,\,n}$), where k represents the training subset and n the feature. These feature weights are compiled into a matrix and evaluated based on three criteria to ensure the stability and robustness of the final model. The three criteria are as follows:

$$\tau_1(\beta_n) = \frac{1}{K}\sum_{k=1}^{K} 1_{[\beta_{k,n}\neq 0]} \tag{2}$$

$$\tau_2(\beta_n) = \frac{1}{K}\left|\sum_{k=1}^{K} sign(\beta_{k,n})\right| \tag{3}$$

$$\tau_3(\beta_n) = t_{K-1}\left(\frac{|\mu(\beta_n)|}{\sqrt{\frac{\sigma^2(\beta_n)}{K}}}\right) \tag{4}$$

where $\mu(\beta_n)$ and $\sigma^2(\beta_n)$ are the mean and variance of feature weights, and $t_{K-1}(.)$ represents the cumulative distribution function of Student's t-distribution with $K-1$ degrees of freedom.

Eq. (2) measures how often the feature weight ($\beta_{k,\,n}$) is non-zero across K models. Eq. (3) examines the consistency of the sign (positive or negative) of feature weights across models. The last one evaluates, using Student's t-test, whether the mean coefficient for a feature across models is significantly different from zero. In the feature process, a feature is kept if it satisfies all three criteria $\tau_1(\beta_n)$, $\tau_2(\beta_n)$, and $\tau_3(\beta_n)$ with respect to their corresponding user-defined cutoff values t_1, t_2, *and* t_3. The overall RENT pipeline in shown in Fig. 5. The full details of this approach can be found in Jenul et al. (2021).

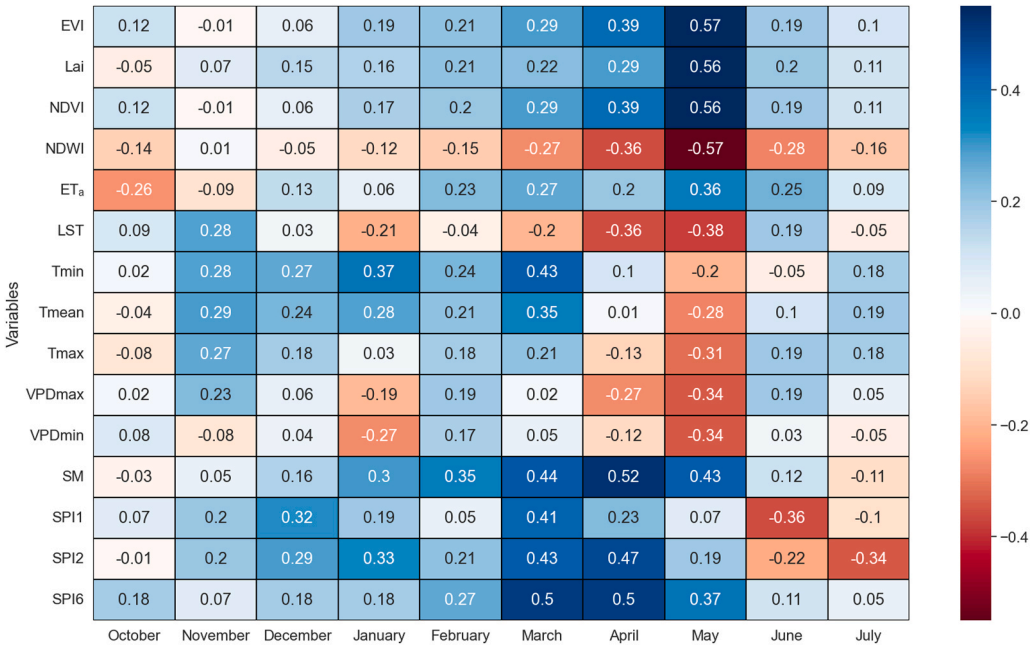


Fig. 7. The correlation coefficients between monthly variables and wheat yield throughout the growing season.

3. Results

3.1. Correlation analysis

The Pearson correlation coefficients between input variables and yield are shown in Fig. 6. They are divided into three groups according to the data sources: climate, satellite-based, and soil properties features. All variables (except SPI and soil properties) were standardized on a county basis to reduce the regional effects of soil, different growing conditions and agronomic practices prior to correlation analysis.

In the climate group, SPI indices and soil moisture showed consistent positive relationships with yield and with each other. On the other hand, temperature-related variables (T_{min}, T_{mean} and T_{max}) and vapor pressure deficits (VPD_{min} and VPD_{max}) were strongly intercorrelated, but their correlations with yield were weak or near zero. These weak correlations suggest that averaging correlations over the growing season may hide the actual effects of temperature and vapor pressure deficit on yield. Thus, we need a more in-depth temporal analysis, which is presented in Fig. 7.

In the second group of Fig. 6, the vegetation indices EVI, LAI, and NDVI were strongly and positively linked to each other and to actual evapotranspiration (ET_a). This means that they are similarly sensitive to crop conditions. NDWI, however, displayed strong but negative correlations with the vegetation indices. This opposite pattern is due to how NDWI is calculated; unlike the other indices, it incorporates a negative element of near-infrared (NIR) reflectance. Additionally, LST showed weak correlations with the other remote sensing variables in this category, implying that it provides unique thermal-related information.

Within the soil-related group, correlations were averaged at different depths. Variables displayed a mix of relationships with yield. The strongest associations with yield were found for SOC ($r = 0.35$) and SC ($r = –0.36$), while CC and pH demonstrated weaker relationships.

In the next step, to examine the temporal dynamics throughout the growing season, Fig. 7 presents the monthly correlation coefficients between monthly features and yield. This figure illustrates that each variable has distinct periods of influence on yield. For instance, vegetation indices like EVI, LAI and NDVI had strong positive correlations with yield from March to May, with the highest value in May. NDWI showed a similar trend, but the correlation values were negative. ET_a was negatively related to yield in October, but it changed to a positive relationship during the rest of the growing season.

The temperature variables (T_{min}, T_{max}, and T_{mean}) showed varied correlations with yield during the growing season. From November to March, they exhibited moderate positive correlations, but this turned negative in April and May, which is during the reproductive stage and a time when heat stress is critical. LST had a moderate negative correlation from January to May, peaking in April and May. This confirms the negative consequences of heat stress during this period. Water-related variables, such as soil moisture, SPI1, SPI2 and SPI6, showed positive correlations from winter through the spring, with the highest correlation of SM, SPI2 and SPI6 occurring in April. It's interesting to note that in the last two months of the growing season (June and July), we found a negative correlation between SPI1, SPI2, and yield. This is probably due to the possible negative effects of heavy rainfall during this time.

