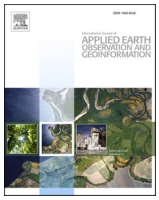
International Journal of Applied Earth Observation and Geoinformation 118 (2023) 103269

Contents lists available at ScienceDirect

International Journal of Applied Earth Observation and Geoinformation

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



A county-level soybean yield prediction framework coupled with XGBoost and multidimensional feature engineering

Yuanchao Li [a,b], Hongwei Zeng [a,b,*], Miao Zhang [a], Bingfang Wu [a,b,*], Yan Zhao [c], Xia Yao [d], Tao Cheng [d], Xingli Qin [a], Fangming Wu [a]

[a] State Key Laboratory of Remote Sensing Science, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100101, China
[b] College of Resources and Environment, University of Chinese Academy of Sciences, Beijing 100049, China
[c] The University of Queensland, Queensland Alliance for Agriculture and Food Innovation, St Lucia, 4067, Australia
[d] National Engineering and Technology Center for Information Agriculture(NETCIA), MARA Key Laboratory for Crop System Analysis and Decision Making, MOE Engineering Research Center of Smart Agricultulre, Jiangsu Key Laboratory for Information Agriculture, Collaborative Innovation Center for Modern Crop Production co-sponsored by Province and Ministry, Nanjing Agricultural University, Nanjing 210095, China

ARTICLE INFO

Keywords:
Soybean
Yield prediction
XGBoost
SHAP
Multidimensional feature engineering

ABSTRACT

Yield prediction is essential in food security, food trade, and field management. However, due to the associated complex formation mechanisms of yield, accurate and timely yield prediction remains challenging in remote sensing-based crop monitoring domains. In this study, a framework of soybean yield prediction integrating extreme gradient boosting (XGBoost) and multidimensional feature engineering was developed at the county level in the United States using publicly available datasets. Excellent accuracy values were obtained for over 959 counties in 12 states throughout the midwestern U.S., with a test coefficient of determination (R^2) of 0.82 and a root-mean-square error (RMSE) of 0.246 t/ha, using our approach. Following a "train–validate–test" assessment strategy, our study shows that XGBoost outperforms other county-level soybean yield prediction models with identical inputs, including linear regression (LR), random forest (RF), k-nearest neighbor (KNN), artificial neural network (ANN), support vector regression (SVR), long short-term memory (LSTM), and deep neural network (DNN). The results show that accurate results of soybean yield prediction can be obtained as early as the pod-setting stage. We implemented the feature importance and Shapley additive explanations (SHAP) algorithms to quantify the impact of input features on the XGBoost model in the training and prediction stages, respectively. The enhanced vegetation index (EVI) at the pod-setting period is the most crucial factor, but the yield prediction is not dependent on only a few key features. Yields were detrended using longer-term historical yield data, and R^2 increased from 0.58 to 0.82 while RMSE decreased from 0.374 t/ha to 0.246 t/ha. We employed multidimensional feature engineering to generate phenology-based features, and R^2 improved from 0.79 to 0.82 while RMSE decreased from 0.268 t/ha to 0.246 t/ha using this approach. The framework can be easily implemented and extended in the future in combination with early crop identification.

1. Introduction

Soybeans are the world's largest source of protein for feed and the second-largest source of vegetable oil (Song et al., 2021). As the top soybean producer and exporter, the U.S. produced 115 million tons of soybeans in 2020, accounting for 39.9% of world soybean exports (USDA/NASS, 2021). Slight changes in U.S. soybean production can cause significant fluctuations in the world soybean market. Therefore, accurate and timely U.S. soybean production forecasts play an important role in international food trade and food security (Fritz et al., 2019). However, it is challenging to accurately predict soybean yield over large areas in a timely manner due to the complex formation mechanism of yield and the complicated impacts of weather, soil, vegetation, and management on yield (Klompenburg et al., 2020).

Physical process-based and machine learning models are the current mainstream yield prediction methods (Archontoulis et al., 2020; Feng et al., 2020; Maimaitijiang et al., 2020). Although process-based models are valid and more interpretable (Jagtap and Jones, 2002; Malik et al.,

* Corresponding authors at: State Key Laboratory of Remote Sensing Science, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100101, China.
E-mail addresses: zenghw@aircas.ac.cn (H. Zeng), wubf@aircas.ac.cn (B. Wu).

https://doi.org/10.1016/j.jag.2023.103269
Received 30 August 2022; Received in revised form 24 January 2023; Accepted 17 March 2023
Available online 27 March 2023

2018), their stringent requirements regarding field and local sensing data make them difficult to apply to large-scale yield prediction (Kang and Ozdogan, 2019; Zhang et al., 2019b). In contrast, machine learning does not require a variety of sensors to be installed in a field to obtain specific crop parameters (Wang et al., 2020c; Zhang et al., 2019b) and instead learns empirical relationships from historical yield records and various yield-related indicators (Ma et al., 2021). Machine learning has a strong ability to mine and learn empirical relationships. With the rapid advances in computer software and hardware, machine learning and deep learning are increasingly applied to large-scale yield predictions (Klompenburg et al., 2020). For example, RF, SVR, convolutional neural network (CNN), recurrent neural network (RNN), LSTM, and 3-dimensional CNN (3DCNN) have been used in crop yield prediction at the county level (Cai et al., 2019; Gavahi et al., 2021; Khaki et al., 2019; Sakamoto, 2020; Sun et al., 2019).

Data and models are critical aspects of machine learning-based yield prediction methods. From a data perspective, the aim is to select or generate suitable features from multiple available raw data sources for the constructed yield prediction model (Elavarasan et al., 2020; Panda et al., 2010). Yield is influenced by the seeds, climate, soil, topography, fertilization, and management. While many studies have used different methods to extract various yield-related indicators, effectively combining indicators and experience for better yield prediction still requires study. Early-stage studies mainly used satellite data for yield prediction (Mkhabela et al., 2011). The establishment and application of vegetation indices derived from the spectral reflectance values of satellite surfaces act as the most frequently used indicators in yield prediction (Bolton and Friedl, 2013). These indicators include the normalized difference vegetation index (NDVI), enhanced vegetation index (EVI), green chlorophyll vegetation index (GCVI), ratio vegetation index (RVI), soil-adjusted vegetation index (SAVI), comprehensive dynamic range vegetation index (WDRVI) and net primary productivity (NPP) (Jaafar and Ahmad, 2015). Climatic indicators are also frequently used to explain the impact of climate on yield (Jiang et al., 2020), such as the standardized precipitation evaporation index (SPEI) and the Palmer drought severity index (PDSI), as well as the growing degree days (GDDs) and killing degree days (KDDs) in combination with crop phenology. Subsequently, most studies have found that approaches using combinations of various data sources, including satellite data, climate data, soil data, field management data, and historical yield data, outperform approaches using single types of data (Cai et al., 2019; Johnson, 2014; Ma et al., 2021; Sakamoto, 2020). Some studies have attempted to automatically extract yield-related information from satellite data using a deep CNN–Gaussian process (GP) model (You et al., 2017). Nevertheless, later studies have shown that using only satellite data is insufficient. Combining features with human domain knowledge in traditional machine learning and deep learning models can provide better predictions (Fan et al., 2022; Han et al., 2020; Hunt et al., 2019; Ma et al., 2021; Schwalbert et al., 2020). In previous studies, a few factors have generally been selected for prediction, and they have not been extensively compared or explained in sufficient depth (Sakamoto, 2020). We used multiple factors in our modeling process to comprehensively assess the effects of different factors on soybean yield prediction.

The model is the second critical aspect in yield prediction. Machine learning approaches without deep network structures and deep learning methods have been widely used in yield prediction (Klompenburg et al., 2020). However, which is the best model and whether deep learning is better than machine learning are inconclusive (Klompenburg et al., 2020). For instance, Cao et al. (2021a) found that RF and LSTM performed well in predicting wheat yields at China's county and field scales. Another study showed that adaptive boosting (AdaBoost) is more accurate than a DNN deep learning model in wheat yield prediction at the county scale in the U.S. (Wang et al., 2020c). Kang et al. (2020) found that the XGBoost algorithm outperformed LSTM and CNN methods regarding accuracy and stability. Yield prediction at the county level faces two challenges: one is limited samples, and the other originates from the black box effect of machine learning (Cao et al., 2021b), especially for deep learning models. Deep learning models applied to yield prediction face the risk of insufficient sample sizes and overfitting (Ma et al., 2021). Although deep learning has achieved better results than machine learning in some areas, this advancement can make the black box problem more prominent (Castelvecchi, 2016). As AI techniques continue to advance, the models used for yield prediction are also changing and more advanced. Therefore, a fair and extensive comparison of machine learning and deep learning is necessary to obtain more accurate prediction results and descriptive studies for future yield predictions.

Organizing features and models well is another exploration direction for yield prediction. Several studies have improved feature engineering and model combinations to achieve more accurate yield prediction. For example, Bocca and Rodrigues (2016) obtained a mean absolute error (MAE) of 4.49 Mg ha^{-1} on the test set when using the "RF + decomposed weather" strategy to predict sugarcane yield. Jiang et al. (2020) proposed an "LSTM + Phenology" approach for maize yield prediction in the Corn Belt and obtained results with an R^2 of 0.76. With the "RF + biomass simulation" strategy, Feng et al. (2020) obtained satisfactory predictions with an RMSE of 0.7 t/ha for wheat in southeastern Australia one month before harvest. This combination needs to be further investigated. These approaches are an example of how physical process-based models and machine learning models can be perfectly combined in the future since both feature engineering and physical process-based models are essentially a concretization of human empirical knowledge in the field of agriculture.