3.2. Feature selection

In this study, we randomly sampled 100 subsets from the training dataset, each having 70 percent of the total training size. In order to ensure reproducibility, different random seeds ranging from 0 to 99 were used during the sampling process. Subsequently, a grid search was performed to find the best combination of elastic net hyperparameters and cutoff values. In our implementation of RENT, we adapted the original two-stage grid search procedure described by Jenul et al. (2021) into a single step for simplicity and visualization purposes. Additionally, elastic net hyperparameter names were changed to match scikit-learn nomenclature (Pedregosa et al., 2011).

The grid search was carried out over the following parameter space: for the elastic net hyperparameters, α ranged from 10^{-4} to 10 in six logarithmic steps, and $\ell_1 ratio \in [0,1]$ with a step size of 0.1. For the stability cutoff values, t_1 and t_2 were bounded by [0.5,1] with the same step size as $\ell_1 ratio$, while t_3 was selected from {0.95,0.975,0.99} as these values correspond to different significance levels in the *t*-test. For each hyperparameter combination, elastic net models were fitted to all subsets to obtain feature weights, and features meeting the three stability criteria were kept. A linear regression model was then trained on the training dataset, using the retained features, and evaluated via Bayesian information criterion (BIC). The optimal configuration, along with its corresponding selected features, was determined based on the lowest BIC observed among all combinations.

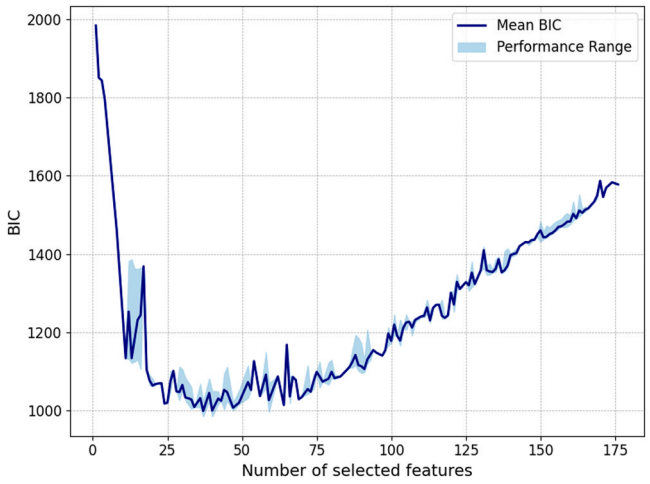


Fig. 8. The optimal number of selected features according to BIC when tuning Elastic Net hyperparameters and cutoff values.

As depicted in Fig. 8, the minimum BIC value occurs when there are 40 features. In the minimum BIC, the values of α, $\mathcal{L}_1 ratio$, t_1, t_2 and t_3 are 0.01, 0.4, 0.9, 0.7 and 0.975, respectively. In this figure, for some values of the number of features, the figure includes more than one BIC performance value. This is because, when tuning hyperparameters and cutoff values, different models sometimes chose different sets of features, even though the total number of selected features was the same.

In Table 2, the selected features are presented. These features are divided into two categories: monthly and static features. Examining the monthly features, early in the growing season (October to December), variables related to water and temperature, such as SM, NDWI, LST, and VPD, stand out as the key features. This shows their important role in crop establishment. In spring (February to April), we observed a phenological change as EVI joined the influential variables, while water-related variables such as SM, SPI, and VPD remain important in affecting yield outcomes. In particular, SPI6 in March tracked cumulative precipitation from growing season onset up to March. In the late growing season (May-July), which corresponds to the heading to maturity stages, LAI emerges as a key feature in May and June. Additionally, satellite-derived ET_a is selected for all late-season months. The inclusion of these features, alongside SM, temperature, SPI, and NDWI, points to the crops' sensitivity to water stress and heat during grain development.

Regarding static features, the average historical yields, clay content, pH, and soil organic carbon content are selected to represent the county's yield level and the soil's physical structure and properties. These static features describe location-specific characteristics, while monthly features display environmental and vegetative conditions over the growing season.

Collectively, this pattern of selected features highlights the need for integrating multi-source and multi-temporal data to capture specific factors influencing yield outcomes.

3.3. Winter wheat prediction results

The performance results of three machine learning algorithms for winter wheat prediction are reported in Table 3. The evaluation of results is conducted under two scenarios: using all available features and feature-selected inputs. These results are categorized separately based on our test years, 2022 and 2023. Across all models and feature sets, the R^2 values ranged from −0.70–0.71, the RMSE from 0.46 to 1.11 t ha^{-1}, and the MAE from 0.37 to 0.94 t ha^{-1}.

Feature selection had a clear positive impact, particularly in 2022, with models using the subset of features mostly leading to better or at least comparable results. In 2022, for example, with the selected features, XGBoost's R^2 improved from 0.65 to 0.71 and its RMSE dropped from 0.51 to 0.46 t ha^{-1}. This demonstrates how well the feature selection technique works in reducing the redundancy in the dataset, which raises model accuracy and lowers the risk of overfitting.

Across both years and feature sets, XGBoost outperformed the other algorithms, with the lowest RMSE and MAE, as well as the highest R^2 values. Remarkably, both RF and XGBoost showed minimal changes in performance when using the full set of features or the selected features, which indicated their inherent ability to handle feature redundancy. On the other hand, LR models had the lowest R^2 and the highest RMSE. Thus, we concluded that non-linear models, namely RF and XGBoost, are more accurate than LR in terms of accuracy.

The 2023 prediction results for XGBoost and RF were mostly worse than those of the previous year, which can be explained by the drought across the U.S. winter wheat belt during the 2022–2023 growing season.

Table 3
Performance comparison of LR, RF and XGBoost using both all features and selected features.

		All features			Selected features		
		RMSE (t ha^{-1})	R^2	MAE (t ha^{-1})	RMSE (t ha^{-1})	R^2	MAE (t ha^{-1})
2022	LR	1.11	−0.70	0.94	0.70	0.33	0.59
	RF	0.49	0.67	0.38	0.49	0.68	0.39
	XGBoost	0.51	0.65	0.40	0.46	0.71	0.37
2023	LR	0.73	0.46	0.59	0.69	0.52	0.56
	RF	0.67	0.55	0.53	0.68	0.53	0.54
	XGBoost	0.59	0.65	0.47	0.60	0.63	0.48

Table 2
Selected variables (Numbers in parentheses indicate the soil depth in centimeters for the corresponding variables e.g., PH(0) denotes pH at 0 cm depth).

	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun	Jul
EVI						✓				
LAI								✓	✓	
NDVI										
NDWI	✓							✓		
ET_a				✓				✓	✓	✓
LST	✓					✓				
T_{min}			✓						✓	✓
T_{mean}										
T_{max}		✓					✓	✓		✓
VPD_{max}		✓	✓					✓		
VPD_{min}		✓	✓		✓	✓				
SM			✓				✓			✓
SPI1			✓	✓	✓				✓	
SPI2									✓	
SPI6						✓				
Static features	Average Yield, CC (30), PH (0), PH (10), SOC (10), SOC (200)									

Table 4
Statistical summary of wheat yield (t ha^{-1}) in the training set (2014–2021) and test sets (2022, 2023).

Dataset	Years	Sample number	Mean (t ha^{-1})	Median (t ha^{-1})	25th Percentile (t ha^{-1})	75th Percentile (t ha^{-1})
Training set	2014–2021	1018	2.72	2.78	2.21	3.25
Test set 1	2022	137	2.60	2.54	2.04	3.21
Test set 2	2023	113	2.61	2.52	1.72	3.37

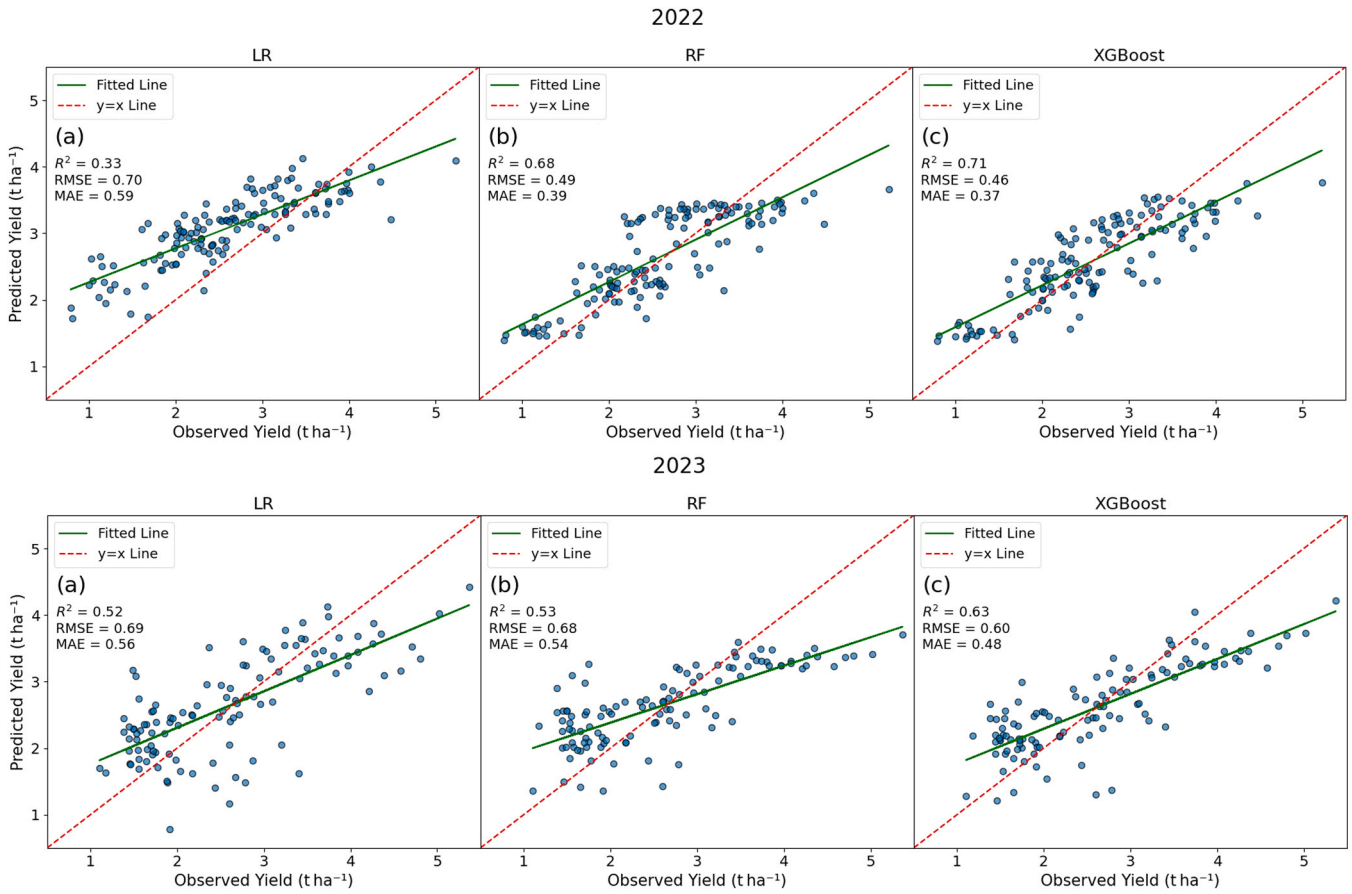


Fig. 9. Comparison of observed and predicted winter wheat yields using selected features for three machine learning models in 2022 and 2023: (a) LR, (b) RF, and (c) XGBoost.

Kansas was the most affected state and experienced the most severe precipitation shortfall since 1896 (Zhang et al., 2024). As a result, Kansas's winter wheat production fell to 5.5 million tons in 2023, the lowest since 1966. These extreme weather conditions altered the data patterns and broadened the range of yield values that were not seen in the training set. As shown in Table 4, the 2023 test has a considerably larger interquartile range (1.72–3.37 t ha^{-1}), although its mean and median are similar to the other sets. This higher variability is likely to have affected the results, as algorithms perform best when training and test data distributions are similar and there is no need to extrapolate beyond the range of training data.

Fig. 9 provides a visual comparison of observed and predicted winter wheat yields of the models that used only selected factors for the test years (2022 and 2023). In each subplot, a red y = x reference line was drawn, which represents perfect predictions, alongside a green fitted regression line that illustrates the general trend of the predictions. The model performs at its best when points closely cluster around the $y = x$ line.

Fig. 9 shows that the predicted values for 2022 were mostly close to the ideal line $y = x$. However, in 2023, the points were further away from this line. Even though the performance of models varied, XGBoost always did better than the other models and had the highest accuracy in both years.