In addition, near real-time prediction is also a key direction for yield prediction research, which will bring real-time feedback to agricultural field management decisions. For example, Sakamoto et al. (2014) achieved near real-time forecasting of maize at national and state levels by constructing a linear regression model for the MODIS wide dynamic range vegetation index (WDRVI). Ma et al. (2021b) made near real-time predictions of maize yield at the county level and assessed uncertainty. However, the question remains as to why the model's prediction accuracy changes over time. This paper will address this question in conjunction with the growing phenology of soybeans and SHAP (Lundberg and Lee, 2017).

Aiming to improve the prediction performance of soybean yield on a large scale and reveal how environmental stresses affect soybean yield, this study attempts to design a new soybean yield prediction framework at the county level by integrating the latest XGBoost model and multidimensional feature engineering. The static and dynamic features of soybean yield, phenological information, and agricultural knowledge are well organized for better soybean yield prediction. The specific aims of this study include the following:

(1) To design a new soybean yield prediction framework at the county level to achieve high-accuracy prediction by combining multidimensional feature engineering and XGBoost.
(2) To quantify the importance of multidimensional features during the training and explain the impact of features on soybean yield prediction.
(3) To determine the optimal period for soybean yield prediction and understand the rationale behind the differences in yield forecasting performance among different models.

2. Materials and methods

2.1. Study area

The study area is located in the midwestern U.S. and includes 959 counties across 12 states (Fig. 1). As the dominant soybean-producing area in the world, the midwestern U.S. accounted for 83.4% and 82.4% of the total soybean production and the number of planted acres

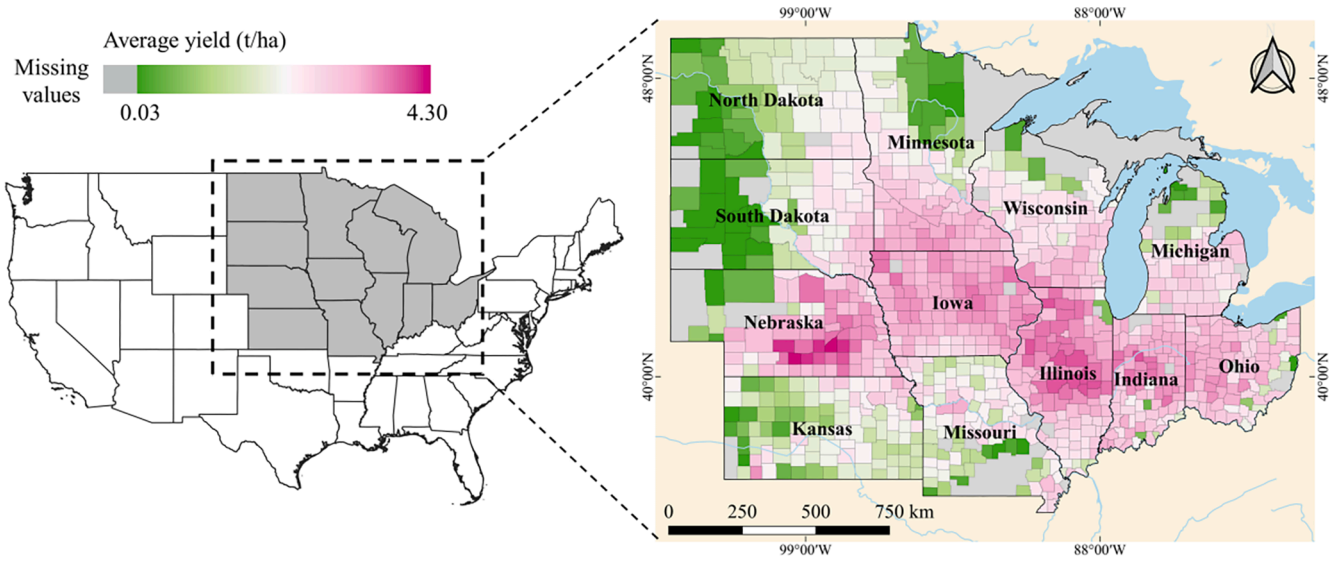


Fig. 1. Geographic location of the study area and average soybean yields of 959 counties in the midwestern United States, 2003–2020.

in the U.S. in 2020, respectively (USDA/NASS, 2021). Soybean yield at the county level ranged from 0.03 t/ha to 4.30 t/ha. The available soybean yield data provide an excellent resource from which a robust yield prediction framework can be built. The midwestern U.S. is one of four statistical regions defined by the U.S. Census Bureau since 1950 (United States Census Bureau, 2013), ensuring the consistency and reliability of statistical methods and data for surveys and censuses. The yield data and environmental factors for soybeans in the midwestern U. S. are highly representative and can enhance the robustness of soybean yield forecasting models. States with different irrigation intensities, such as Nebraska, which accounts for 14.8% of irrigated U.S. farmland, and Iowa and Illinois, which account for <0.5% of irrigated farmland, are covered in the study area. The study area also covers areas affected by various hazards. For example, six states in the plains and the Midwest (Arkansas, Indiana, Iowa, Kansas, Missouri, and Nebraska) experienced a severe, widespread drought in 2012 (Smith and Matthews, 2015).

2.2. Data

2.2.1. Yield data at the county level and processing

Historical statistical county-level soybean yield records from 2003 to 2020 in the midwestern U.S. were acquired from the Quick Stats Database of the United States Department of Agriculture (USDA), National Agricultural Statistics Service (USDA/NASS, 2021). The duration of the yield data was synchronized with the earliest start time among the explanatory variables, such as the fraction of photosynthetically active radiation (FPAR) and leaf area index (LAI), which were available as early as July 4th, 2002 (Myneni et al., 2015). The soybean yield data for each county were not always recorded annually due to the rotation of maize and soybeans. The final yield data sample size was 10,978 after cleaning by executing quality control on the explanatory variables and identifying the soybean planting regions.

Understanding yield data is a prerequisite for subsequent modeling and prediction tasks. Several statistical methods were used in the temporal and spatial analysis of the historical yield records. The augmented Dickey–Fuller (ADF) (Mushtaq, 2011) test was used to check the stationarity of the historical yield time series from 1980 to 2019. The ADF statistic was −1.99 (p = 0.29), meaning that the soybean yield time series were nonstationary. A significant increasing trend was observed in the yield time series from 1980 to 2020 (r = 0.52, slope = 0.03 t/ha, p < 0.001) (Fig. 2a). The increasing trend in yield could be attributed to technological advancement, including advances in genetics, seed variety, the use of fertilizer, and field management (Fuglie, 2007; Shahhosseini et al., 2020). However, these data are difficult to collect due to time-consuming and labor-intensive collection processes (Hussain and Thapa, 2012) or commercial confidentiality (Blair, 1999). Some studies used the annual average yield and year as a factor to reflect technological advances ((Ma et al., 2021)). Some researchers believed that the trend should be removed before applying some specific predetermined function, such as a simple linear regression model (Quiring and

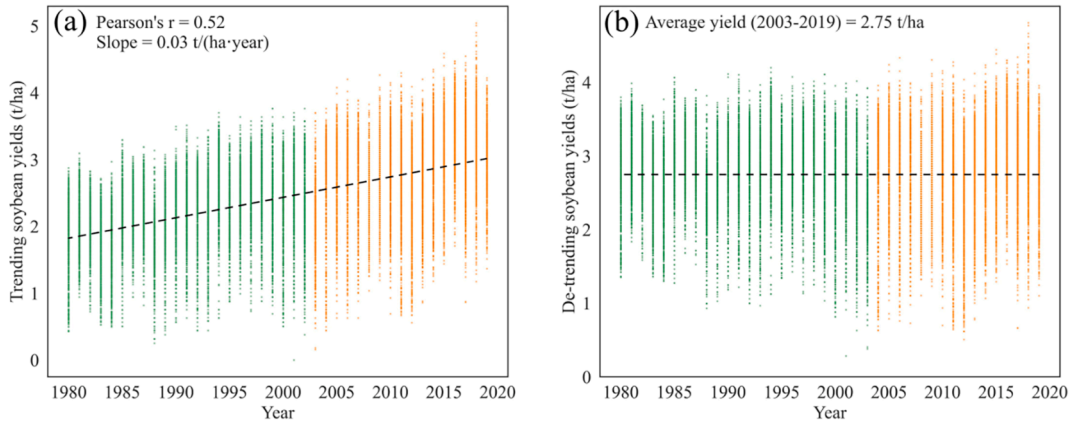


Fig. 2. (a) Linear trend of soybean yield from 1980 to 2020 and (b) detrended soybean yield data from 1980 to 2020.

Papakryiakou, 2003) or a second-order polynomial regression model (Hlavinka et al., 2009). If the trend is not removed, it will affect the interannual prediction caused by climate fluctuations. In this study, detrended soybean yield data from 2003 to 2020 were obtained by using global linear regression (Fig. 2b) with the following equation.

$$Yield_{de,i} = Yield_i - slope^{*}year - b \quad (1)$$

where $Yield_{de,i}$ and $Yield_i$ indicate the postdetrend and predetrend soybean yields of county i, respectively, and $slope$ and b are 0.03 t/(ha·year) and −61.46 t/ha, respectively.

The quartiles and the shape of the detrended yield distribution from 2003 to 2020 are shown in Fig. 3a. The cumulative probability distribution of the yield residuals derived from Monte Carlo simulations (Schwalbert et al., 2020) is shown in Fig. 3b. The average cumulative probabilities for 2003, 2008, 2012, and 2016 were 0.25, 0.39, 0.33, and 0.71, respectively, where the average yield residuals for 2003, 2008, and 2012 were negative, while the average yield residuals for 2016 were positive. The average cumulative probability indicates that the incidence periods of low-yielding soybean years in 2003, 2008, and 2012 were once every 4, 2.5, and 3 years, respectively, while the high-yielding soybean year in 2016 occurred approximately once every 3.5 years. The temporal analysis shows significant trending and volatility in the yield data, which are characteristics that can increase uncertainty in data processing methods and model evaluation steps.