The plots in Fig. 9 illustrate that the models tend to overestimate low yields and underestimate high ones. This pattern is likely due to the limited number of extreme wheat yield samples in the training set. This systematic bias suggests a potential need for further data balancing or model adjustments to improve performance at the tails of the yield distribution.

The comparison of R^2 and RMSE for models trained with the selected features, during the whole study period is shown in Fig. 10. The performance of all models declined considerably from training to test years, as R^2 values decreased and RMSE rose accordingly. LR had the lowest accuracy even when it was tested against the training years, while RF and XGBoost maintained high and stable performance during the training. XGBoost had the highest R^2 and the lowest RMSE in both training and test years.

3.4. Spatial distribution of errors in winter wheat forecasts

Fig. 11 displays the observed winter wheat yields and prediction errors from three models (LR, XGB, RF) for 2022 and 2023, using

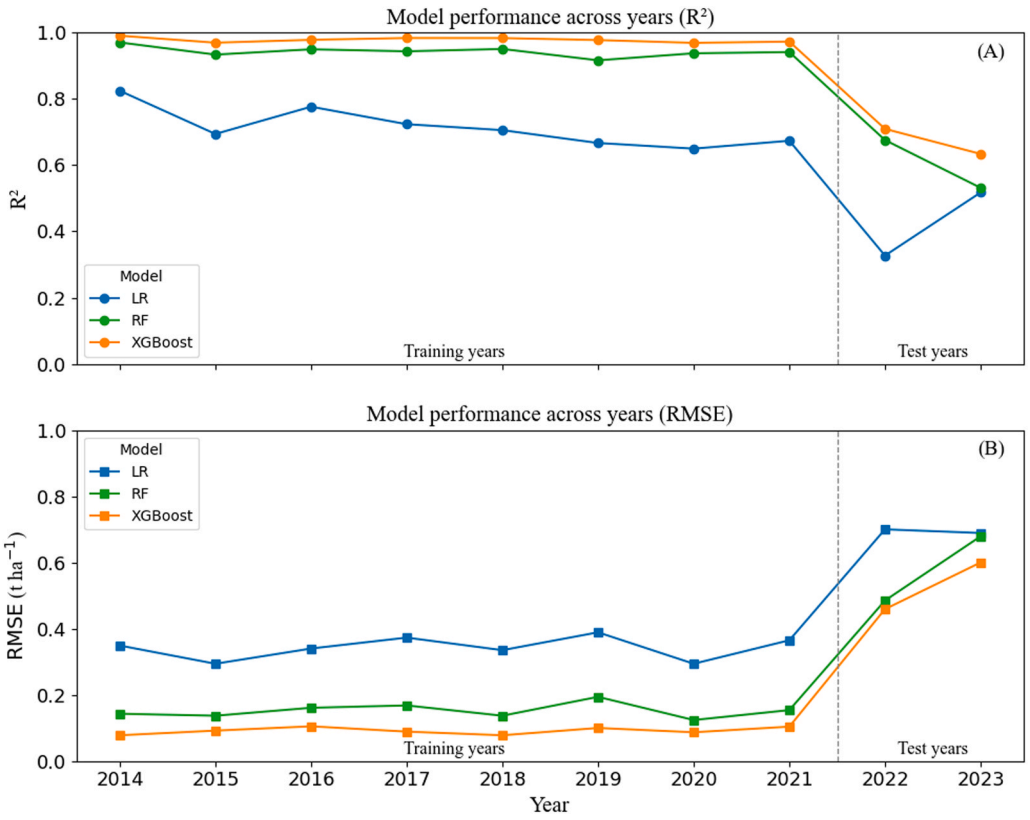


Fig. 10. Performance of the models from 2014 to 2023 with selected features as inputs: (A) R^2 comparison across models; (B) RMSE comparison across models.

selected factors as inputs. The observed yield maps (Fig. 11 (a) and (e)) show a lack of consistent spatial patterns, although a slight west-to-east yield increase is noticeable, particularly in 2023. The wheat yield varied greatly, ranging from less than 1 t ha^{-1} to about 4 t ha^{-1}.

Regarding the error values, positive values (red tones) indicated an overestimation of yield, while negative values (blue tones) reflect underestimation by the models. Generally, lighter colors on the error maps mean that the predictions are more accurate. As shown by the lighter color tones, XGBoost models had lower error magnitudes in both years (Fig. 9(d) and (h)), suggesting superior predictive performance compared to the LR and RF models.

A notable observation is that yields were underestimated in the central counties, while they were overestimated in the northeastern area in 2022 and the eastern region in 2023. In Fig. 11, these geographically clustered errors are marked by black ellipses. Such homogeneous error patterns suggest that the models have difficulty accounting for the spatial variability of yield.

4. Discussion

We adopted multiple sources of data, including VIs, ET_a, climate and soil data, to predict winter wheat yield at the county scale. Correlation analysis in conjunction with a feature-selection method was employed to identify key features and investigate the temporal significance of time-series variables. Three machine learning models were implemented to perform wheat yield prediction.

4.1. Input variables analysis

We inspected the interaction of input variables with wheat yield using two methods, correlation analysis and feature selection. Regarding feature importance, most studies have focused on the overall relationship between time-series variables and yield. However, our study presents the temporal correlation of individual time-series variables, providing deeper insights into agricultural data analysis. The correlation analysis revealed that vegetation indices (VIs) are most significant after the green-up stage (February to March) and peak during the heading to early grain-filling stages (May). In contrast, weaker correlations are observed in the early growing season and harvest periods. These results are consistent with the findings of prior studies (Joshi et al., 2023; Panek et al., 2020). Similarly, Cai et al. (2019) showed that the contribution of satellite data saturates at the peak of the growing season. Moreover, Actual evapotranspiration showed a moderate positive correlation from mid-growing season to the maturity stage, but demonstrated a negative correlation in October. This negative relationship can be attributed to the fact that, in the early stage of wheat growth, evapotranspiration mainly consists of soil evaporation rather than plant transpiration. As a result, if ET_a is excessively high, soil moisture may be depleted, potentially causing water stress that adversely affects germination and root establishment.

Regarding climate and soil data, soil moisture consistently showed a positive correlation throughout the growing season to the end of the flowering stage. SPI1 also had a similar trend, but, interestingly, showed a moderate negative relationship with yield in June. This finding aligns with Joshi et al. (2023), who reported that heavy rainfall just before the harvest may cause wheat kernels to lose weight, ultimately leading to yield loss. Temperature variables mostly had a positive linkage to wheat yield before April, whereas a negative correlation was found during the heading to grain filling stages. A similar result was reported by Jarlan et al. (2014), who found a positive correlation in the early stage and a negative one in the grain-filling stage of wheat growth in Morocco. Generally, most variables reached their peak correlations with yield during April and May. This critical period spans from jointing to grain-filling stages. Identifying such a yield-sensitive window prior to harvest could be valuable for both decision-makers and farmers. Monitoring this period not only helps authorities strategically plan for supply and demand but also enables farmers to adapt their within-season management practices, such as modifications to fertilization

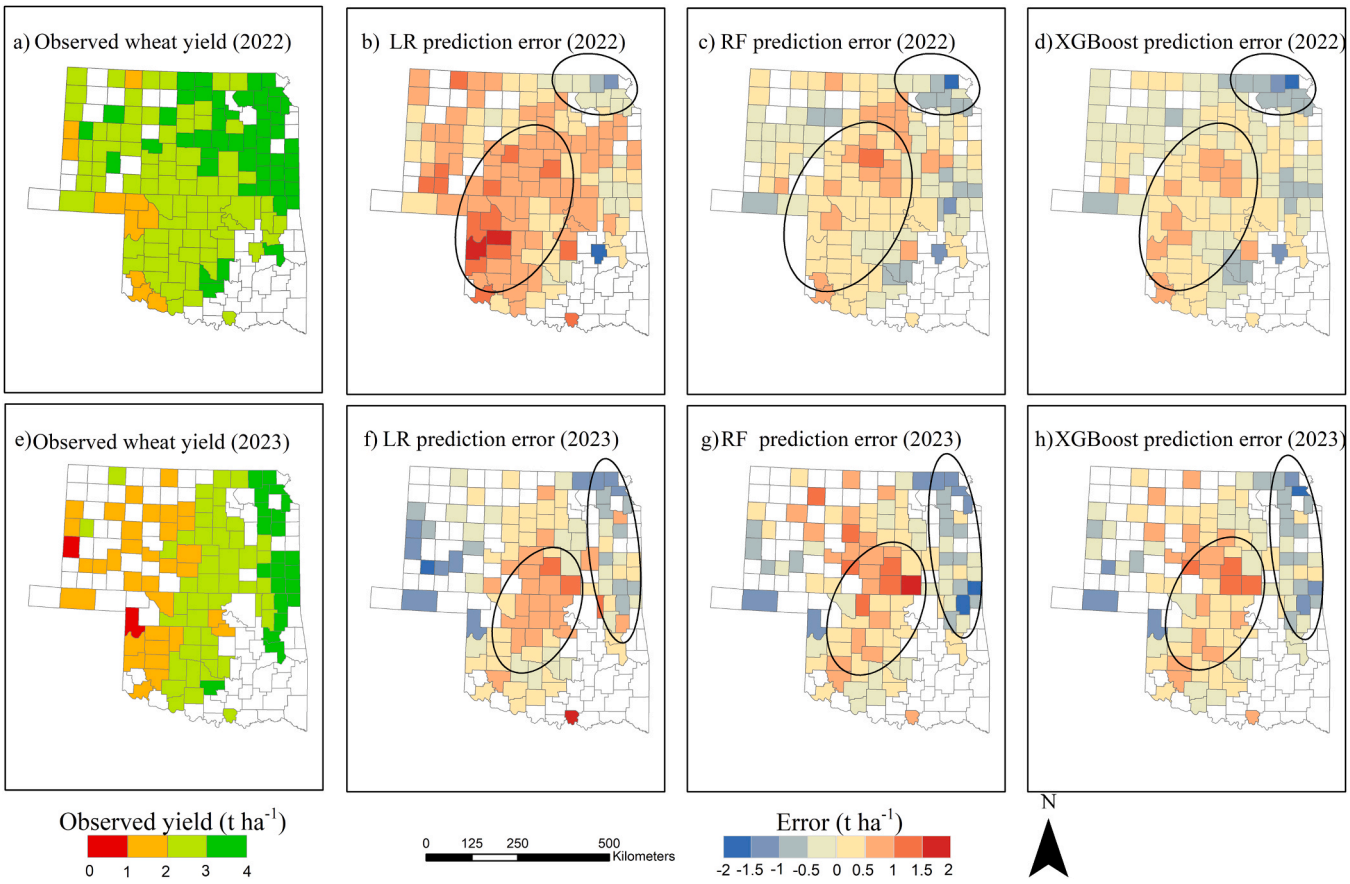


Fig. 11. Spatial representation of prediction errors for 2022 and 2023: (a, e) Observed yield; (b, f) LR prediction errors; (c, g) RF prediction errors; (d, h) XGBoost prediction errors. Panels in the top row are for 2022, and those in the bottom row are for 2023.

or irrigation. In future research, incorporating data with higher temporal and spatial resolution for this period may lead to more accurate predictions of final yield.

The feature selection method used in this study considers both the stability of the selected features and model performance to select impactful features. This approach distinguishes our study from most previous works, as maximizing model performance has mainly been the only criterion in most previous studies. While correlation analysis just informs us about how features are related to each other, an efficient feature selection technique can hand-pick independent features with the strongest influence on the target variable.

Our selected features mostly align with agronomic knowledge. Vegetation indices such as EVI, NDWI, and LAI were chosen from mid to late growing season, as these features in this period, especially during heading to grain filling, proved to be indicators of biomass and canopy health in numerous studies. Among selected variables, ET_a is chosen in four months, January and May to July, demonstrating its significance in yield prediction. This is because actual evapotranspiration is a reflection of various factors, including temperature, precipitation, solar radiation, etc. Likewise, soil moisture and different SPI timescales emerged several times as selected variables, together covering almost the entire growing season. This illustrates the importance of water availability for crop yield.

4.2. Performance

In our framework, we combined various datasets with different spatial and temporal resolutions and unified their resolutions to predict winter wheat yield at the county scale. These data were filtered out using a feature selection method before being fed to the model. Then, the machine learning models were trained from 2014 to 2021 and tested in two separate years, 2022 and 2023. Based on the results, the best model for year 2022 could explain up to 71 % and for year 2023, up to 65 %. This variability in the performance of models across different years can be attributed to different environmental factors, as well as statistical differences in training and test data.