Fig. 3c shows the distributions of the soybean yield data for the 12 states. Significant spatial heterogeneity occurred in the soybean yield among the different states. Illinois, Indiana, Iowa, and Nebraska had the highest soybean yields, averaging over 3.0 t/ha, while North Dakota had the lowest soybean yield, averaging just under 2.0 t/ha. Illinois, Kansas, Minnesota, and South Dakota had narrow and long violin plots, indicating relatively high county-to-county variability in the soybean yield in these states. Indiana, Iowa, North Dakota, and Ohio had wide and short violin distributions, indicating less variability in the soybean yield between the counties in these states. These two differences reflect the regional polarization of yields in space and represent a challenge to the single-model approach.

2.2.2. Phenological information

In this study, phenological data were collected from the weekly reports of the Economics, Statistics and Market Information System, USDA. The phenological data cover the entire growing season of soybeans (April to November), which includes planting, fruiting, and harvesting denoted by area percentage in major producing states (https://usda.library.cornell.edu/). The phenological stages of soybeans are shown in Fig. 4. The nonprobability crop progress survey included inputs from approximately 3600 respondents to subjectively estimate crop progress levels at various stages of development. In this study, daily frequency phenology data were generated by resampling the week-frequency phenology data using cubic polynomial interpolation. Then, the state phenological period was defined as the time during which phenological progress was between 25% and 75%. Fig. 4 shows the phenological stages in 12 states in 2020. The phenology of soybeans exhibited wide temporal–spatial variations due to geographical location, climate pattern, and weather condition differences. For example, soybeans in Iowa were planted 20 days earlier than in North Dakota in 2020. In addition, the soybean growing progress in the same state can occur as much as half a month earlier or later in different years. Such a time interval exceeds the satellite revisiting period and is sufficient to influence the satellite monitoring results.

2.2.3. Soybean layer

The soybean layer was used to identify the soybean planting area for

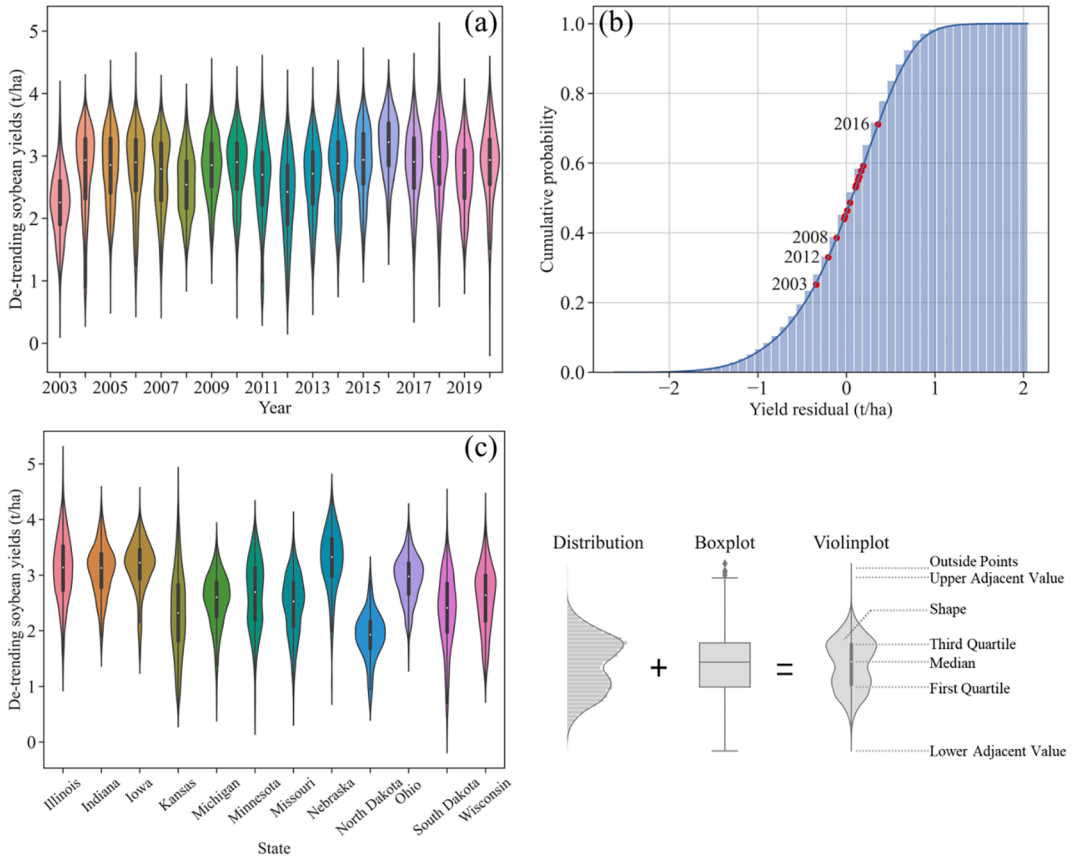


Fig. 3. (a) Violin plot of soybean yield from 2003 to 2019; (b) cumulative probability distribution of residuals with average yield; and (c) violin plot of soybean yield in 12 states.

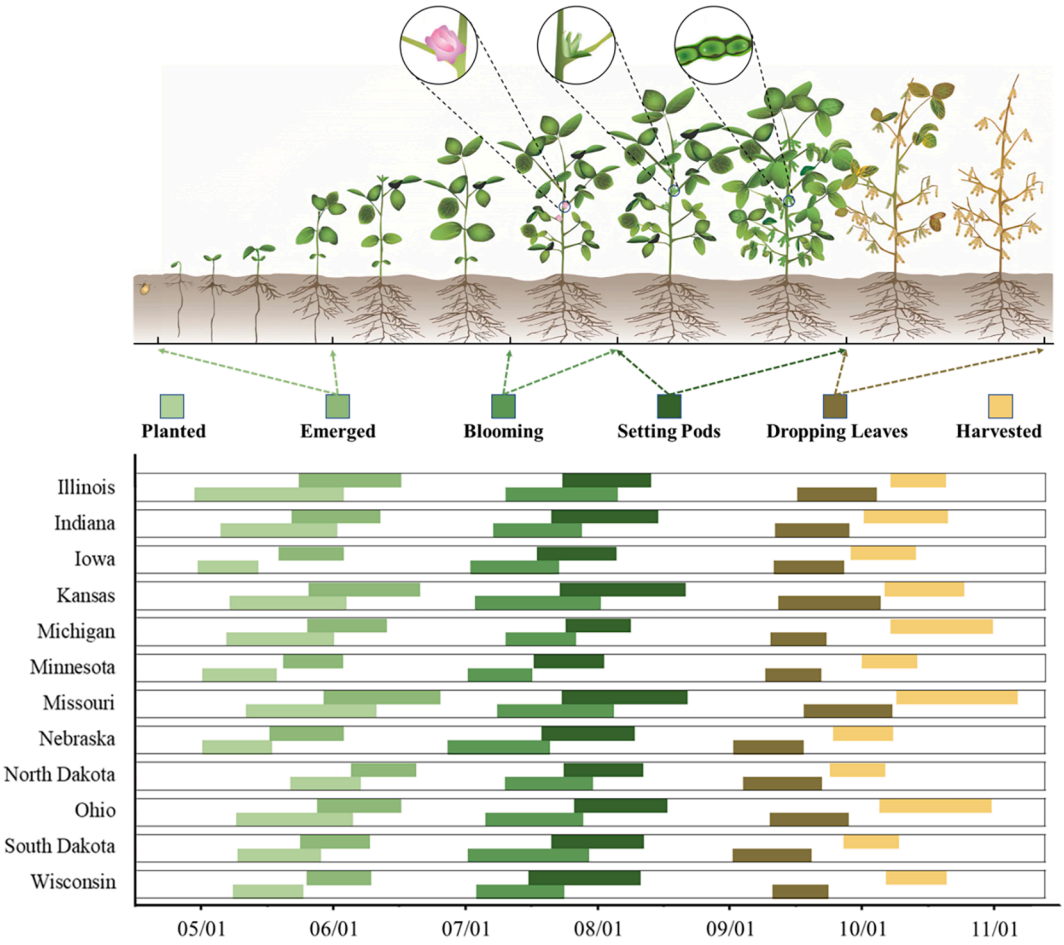


Fig. 4. Soybean planting, emerging, blooming, pod setting, leaf dropping and harvesting stage for 12 states in 2020 (Note: Soybean growth diagram reference: https://www.dekalbasgrowdeltapine.com).

each year. In this study, the cropland data layer (CDL) (Boryan et al., 2011) from 2008 to 2020 and CSDL data (Wang et al., 2020a) from 2003 to 2007 were used as the soybean CDL. The CDL is a crop-specific land cover data layer created annually for the continental U.S. using moderate-resolution satellite imagery and extensive agricultural ground truths. It is used in numerous agricultural yield prediction studies in the U.S. However, CDL data were not available in some states before 2008. To address the issue of missing CDL data prior to 2008 in this study, CSDL data published by Wang et al. (2020a) were obtained. These data were obtained using machine learning algorithms with CDL data as training samples.

2.2.4. Satellite products and meteorological data

Table 1 shows a summary of the satellite data and climatic data used in our study. The satellite data include seven land surface reflectance bands of MOD09A1 (Vermote, 2015) and their vegetation indices generated from reflectance bands, including the land surface water index (LSWI), NDVI, EVI, GCVI, RVI, SAVI, and WDRVI. Land surface temperature (LST) data include the LSTs at day (LST_{Day}) and night (LST_{Night}) from MOD11A2 (Wan et al., 2015) and the difference (LST_{dif}) between LST_{Day} and LST_{Night}. FPAR and LAI data were extracted from MCD15A3H (Myneni, 2015). Total evapotranspiration (ET) and average latent heat flux (LE) were extracted from MOD16A2 (Running et al., 2017). The pixels of MODIS products inflected by clouds, the atmosphere, and other noise were removed by the "QA" or "QC" bands. Climatic data consist of precipitation, temperature, humidity, pressure, and radiation data from phase 2 of the North American Land Data Assimilation System (NLDAS-2) (Cosgrove et al., 2003) and GRIDMET drought indices (Abatzoglou, 2013), including standardized precipitation index (SPI) for 14 days (SPI_14d), 30 days (SPI_30d), and 90 days (SPI_90d), evaporative drought demand index (EDDI) for 14 days (EDDI_14d), 30 days (EDDI_30d), and 90 days (EDDI_90d), standardized precipitation evapotranspiration index (SPEI) for 14 days (SPEI_14d), 30 days (SPEI_30d), and 90 days (SPEI_90d), palmer drought severity index (PDSI) and palmer Z index (Palmer_Z).