One of the strengths of this study is the feature selection approach that led to a noticeable improvement in most models despite using less than a fourth of the candidate features. The benefits of dimensionality reduction are also discussed in prior studies (Li et al., 2022, 2023). While the improvement in the model's accuracy was subtle for XGBoost and RF, it was substantial for LR. Such a result reflects that XGBoost and RF are less prone to overfitting. Moreover, our results indicated a distinct spatial pattern, with yields being underestimated in several high-producing areas and overestimated in some low-yielding regions, leading to geographically clustered error trends. Therefore, some contiguous counties exhibited similar errors, which are likely due to a lack of extreme yield values in the training data as well as the inability of traditional machine learning models to capture spatial yield variability, as reported in other studies. The underestimation in high-yield regions was also observed by Wang et al. (2020). Addressing this issue could involve incorporating spatial machine learning models, which may help improve the accuracy of predictions across regions with varying conditions.

4.3. Study limitations

There are some limitations in this study that hinder the modelling ability. One possible source of uncertainty in models can be attributed to not including some influential agronomic and environmental factors

such as pests, fertilizers, management practices, irrigation, etc. Further studies can incorporate these variables, not only to improve forecasts but also as a means of better understanding the final yield dynamics. Also, while having a general prediction model can provide broad applicability, the relationships between predictive variables and the target are not spatially uniform across the geographical area in the two states. To address this issue, a spatially-aware machine learning model needs to be incorporated to better account for geographical heterogeneity. The same challenge is also true for the feature selection, as the influential variables may differ across regions. Thus, the feature selection process needs to be tailored to agronomically similar areas.

Finally, while aggregating all datasets to a monthly resolution helped to reduce data dimensionality, this rather coarse temporal scale restricts the capacity to completely capture the dynamics of crop growth. Within days or weeks, the wheat phenological stages can change rapidly and may be exposed to extreme weather or abnormal conditions. Thus, a finer temporal resolution, such as weekly or bi-weekly, would probably be more suitable for tracking these changes and may improve the sensitivity of the model to short-term variations.

5. Conclusion

In this paper, a framework with integrated soil, satellite-based, and climate data was introduced to predict winter wheat yield using machine learning algorithms. One notable finding that emerged is that ET_a proved to add critical information as it demonstrated a moderate positive relationship to crop yield and was selected multiple times across the growing season in the feature selection procedure. Also, the correlation results showed that two to three months before harvest are significantly important in determining the final yield, as most variables had strong correlation with crop yield in April and May compared to other months. Moreover, the feature selection technique implemented in this study effectively identified key predictors while maintaining and, in most cases, enhancing model performance. It also contributed to mitigating overfitting and reducing multicollinearity. Importantly, the month-by-month analysis of variable importance provides decision-makers with more nuanced insights into the temporal dynamics influencing wheat yield. In terms of modelling performance, XGBoost outperformed both LR and RF in both test years. Also, the spatial analysis of models' errors showed significant geographical clustering, especially in low-yield and high-yield regions. This highlights the need for incorporating spatially-aware modeling approaches to further improve yield predictions.

CRediT authorship contribution statement

Seyed Arash Khosravani Shariati: Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Formal analysis, Data curation. **Ali Abbasi:** Writing – review & editing, Validation, Supervision, Project administration, Investigation, Conceptualization.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

All research data is available online and the sources of all used data are cited in the manuscript.

References

Abdel-salam, M., Kumar, N., Mahajan, S., 2024. A proposed framework for crop yield prediction using hybrid feature selection approach and optimized machine learning. Neural Comput. Appl. 36, 20723–20750. https://doi.org/10.1007/s00521-024-10226-x.

Adhikari, K., Smith, D.R., Hajda, C., Kharel, T.P., 2023. Within - field yield stability and gross margin variations across corn fields and implications for precision conservation. Precis Agric. 24, 1401–1416. https://doi.org/10.1007/s11119-023-09995-7.

Asseng, S., Zhu, Y., Basso, B., Wilson, T., Cammarano, D., 2014. Simulation modeling: applications in cropping systems. Encycl. Agric. Food Syst. 5, 102–112. https://doi.org/10.1016/B978-0-444-52512-3.00233-3.

Baffour-Ata, F., Antwi-Agyei, P., Nkiaka, E., Dougill, A.J., Anning, A.K., Kwakye, S.O., 2021. Effect of climate variability on yields of selected staple food crops in northern Ghana. J. Agric. Food Res. 6, 100205. https://doi.org/10.1016/j.jafr.2021.100205.

Balaghi, R., Tychon, B., Eerens, H., Jlibene, M., 2008. Empirical regression models using NDVI, rainfall and temperature data for the early prediction of wheat grain yields in Morocco. Int. J. Appl. Earth Obs. Geoinf. 10, 438–452. https://doi.org/10.1016/j.jag.2006.12.001.

Basso, B., Cammarano, D., Carfagna, E., 2013. Review of crop yield forecasting methods and early warning systems. The First Meeting of the Scientific Advisory Committee of the Global Strategy to Improve Agricultural and Rural Statistics, pp. 1–56.

Bazrafshan, O., Ehteram, M., Moshizi, Z.G., Jamshidi, S., 2022. Evaluation and uncertainty assessment of wheat yield prediction by multilayer perceptron model with bayesian and copula bayesian approaches. Agric. Water Manag 273. https://doi.org/10.1016/j.agwat.2022.107881.

H. Beaudoing and M. Rodell, 2020. GLDAS Noah Land Surface Model L4 3 hourly 0.25 x 0.25 ° V2.1. NASA/GSFC/HSL, Greenbelt, Maryland, USA,Goddard Earth Sciences Data and Information Services Center (GES DISC). 10.5067/E7TYRXPJKWOQ.

Becker-Reshef, I., Vermote, E., Lindeman, M., Justice, C., 2010. A generalized regression-based model for forecasting winter wheat yields in Kansas and Ukraine using MODIS data. Remote Sens. Environ. 114, 1312–1323. https://doi.org/10.1016/j.rse.2010.01.010.

Bi, H., Ma, J., Zheng, W., Zeng, J., 2016. Comparison of soil moisture in GLDAS model simulations and in situ observations over the Tibetan Plateau. J. Geophys. Res. Atmospheres 121, 2658–2678. https://doi.org/10.1002/2015JD024131.

Bokusheva, R., Kogan, F., Vitkovskaya, I., Conradt, S., Batyrbayeva, M., 2016. Satellite-based vegetation health indices as a criteria for insuring against drought-related yield losses. Agric. Meteor. 220, 200–206. https://doi.org/10.1016/j.agrformet.2015.12.066.

Boryan, C., Yang, Z., Mueller, R., Craig, M., 2011. Monitoring US agriculture: the US department of agriculture, national agricultural statistics service, cropland data layer program. Geocarto Int. 26, 341–358. https://doi.org/10.1080/10106049.2011.562309.