2.2.5. Soil, irrigation, and location data

In this study, soil data were collected from the soil survey geographic database (SSURGO) of the Natural Resources Conservation Service Soils, USDA (Walkinshaw et al., 2021). Soil data at a depth of 0–25 cm include chemical (calcium carbonate content, cation exchange capacity, electrical conductivity, pH, sodium adsorption ratio, and maximum organic soil matter), physical (available water holding capacity, drainage class, soil texture, sand content, silt content, clay content, and soil type), and land capability (nonirrigated land capability class, irrigated land capability class) properties (Table 2). The county soil texture was aggregated by the majority for each county, while other soil properties of the county were obtained by aggregating the mean of all pixels. In addition, we collected irrigated and nonirrigated cropland area data for 2002, 2007, 2012, and 2017 from the USDA; census data were taken once every five years in these areas (USDA/NASS, 2021). Because there are many missing data at the county level for the noncensus year, the nearest values of the percentage of irrigated cropland (PIC), which is defined as the ratio of irrigated cropland area to cropland area and represents the county's irrigation level, for five years was assigned to the available PIC as an explanatory variable. PICs for 2003 and 2004 were derived from

Table 1
Summary of the satellite data and climatic data.

Data source	Feature	Formulation/Wavelength/Description	Frequency	Resolution
MOD09A1	Red	620–670 nm	8-day	500 m
	Nir	841–876 nm		
	Blue	459–479 nm		
	Green	545–565 nm		
	Nir_1	1230–1250 nm		
	Swir1	1628–1652 nm		
	Swir2	2105–2155 nm		
	LSWI	(Nir-Swir1)/(Nir + Swir1)		
	NDVI	(Nir-Red)/(Nir + Red)		
	EVI	2.5× (Nir-Red)/(Nir + 6*Red-7.5*Blue + 1)		
	GCVI	Nir/Green – 1		
	RVI	Nir/Red		
	SAVI	1.5×(Nir-Red)/(Nir + Red + 0.5)		
	WDRVI	(α*Nir – Red)/(α*Nir + Red), α = 0.1		
MOD11A2	LST_{Day}	Day land surface temperature	8-day	1000 m
	LST_{Night}	Night land surface temperature		
	LST_{dif}	$LST_{Day} - LST_{Night}$		
MCD15A3H	FPAR	Fraction of photosynthetically active radiation	4-day	500 m
	LAI	Leaf area index		
MOD16A2	ET	Evapotranspiration	8-day	500 m
	LE	Latent heat flux		
GRIDMET DROUGHT	SPI_14d	SPI where precipitation was aggregated for the last 14 days	daily	4638.3 m
	SPI_30d	SPI where precipitation was aggregated for the last 30 days		
	SPI_90d	SPI where precipitation was aggregated for the last 90 days		
	EDDI_14d	EDDI where PET was aggregated for the last 14 days		
	EDDI_30d	EDDI where PET was aggregated for the last 30 days		
	EDDI_90d	EDDI where PET was aggregated for the last 90 days		
	SPEI_14d	SPEI where climatic water balance was aggregated for the last 14 days		
	SPEI_30d	SPEI where climatic water balance was aggregated for the last 30 days		
	SPEI_90d	SPEI where climatic water balance was aggregated for the last 90 days		
	PDSI	Palmer Drought Severity Index		
	Palmer_Z	Palmer Z Index		
NLDAS-2	Precipitation	Hourly total precipitation	hourly	13915 m
	Temp	Air temperature at 2 m above the surface		
	Humidity	Specific humidity at 2 m above the surface		
	Pressure	Surface pressure		
	Shortwave	Surface downward shortwave radiation		
	Longwave	Surface downward longwave radiation		

2002 data; PICs for 2005, 2006, 2008, and 2009 were derived from 2007 data; PICs for 2010, 2011, 2013, and 2014 were derived from 2012 data; and PICs for 2015, 2016, 2018, 2019, and 2020 were derived from 2017 data. Furthermore, the spatial information of the shapefile data itself, including the latitude and longitude information of each county centroid and the state where it is located, was also mined as a variable to depict the regional characteristics of yield.

2.3. Soybean yield prediction framework

The soybean yield prediction flowchart is presented in Fig. 5. The flowchart can be divided into three parts: data preprocessing, multidimensional feature engineering, and model optimization and comparison. Data preprocessing has been described in detail in Section 2.2. Multidimensional feature engineering consists of ten-day time series variables, phenology-based variables combined with soybean phenology and time series variables, and static variables. These new and original variables were concatenated with the detrended yields to generate inputs to the model. In model optimization and comparison, the model was constructed, then trained and validated, and the parameters were optimized in an iterative process. The optimized model was saved and used to predict soybean yields for a new test set. The prediction process is entirely consistent with reality for the 2020 test set, and near real-time prediction and importance analysis of features are performed based on the optimized model. The test results for each year were used for the stability assessment of XGBoost and comparison with other models. The importance, correlation, and optimal number of features will be further explored based on the optimization model to clarify the impact on yield.

2.3.1. Construction of multidimensional features

Multidimensional feature engineering aims to enrich and mine more features that may be useful for yield prediction and helps to improve the yield prediction ability of the utilized model (Heaton, 2016). It includes the construction of 10-day time series variables, the construction of phenology-based variables, and the processing of static variables. This study used a normalization method to convert numerical variables to eliminate the effect of variable magnitude differences on the model. Machine learning algorithms require that the individual dynamic numerical variables conform to the standard normal distribution; otherwise, they might reduce the model's predictive performance (Raju et al., 2020). Here, we used the Z score method (Jain et al., 2005) to transform all numerical variables into a standard normal distribution (Eq. (2)):

$$Z = \frac{x_i - \mu}{\sigma} \tag{2}$$

where x_i is the i^{th} variable, and μ and σ are the mean and standard de-

Table 2
Summary of the soil and irrigated cropland data.

Category	Variable	Description/Units	Range	Resolution
Soil	$CaCO_3$	Calcium carbonate content, kg/m^2	0–1849.0	800 m
	CEC	Cation exchange capacity, cmol/kg	0–250.6	800 m
	EC	Electrical conductivity, dS/m	0–411.3	800 m
	pH	pH	3.1–10.1	800 m
	SAR	Sodium adsorption ratio	0–9435.7	800 m
	Max_OM	Maximum soil organic matter percentage	0–1	800 m
	PAWS	Available water holding capability, cm	0–15	800 m
	Drainage	Drainage class	Integers 1–8	800 m
	Texture	Proportion of sand, silt, and clay	Integers 1–12	800 m
	Sand	Sand percentage by weight	0–99.8 (%)	800 m
	Silt	Silt percentage by weight	0–90.4 (%)	800 m
	Clay	Clay percentage by weight	0–88.4 (%)	800 m
	Soil type	Soil classification	Integers 0–78	800 m
	N_class	Nonirrigated land capability class	Integers 1–8	800 m
	I_class	Irrigated land capability class	Integers 1–8	800 m
Irrigation	PIC	Percentage irrigation of cropland	0–1	County level
Location	Latitude	Latitude of county centroid	36.2°–48.8° N	County level
	Longitude	Longitude of county centroid	80.7°–103.5° W	County level
	Position Code	State name	Integers 0–11	State level

viation, respectively. Finally, a total of 1,129 variables were generated by multidimensional feature construction.

(1) Construction of 10-day time series variables.

The construction of annual 10-day time series of features includes satellite and climate data between April and November (Table 3). Remote sensing reflectance, indices, and products were processed by quality checking, soybean layer masking, 10-day composite, temporal reconstruction (Eq. (3)), and county aggregation. First, the quality band masked the pixels of remote sensing reflectance, indices, and products contaminated by clouds. Second, pixels without soybean planting were masked using the soybean layer. Third, all datasets between April and November were composited at each pixel with a 10-day average. Finally, linear moving interpolation (Lepot et al., 2017) was employed to fill in the missing values for each pixel for every set of ten days that were missed due to the impact of clouds. The mathematical formula for temporal reconstruction in pixel location i is denoted as

$$y_i(t) = \Delta t_1 \frac{y_i(t+\Delta t_2) - y_i(t-\Delta t_1)}{\Delta t_1 + \Delta t_2} + y_i(t-\Delta t_1) \quad (3)$$

where $y_i(t)$ is the missing pixel value at time t and Δt_1 and Δt_2 are the time periods backward and forward to the first available value from time t, respectively. Finally, average aggregation was performed for each county and output for all soybean growing areas. Climate data were composited at each pixel with a 10-day average between April and November. A total of 936 variables were generated by the 10-day time series method.

(2) Construction of phenology-based variables.

A series of new phenology-based variables for vegetation indices, land surface temperature, and climate were generated by combining time-based features with the phenological stages using mean and sum operations at six soybean growing stages (Table 4). The phenology-based approach avoids the temporal limitations of the original time-based features and allows the model to learn how crop variables affect yields at different phenological stages. Water stress is the primary limiting factor for achieving maximum soybean yields (Lesk et al., 2016). We used drought data to quantify the water stress of soybeans at six growing stages. Here, drought conditions at each growing stage were categorized into no drought (more than −1.00), abnormally dry (−1.00 to −1.99), moderate drought (−2.00 to −2.99), severe drought (−3.00 to −3.99), extreme drought (−4.00 to −4.99), and exceptional drought (-5.00 or less) based on the PDSI value (Svoboda et al., 2002). A total of 192 phenology-based variables were generated.

(3) Static variables processing.

Machine learning algorithms require categorical features to be converted into numerical variables for model building. Count encoding, label encoding, and one-hot encoding (Buitinck et al., 2013) were adopted for categorical variable transforms (Table 5). Each county was divided into irrigated (PIC > 10%) and nonirrigated (PIC < 10%) categories based on the PIC value and then converted into numerical variables by label encoding, i.e., irrigated county (1) and nonirrigated county (0). A total of 15 soil properties were used as variables in modeling. Here, the soil type variable, which has many categories and is unordered, was converted into numerical variables by using count encoding. The position code containing information on the 12 states was transformed into 12 0–1 vectors using one-hot encoding.

2.3.2. Regressor: XGBoost

Overfitting and multicollinearity are common issues for machine learning and deep learning algorithms. Here, XGBoost (Chen and Guestrin, 2016), a tree-based model, was selected as a regressor to build a soybean yield prediction model after comprehensively considering the risk of overfitting, multicollinearity, and training speed. The classification and regression trees (CARTs) in XGBoost are constructed sequentially, and each new tree corrects the previous trees' errors. In other words, the prediction $\hat{y}_i$ in Eq. (4) is the sum of the outputs of all the trees. As a result of its sequential structure, XGBoost is unaffected by highly correlated features, reducing the feature multicollinearity problem (Parsa et al., 2020; Tianqi Chen et al., 2021). XGBoost uses an early stopping approach and L1/L2 regularization terms in the loss function to reduce the overfitting problem. XGBoost is designed with parallelization in computing feature gain, depth-first tree pruning, and algorithmic enhancements to ensure high accuracy and increase the training speed. Mathematically, XGBoost can be formulated as the following equation:

$$\hat{y}_i = \theta(x_i) = \sum_{k=1}^{K} f_k(x_i), f_k \in \mathscr{F} \quad (4)$$

where $\mathbf{x}_i$ is the vector of the i^{th} samples, K is the number of trees, f_k is a function in functional space $\mathscr{F}$, and $\mathscr{F} = \{f(\mathbf{x}) = w_{q(\mathbf{x})}\}(q : \mathbb{R}^m \to T, w \in \mathbb{R}^T)$ is the set of all possible CARTs. Each f_k has an independent tree structure q and leaf weights w. The objective function of XGBoost to be optimized is shown in Eq. (5).

$$\mathscr{L}(\theta) = \sum_{i}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \quad (5)$$

where l is a differentiable convex loss function that measures the difference between the prediction $\hat{y}_i$ and the target y_i and Ω is a regulari-

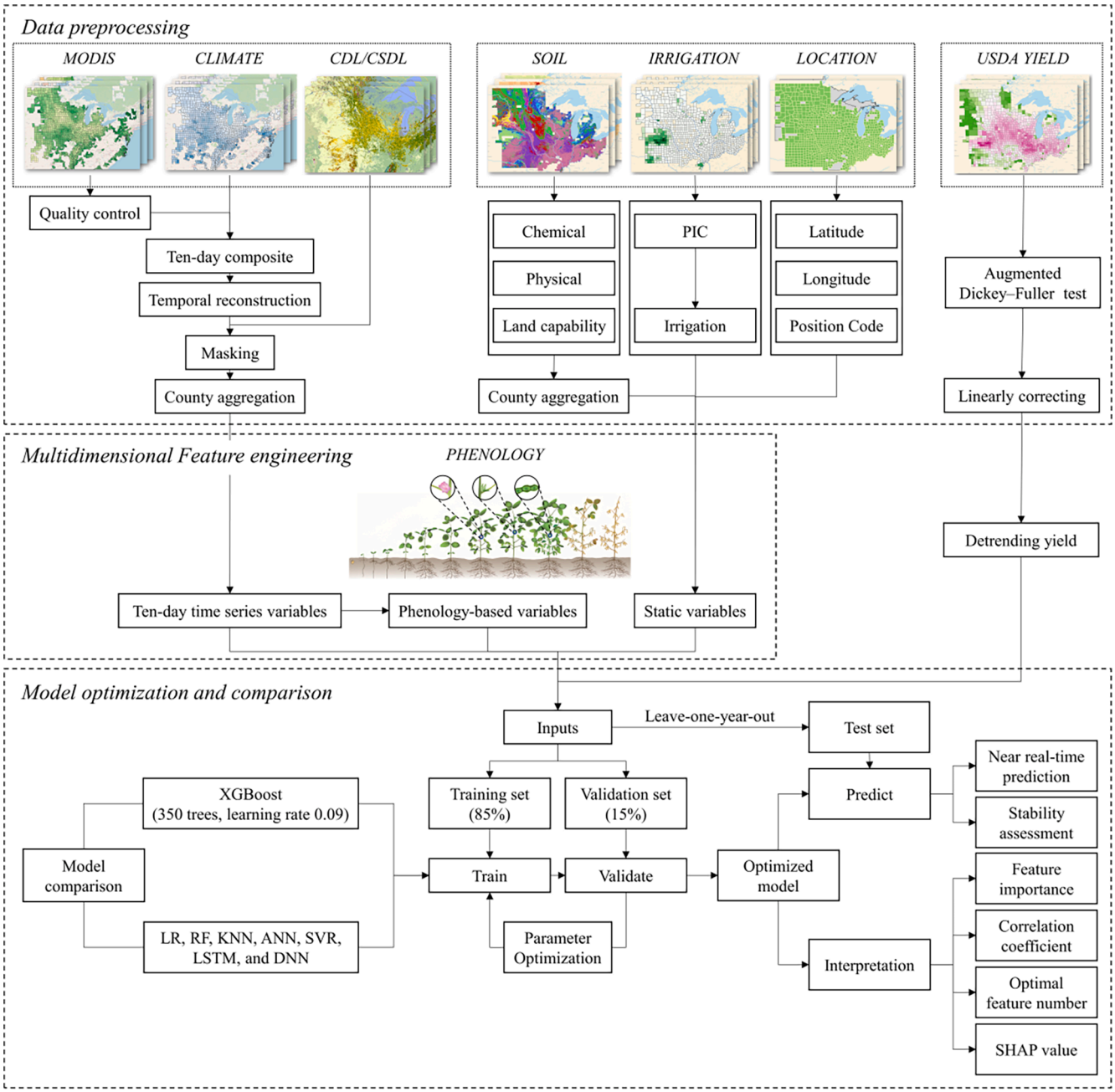


Fig. 5. Flowchart of the soybean yield prediction framework.

zation term penalizing the complexity of the model. Eq. (5) is decomposed in an additive way into an optimization for each tree since it cannot train it directly. For the prediction $\hat{y}_i^{\,t}$ of the i^{th} samples at the t^{th} iteration, a new tree f_t needs to be added to minimize the following objectives (Eq. (6)).

$$\mathscr{L}^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t) \tag{6}$$

which means that the new tree f_t is greedily added, which can most greatly improve XGBoost. Second-order approximation with Taylor expansion of the loss function is computed to speed up the optimization (Eq. (7)).

$$\widetilde{\mathscr{L}}^{(t)} = \sum_{i=1}^{n} \left[g_i f_t(\mathbf{x}_i) + \frac{1}{2} h_i f_t^2(\mathbf{x}_i) \right] + \Omega(f_t) \tag{7}$$

where $g_i = \partial_{\hat{y}^{(t-1)}} l(y_i, \hat{y}^{(t-1)})$ and $h_i = \partial^2_{\hat{y}^{(t-1)}} l(y_i, \hat{y}^{(t-1)})$ are first- and second-order gradient statistics on the loss function, respectively. The complexity of the model Ω is defined in XGBoost as

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \left\| w \right\|^2 \tag{8}$$

where T is the number of leaves, w is the leaf weight, and parameters γ and λ determine the strength of the penalization. This complexity control approach is effective in practice in preventing overfitting, whereas traditional tree learning processing methods emphasize only improving impurities and regularization handled less carefully or simply ignored (Eq. (8)).

The performance of XGBoost is compared with the other seven methods, including five machine learning models (LR, RF, KNN, ANN, and SVR) and two deep learning models (LSTM and DNN). The LR model

Table 3
Summary of 10-day time series variables.

Category	Features	Number Features	Variables
Satellite	Red, Nir, Blue, Green, Nir, Swir1, Swir2	7	7 × 24
	NDVI, EVI, LSWI, GCVI	4	4 × 24
	RVI, SAVI, WDRVI, FPAR, LAI	5	5 × 24
	LST_{Day}, LST_{Night}, LST_{dif}	3	3 × 24
	ET, LE	2	2 × 24
Climate	SPI_14d, SPI_30d, SPI_90d	3	3 × 24
	EDDI_14d, EDDI_30d, EDDI_90d	3	3 × 24
	SPEI_14d, SPEI_30d, SPEI_90d	3	3 × 24
	PDSI, Palmer_Z	2	2 × 24
	Precipitation, Temp (air), Humidity, Pressure, Shortwave, Longwave	6	6 × 24
	Drought based on Palmer Drought Severity Index (PDSI)	1	1 × 24

Table 4
Overview of new phenology-based variables.