Bouras, E.H., Jarlan, L., Er-Raki, S., Balaghi, R., Amazirh, A., Richard, B., Khabba, S., 2021. Cereal yield forecasting with satellite drought-based indices, weather data and regional climate indices using machine learning in morocco. Remote Sens. 13. https://doi.org/10.3390/rs13163101.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32. https://doi.org/10.1023/A:1010933404324.

Cai, Y., Guan, K., Lobell, D., Potgieter, A.B., Wang, S., Peng, J., Xu, T., Asseng, S., Zhang, Y., You, L., Peng, B., 2019. Integrating satellite and climate data to predict wheat yield in Australia using machine learning approaches. Agric. Meteor. 274, 144–159. https://doi.org/10.1016/j.agrformet.2019.03.010.

Chen, T., Guestrin, C., 2016. XGBoost: A scalable tree boosting system. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 13-17-August, pp. 785–794. https://doi.org/10.1145/2939672.2939785.

Chen, K., O'Leary, R.A., Evans, F.H., 2019. A simple and parsimonious generalised additive model for predicting wheat yield in a decision support tool. Agric. Syst. 173, 140–150. https://doi.org/10.1016/j.agsy.2019.02.009.

Daly, C., Halbleib, M., Smith, J.I., Gibson, W.P., Doggett, M.K., Taylor, G.H., Curtis, J., Pasteris, P.P., 2008. Physiographically sensitive mapping of climatological temperature and precipitation across the conterminous United States. Int. J. Climatol. Int. J. Clim. https://doi.org/10.1002/joc.1688.

Fan, J., McConkey, B., Wang, H., Janzen, H., 2016. Root distribution by depth for temperate agricultural crops. Field Crops Res 189, 68–74. https://doi.org/10.1016/j.fcr.2016.02.013.

FAO, 2024. Hunger numbers stubbornly high for three consecutive years as global crises deepen – UN report. ⟨https://www.fao.org/newsroom/detail/hunger-numbers-stubbornly-high-for-three-consecutive-years-as-global-crises-deepen–un-report/en⟩ (accessed 1 August 2025).

Fu, H., Lu, J., Li, J., Zou, W., Tang, X., Ning, X., Sun, Y., 2025. Winter wheat yield prediction using satellite remote sensing data and deep learning models. Agronomy 15, 1–21. https://doi.org/10.3390/agronomy15010205.

Hengl, T., 2018. Sand content in \% (kg / kg) at 6 standard depths (0, 10, 30, 60, 100 and 200 cm) at 250 m resolution. https://doi.org/10.5281/zenodo.2525662.

Hengl, T., 2018. Clay content in \% (kg / kg) at 6 standard depths (0, 10, 30, 60, 100 and 200 cm) at 250 m resolution. https://doi.org/10.5281/zenodo.2525663.

Hengl, T., 2018c. Soil pH in H2O at 6 standard depths (0, 10, 30, 60, 100 and 200 cm) at 250 m resolution. https://doi.org/10.5281/zenodo.2525664.

Hengl, T., Wheeler, I., 2018. Soil organic carbon content in x 5 g / kg at 6 standard depths (0, 10, 30, 60, 100 and 200 cm) at 250 m resolution. https://doi.org/10.5281/zenodo.2525553.

Jarlan, L., Abaoui, J., Duchemin, B., Ouldbba, A., Tourre, Y.M., Khabba, S., Le Page, M., Balaghi, R., Mokssit, A., Chehbouni, G., 2014. Linkages between common wheat yields and climate in Morocco (1982-2008). Int J. Biometeorol. 58, 1489–1502. https://doi.org/10.1007/s00484-013-0753-9.

Jenul, A., Schrunner, S., Liland, K.H., Indahl, U.G., Futsaether, C.M., Tomic, O., 2021. Rent - repeated elastic net technique for feature selection. IEEE Access 9, 152333–152346. https://doi.org/10.1109/ACCESS.2021.3126429.

Ji, L., Senay, G.B., Friedrichs, M., Schauer, M., Boiko, O., 2021. Characterization of water use and water balance for the croplands of Kansas using satellite, climate, and irrigation data. Agric. Water Manag. 256, 107106. https://doi.org/10.1016/j.agwat.2021.107106.

Joshi, A., Pradhan, B., Chakraborty, S., Behera, M.D., 2023. Winter wheat yield prediction in the conterminous United States using solar-induced chlorophyll fluorescence data and XGBoost and random forest algorithm. Ecol. Inf. 77, 102194. https://doi.org/10.1016/j.ecoinf.2023.102194.

Li, Z., Chen, Z., Cheng, Q., Duan, F., Sui, R., Huang, X., Xu, H., 2022. UAV-based hyperspectral and ensemble machine learning for predicting yield in winter wheat. Agronomy 12, 1–26. https://doi.org/10.3390/agronomy12010202.

Li, A., Liang, S., Wang, A., Qin, J., 2007. Estimating crop yield from multi-temporal satellite data using multivariate regression and neural network techniques. Photo Eng. Remote Sens. 73, 1149–1157. https://doi.org/10.14358/PERS.73.10.1149.

Li, Z., Zhou, X., Cheng, Q., Zhai, W., Mao, B., Li, Y., Chen, Z., 2023. An integrated feature selection approach to high water stress yield prediction. Front. Plant Sci. 14, 1–17. https://doi.org/10.3389/fpls.2023.1289692.

Lischeid, G., Webber, H., Sommer, M., Nendel, C., Ewert, F., 2022. Machine learning in crop yield modelling: A powerful tool, but no surrogate for science. Agric. Meteor. 312, 108698. https://doi.org/10.1016/j.agrformet.2021.108698.

Lobell, D.B., Burke, M.B., 2010. On the use of statistical models to predict crop yield responses to climate change. Agric. Meteor. 150, 1443–1452. https://doi.org/10.1016/j.agrformet.2010.07.008.

McKee, T.B., Doesken, N.J., Kleist, J., 1993. The Relationship of Drought Frequency and Duration to Time Scales. J. Surg. Oncol. 105, 179–184. https://doi.org/10.1002/jso.23002.