Category	Features	Phenology	Composite method	Number	
				Features	Variables
Satellite	NDVI, EVI, LSWI, GCVI	6 stages*	Mean	4	4 × 6
	RVI, SAVI, WDRVI, FPAR, LAI	6 stages	Mean	5	5 × 6
	LSTDay, LSTNight, LSTdif	6 stages	Mean	3	3 × 6
	ET, LE	6 stages	Mean	2	2 × 6
Climate	SPI_14d, SPI_30d, SPI_90d	6 stages	Mean	3	3 × 6
	EDDI_14d, EDDI_30d, EDDI_90d	6 stages	Mean	3	3 × 6
	SPEI_14d, SPEI_30d, SPEI_90d	6 stages	Mean	3	3 × 6
	PDSI, Palmer_Z	6 stages	Mean	2	2 × 6
	Precipitation, Humidity, Pressure, Shortwave, Longwave	6 stages	Sum	5	5 × 6
	Temp (air)	6 stages	Σ (Temp – 10 °C)	1	1 × 6
	Drought	6 stages	PDSI value	1	1 × 6

Six stages*: planting, emerging, blooming, pod setting, leaf dropping and harvesting stages.

Table 5
Summary of static variables and processing.

Category	Variables	Encoding type	Number of variables
Irrigation	PIC	–	1 × 1
	Irrigation	Label encoding	1 × 1
Soil (0–25 cm)	$CaCO_3$, CEC, EC, pH, SAR, Max_OM	–	6 × 1
	PAWS, Drainage, Texture, Sand, Silt, Clay	–	6 × 1
	Soil type	Count encoding	1 × 1
	N_class, I_class	–	2 × 1
Location	Latitude, Longitude	–	2 × 1
	Position Code (12 states)	One-hot encoding	1 × 12

was used as the benchmark method because it is one of the simplest model forms for constructing relationships between soybean yield and multiple features (Gao et al., 2018; Montgomery et al., 2021). An RF is a tree-based model that ensembles multiple weak decision trees to generate a powerful regression model. An RF is simple to construct and train, has high computational efficiency, and is insensitive to outliers and is resistant to overfitting (Breiman, 2001). KNN uses the distances or similarities between samples to find neighbors and refers to the neighbor's yields to predict unknown yields (Keller et al., 1985). An ANN with three layers is used to fit the nonlinear relationships between the yield and the features (Wang, 2003). SVR uses a kernel function to map the nonlinear nonseparable low-dimensional feature space to the linear separable high-dimensional feature space and to construct a hyperplane in the high-dimensional space based on a support vector to achieve the best prediction of yield (Noble, 2006). With multiplicative gate units (Noble, 2006), LSTM has achieved major advances in time series prediction and is widely used in yield prediction (Klompenburg et al., 2020). A DNN is a fully connected layer with increased network depth that achieves stronger nonlinear representation (Liu et al., 2017). These model hyperparameters, including their structural and internal hyperparameters and optimization parameters, were partly optimized automatically by model tuning and partly determined empirically. Moreover, we used the average information gain of each decision tree splitting node of XGBoost to assess the importance levels of different features. A larger information gain indicated a more important role possessed by the corresponding feature in the model's training.

2.3.3. Accuracy assessment method

We adopted a "train–validate–test" assessment strategy to evaluate the performance. First, the soybean yield data in 2020 for all states were regarded as the test set. This set was not included in the slope calculation of the historical yield and the superparameter tuning process of the model, which ensured that the yield prediction was entirely consistent with the actual situation. Second, the yield data from 2003 to 2019 were randomly divided into training and validation sets at a ratio of 85% vs. 15%. Since the yield data were detrended and year-independent, this study did not divide the training and validation sets by year. This "train–validate–test" strategy ensured that the test set was wholly invisible and that the optimal parameters were obtained by adjusting the model according to the validation results. In addition, previous studies have shown that the same method can be used to obtain predicted results with different accuracies in different years due to fluctuations in environmental factors (Ma et al., 2021; Sakamoto, 2020). For example, the same model for the United States performed worse in 2012 than in other years due to drought (Kang et al., 2020; Sun et al., 2019). Therefore, the leave-one-year-out strategy has been widely used when assessing the model's stability in different years (Cai et al., 2019; Kim et al., 2019; Ma et al., 2021; Sakamoto et al., 2014; Schwalbert et al., 2020; Srivastava et al., 2022). The RMSE and R^2 were adopted to assess the accuracy.

2.3.4. Explanation of each prediction using SHAP

The feature importance analysis of XGBoost only expressed which features are essential during the training for the node partitioning of XGBoost but cannot reflect the change in input features on the prediction results. We adopted Shapley additive explanations (SHAP) (Lundberg and Lee, 2017) to quantify the contribution of features to the prediction results. SHAP is a framework proposed in 2017 that uses the game theoretic approach to explain the output of any machine learning model. The SHAP value is the difference between the sample prediction and the average prediction and directly represents each feature on the

predicted yield. This is an excellent way to strengthen the interpretation of predicted results based on ML.

3. Results

3.1. Soybean prediction result

Fig. 6 shows the trend of validation R^2 and test R^2 from planting to harvesting for the test year 2020. XGBoost performed better as the amount of incorporated dynamic information grew. The slope of R^2 growth varied in different periods. From early April to the end of August, the slope of R^2 showed a rapidly increasing trend, indicating a significant improvement in the model prediction performance. After the pod-setting stage, the validation R^2 and test R^2 were close to the maximum; both remained stable with only slight changes. The change in validation R^2 and test R^2 indicated that good performance could be obtained after the pod-setting stage, indicating that accurate in-season forecasts of soybean yield at the county level in the United States could be achieved as early as the end of pod-setting stage. The test RMSE and R^2 on August 30th were 0.263 t/ha and 0.79, respectively. The best test results occurred in early October, when the test RMSE and R^2 were 0.246 t/ha and 0.82, respectively.

We divided the counties into nine classes according to their statistical and predicted yields ($0 \leq$ yield < 0.5, $0.5 \leq$ yield < 1.0, $1.0 \leq$ yield < 1.5, $1.5 \leq$ yield < 2.0, $2.0 \leq$ yield < 2.5, $2.5 \leq$ yield < 3.0, $3.0 \leq$ yield < 3.5, $3.5 \leq$ yield < 4.0, and yield > 4.0). A clear spatial aggregation pattern of soybean yield was observed in the statistical data from NASS (Fig. 7a). North Dakota and Kansas were areas with low soybean yields. The areas with high soybean yields were centered in the Corn Belt (e.g., Minnesota, Iowa, and Illinois) and Nebraska. The spatial aggregation pattern of the soybean yield predictions based on XGBoost (Fig. 7b) was highly consistent with that of NASS. The spatial locations of high- and low-yield areas overlapped greatly with the statistical data of NASS. Furthermore, it was not affected by statistical outliers, e.g., in the only county where the NASS statistical yield had not yet exceeded 0.5 t/ha.

3.2. Model interpretation

Fig. 8 shows the top 20 dynamic features ranked in descending order from top to bottom according to the average information gain. Among all the dynamic features, the EVI was the most important feature for yield prediction because it had the highest information gain. Regarding the dynamic features derived from satellite images, including vegetation and moisture indices and raw surface reflectance bands, the vegetation indices (e.g., EVI, GCVI, and LSWI) contributed more to the yield prediction than individual surface reflectance bands. Among the weather features, atmospheric pressure, air temperature, and atmospheric humidity were the most important variables for yield prediction, but their contributions were smaller than those of the vegetation indices. Overall, the influence of the dynamic variables on soybean yield prediction exhibited vegetation indices as the dominant factors and climate variables as secondary factors.

Our study further explored the contribution differences among dynamic features such as the growth date and phenology and found that vegetation variables between mid-July and late August contributed more to soybean yield prediction than the other periods. Mid-July to the end of August is the pod-setting period for soybeans, and the importance levels of dynamic features indicate that pods and leaves largely determine soybean yield. The contributions of the EVI, GCVI, SAVI, and near-infrared (NIR) reflectance also indicate that the pod-setting period of soybeans is the most important period for yield prediction. In contrast, the importance of the climatic features was evenly distributed throughout the soybean growing period. For example, the contribution peaks of atmospheric pressure occurred in mid-May and mid-September, and this feature had little impact on soybean yield in other periods. Although the overall contribution of atmospheric humidity ranked highly, the peak of its contribution did not occur throughout the growth period. Weather events had an enormous impact on soybean yield, but the occurrence of abnormal weather was uncertain, resulting in randomness in the contributions of meteorological features over time.

Furthermore, the 20 most important dynamic variables are shown in Fig. 9a and are mainly dominated by the vegetation indices before and

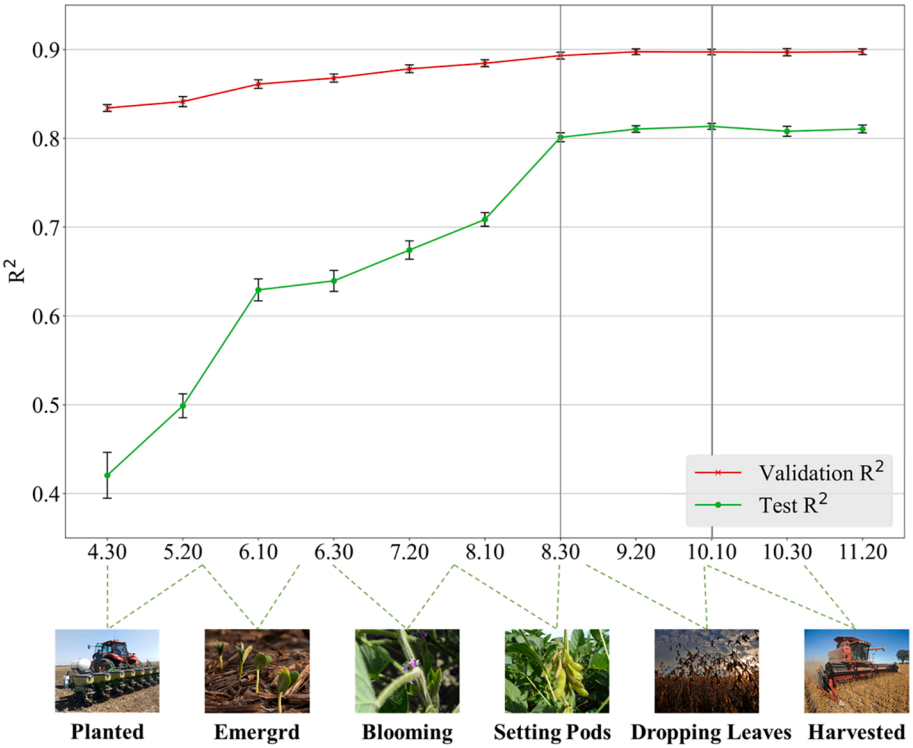


Fig. 6. Trends in validation R^2 and test R^2 from the planting to harvesting stages of soybean.

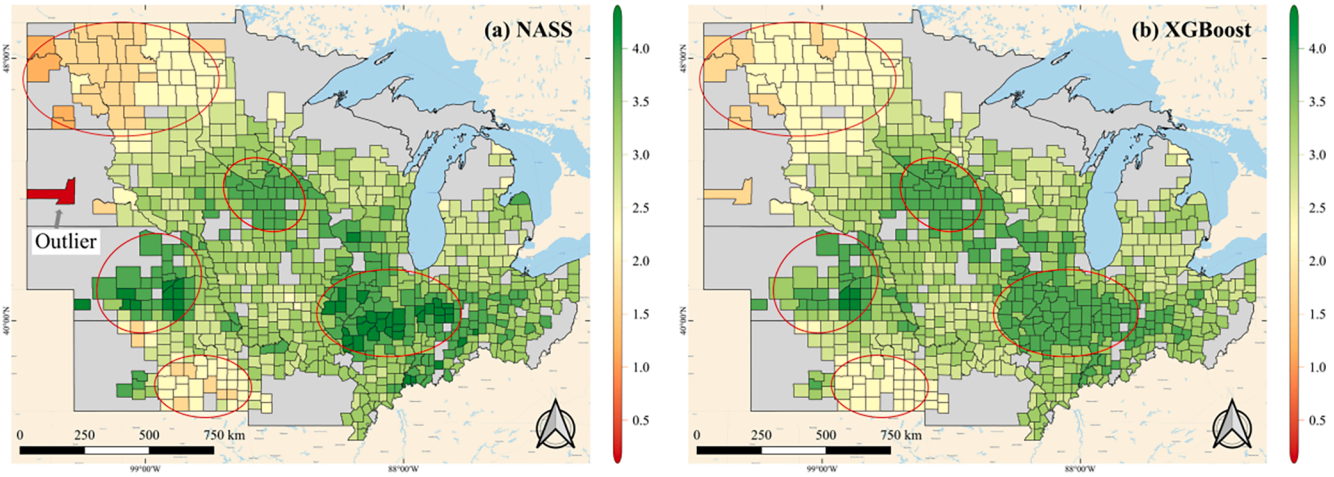


Fig. 7. (a) NASS soybean yield (t/ha) in 2020 and (b) predicted soybean yield (t/ha) in 2020.

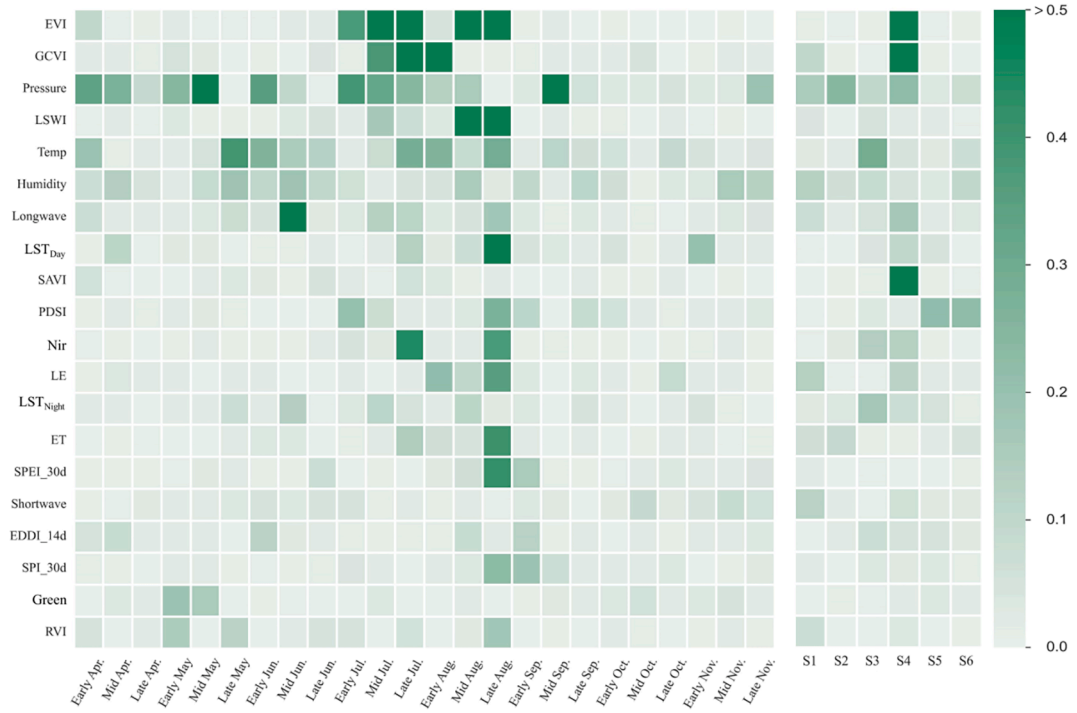


Fig. 8. Average information gains of the top 20 features (S1–S6 indicate the emerging, blooming, pod setting, leaf dropping and harvesting stages in order).

after pod-setting, such as the EVI in pod-setting with an information gain of 4.10 bits. In contrast, the climate variables contributed less to soybean prediction than the vegetation indices, e.g., the highest daytime surface temperature in late August (LST_{Day}_Late Aug.) contributed only 0.68 bits. The model successfully learned the spatial distribution patterns of yields by their positions and longitudes. The top 20 static variables for soybean yield prediction are shown in Fig. 9b. Location and management features were more important for soybean yield prediction than soil features. The results indicated that soybean yield was more influenced by geographic location and that longitude was the most important location variable. Yields were closely tied to longitude, with higher soybean yields in the east and lower soybean yields in the west. The model also successfully captured the effect of irrigation on county soybean yield, with information gains of>0.25 for the PIC and irrigation intensity. The remaining soil variables, such as calcium carbonate ($CaCO_3$) content, soil type (Texture), and maximum organic matter content (Max_OM), had little effect on yield prediction (Fig. 9c).

When comparing the dynamic and static variables, we found that 17 of the top 20 were dynamic variables, and 3 were static variables (ranking 2nd, 15th, and 17th), indicating that dynamic variables contributed more to yield prediction than static variables. Nine of the top 10 most important variables were vegetation indices or moisture variables, indicating the dominance of vegetation indices and moisture features in soybean yield prediction. As the most important feature, the EVI at pod-setting had an average information gain of 4.1 bits, but it did not dominate absolutely. Thus, U.S. soybean yield prediction does not depend on the few highest-ranked variables but depends on many variables.

Fig. 10a shows the average SHAP value of the top 20 dynamic and static variables when the trained model makes a specific prediction. Similarly, a higher SHAP value indicates that the feature is more important for the yield prediction results. Fig. 10a shows that the most important feature in soybean yield prediction is still the EVI at pod setting, when soybean predictions reach stability and accuracy (Fig. 6).

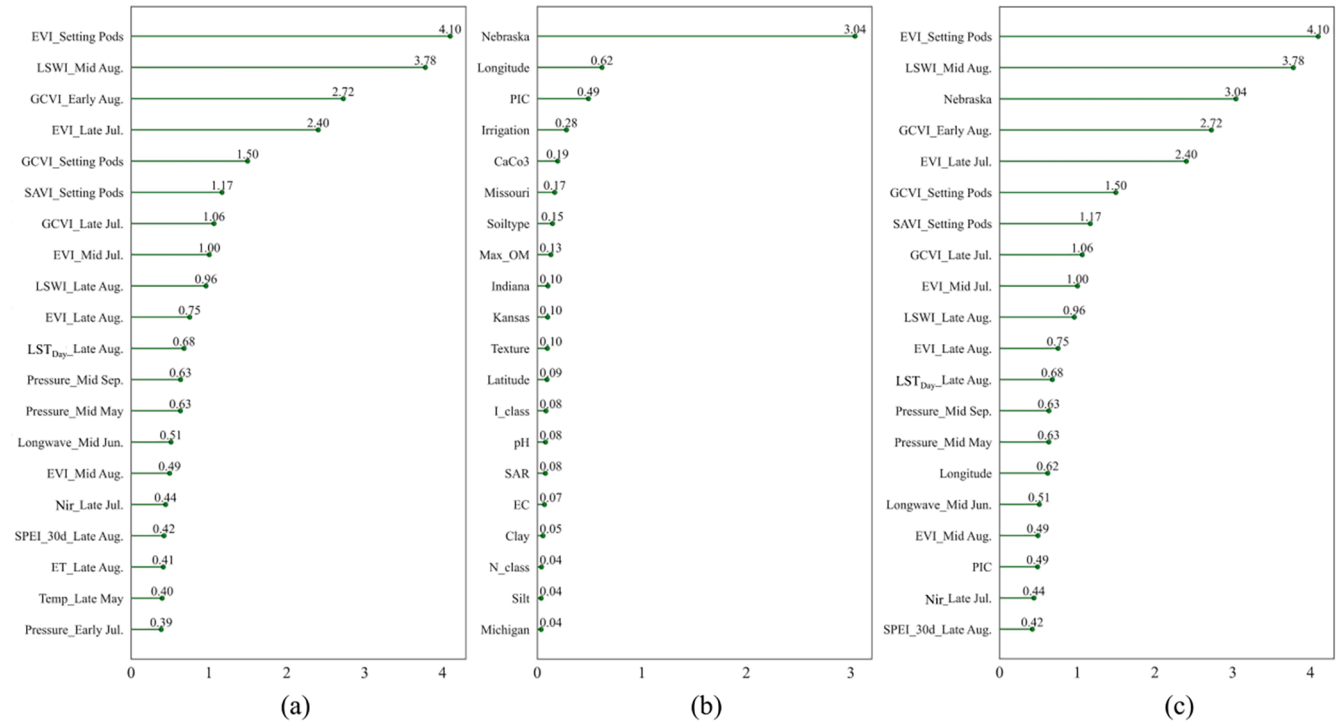


Fig. 9. Average information gains: (a) top 20 dynamic variables, (b) top 20 static variables, and (c) top 20 variables among all dynamic and static variables.