Melton, F.S., Huntington, J., Grimm, R., Herring, J., Hall, M., Rollison, D., Erickson, T., Allen, R., Anderson, M., Fisher, J.B., Kilic, A., Senay, G.B., Volk, J., Hain, C., Johnson, L., Ruhoff, A., Blankenau, P., Bromley, M., Carrara, W., Daudert, B., Doherty, C., Dunkerly, C., Friedrichs, M., Guzman, A., Halverson, G., Hansen, J., Harding, J., Kang, Y., Ketchum, D., Minor, B., Morton, C., Ortega-Salazar, S., Ott, T., Ozdogan, M., ReVelle, P.M., Schull, M., Wang, C., Yang, Y., Anderson, R.G., 2022. OpenET: filling a critical data gap in water management for the western United States. J. Am. Water Resour. Assoc. 58, 971–994. https://doi.org/10.1111/1752-1688.12956.

Myneni, R., Knyazikhin, Y., Park, T., 2021. MODIS/Terra+Aqua Leaf Area Index/FPAR 4-Day L4 Global 500m SIN Grid V061 [Data set]. NASA EOSDIS Land Processes Distributed Active Archive Center. https://doi.org/10.5067/MODIS/MCD15A3H.061.

Naghdyzadegan Jahromi, M., Zand-Parsa, S., Razzaghi, F., Jamshidi, S., Didari, S., Doosthosseini, A., Pourghasemi, H.R., 2023. Developing machine learning models for wheat yield prediction using ground-based data, satellite-based actual evapotranspiration and vegetation indices. Eur. J. Agron. 146, 126820. https://doi.org/10.1016/j.eja.2023.126820.

Panek, E., Gozdowski, D., Stępień, M., Samborski, S., Ruciński, D., Buszke, B., 2020. Within-field relationships between satellite-derived vegetation indices, grain yield and spike number of winter wheat and triticale. Agronomy 10, 1–18. https://doi.org/10.3390/agronomy10111842.

Pede, T., Mountrakis, G., Shaw, S.B., 2019. Improving corn yield prediction across the US Corn Belt by replacing air temperature with daily MODIS land surface temperature. Agric. Meteor. 276277, 107615. https://doi.org/10.1016/j.agrformet.2019.107615.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, É., 2011. Scikit-learn: machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830.

Petersen, L.K., 2018. Real-time prediction of crop yields from MODIS relative vegetation health: a continent-wide analysis of Africa. Remote Sens 10, 1–31. https://doi.org/10.3390/rs10111726.

Reynolds, M.P., Braun, H.J., 2022. Wheat Improvement: Food Security in a Changing Climate, pp. 1–629. https://doi.org/10.1007/978-3-030-90673-3.

Schaaf, C., Wang, Z., 2021. "MODIS/Terra+Aqua BRDF/Albedo Nadir BRDF Adjust. Ref. Dly. L3 Glob. 500m V061 [Data Set.]. NASA EOSDIS Land Process. Distrib. Act. Arch. Cent. URL. https://doi.org/10.5067/MODIS/MCD43A4.061.

Schwalbert, R.A., Amado, T., Corassa, G., Pott, L.P., Prasad, P.V.V., Ciampitti, I.A., 2020. Satellite-based soybean yield forecast: integrating machine learning and weather data for improving crop yield prediction in southern Brazil. Agric. Meteor. 284, 107886. https://doi.org/10.1016/j.agrformet.2019.107886.

Senay, G.B., Bohms, S., Singh, R.K., Gowda, P.H., Velpuri, N.M., Alemu, H., Verdin, J.P., 2013. Operational evapotranspiration mapping using remote sensing and weather datasets: a new parameterization for the SSEB approach. J. Am. Water Resour. Assoc. 49, 577–591. https://doi.org/10.1111/jawr.12057.

Senay, G.B., Friedrichs, M., Morton, C., Parrish, G.E.L., Schauer, M., Khand, K., Kagone, S., Boiko, O., Huntington, J., 2022. Mapping actual evapotranspiration using Landsat for the conterminous United States: Google Earth Engine implementation and assessment of the SSEBop model. Remote Sens. Environ. 275, 113011. https://doi.org/10.1016/j.rse.2022.113011.

Sridhara, S., Ramesh, N., Gopakkali, P., Das, B., Venkatappa, S.D., Sanjivaiah, S.H., Singh, K.K., Singh, P., Al-Ansary, D.O., Mahmoud, E.A., Elansary, H.O., 2020. Weather-based neural network, stepwise linear and sparse regression approach for rabi sorghum yield forecasting of Karnataka, India. Agronomy 10, 1–24. https://doi.org/10.3390/agronomy10111645.

Tian, H., Wang, P., Tansey, K., Zhang, J., Zhang, S., 2021. An LSTM neural network for improving wheat yield estimates by integrating remote sensing data and meteorological data in the Guanzhong. Agric. Meteor. 310, 108629. https://doi.org/10.1016/j.agrformet.2021.108629.

United Nations, 2025. Population. ⟨https://www.un.org/en/global-issues/population⟩.

USDA NASS, 2023. Quick Stats. ⟨https://quickstats.nass.usda.gov/⟩.

USDA NASS, 2024. Crop Production 2023 Summary. ⟨https://downloads.usda.library.cornell.edu/usda-esmis/files/k3569432s/ns065v292/8910md644/cropan24.pdf⟩.

Wan, Z., Hook, S. and Hulley, G., 2021. MODIS/Terra Land Surface Temperature/Emissivity Daily L3 Global 1km SIN Grid V061 [Data set]. NASA EOSDIS Land Processes Distributed Active Archive Center. URL https://doi.org/10.5067/MODIS/MOD11A1.061.

Wang, Y., Zhang, Z., Feng, L., Du, Q., 2020. Combining multi-source data and machine learning approaches to predict winter wheat yield in the conterminous United States. Remote Sens. https://doi.org/10.3390/rs12081232.

Yang, S., Zeng, J., Fan, W., Cui, Y., 2022. Evaluating root-zone soil moisture products from GLEAM, GLDAS, and ERA5 based on in situ observations and triple collocation method over the Tibetan Plateau. J. Hydrometeorol. 23, 1861–1878. https://doi.org/10.1175/JHM-D-22-0016.1.

Zhang, N., Zhao, C., Quiring, S.M., Li, J., 2017. Winter wheat yield prediction using normalized difference vegetative index and agro-climatic parameters in Oklahoma. Agron. J. 109, 2700–2713. https://doi.org/10.2134/agronj2017.03.0133.

Zhang, L., Zhao, H., Wan, N., Bai, G., Kirkham, M.B., Nielsen-Gammon, J.W., Avenson, T. J., Lollato, R., Sharda, V., Ashworth, A., Gowda, P.H., Lin, X., 2024. An unprecedented fall drought drives Dust Bowl–like losses associated with La Niña events in US wheat production. Sci. Adv. 10, 1–8. https://doi.org/10.1126/sciadv.ado6864.