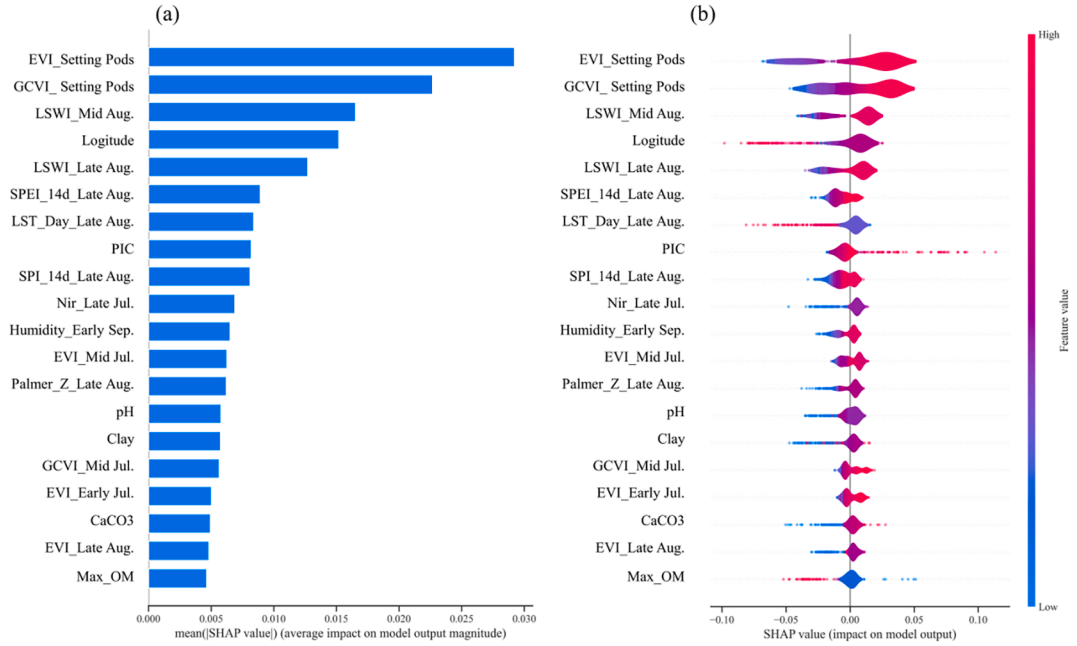


Fig. 10. Average SHAP values for the top twenty features at yield prediction.

This result indicates that the knowledge learned during model training is used well for the prediction, which is the basis for the R^2 of 0.82 achieved by the model on the test set. The violin plot of the SHAP values represents the impact of the top 20 features on the yield prediction (Fig. 10b), where each point is a county's yield prediction in 2020, and we can see how much each feature contributes to the forecasted yield. Here, the EVI at pod setting with high values positively impacts the predicted yield. The same trend was observed in the GCVI at pod setting, LSWI at mid-August, and PIC. In contrast, higher longitude and higher daytime surface temperature will decrease the yield prediction, which is in line with the findings of spatial differences in mean yield (Fig. 1) and temperature stress (Patel et al., 2012).

Fig. 11 shows the correlations among the top 20 variables. The high correlations between the vegetation and water indices with high information gains indicate the important roles of vegetation and water indices in soybean yield prediction. The correlation coefficients between the daytime surface temperature (LST_{Day}_Late Aug) and all vegetation indices in late August were negative. Its correlation coefficient with soybean yield was −0.48, implying that higher daytime surface temperature is associated with lower soybean yield.

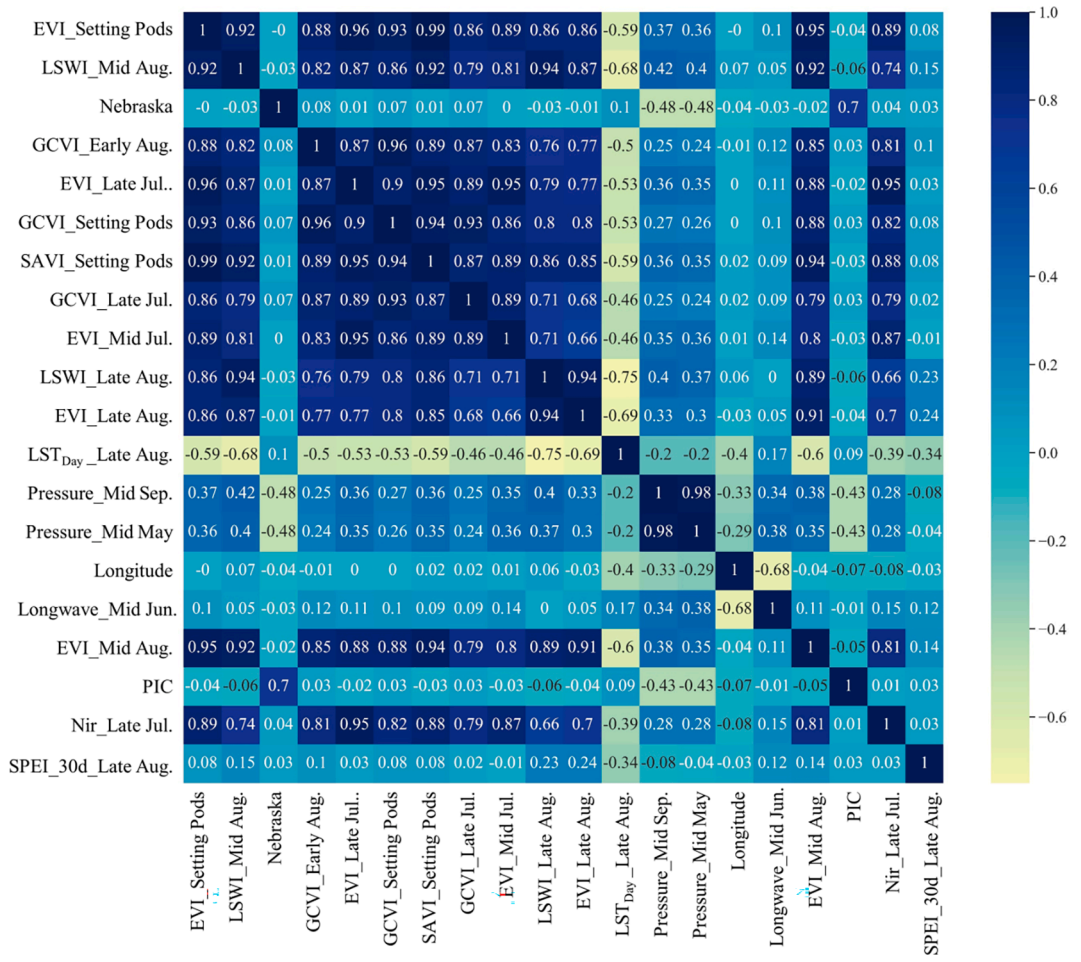


Fig. 11. Correlation coefficient matrix among the top 20 variables.

Based on XGBoost's variable importance ranking, we started with the most important variable and added the first of the remaining variables one by one to explore the effect of the number of variables on the model. The curves of the validation and testing RMSEs of XGBoost as the number of variables increased are shown in Fig. 12. The trend of RMSE shows that the RMSEs for validation and testing reach a relatively stable value when the number of variables reaches 200. This suggests that the soybean yield prediction model designed in our study requires at least 200 of the most important variables to produce relatively accurate and stable results. According to the R^2 and RMSE metrics, XGBoost predicted the best results at 250 variables, corresponding to a validation RMSE and validation R^2 of 0.193 t/ha and 0.90, respectively, while the test RMSE and test R^2 were 0.246 t/ha and 0.82, respectively.

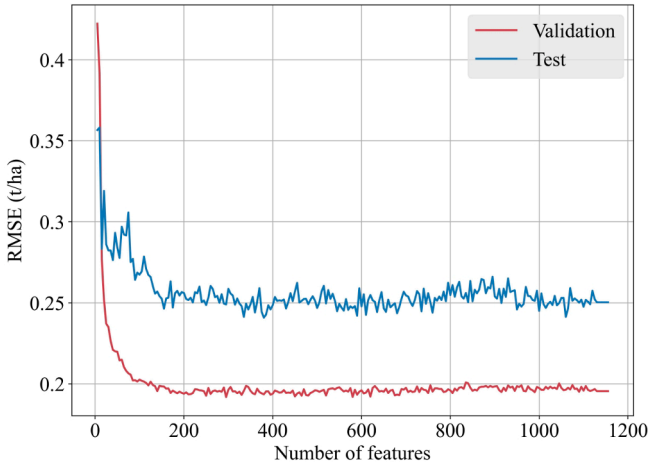


Fig. 12. Trends of the validation and test RMSE metrics with increasing number of variables.

3.3. Performance of XGBoost compared with that of other models

The performance differences among the six machine learning algorithms regarding county-level soybean yield prediction in the midwestern U.S. are shown in Table 6. The training and validation data were split by setting three random seeds, 9, 99, and 999, and the results were averaged. Here, the training metrics indicate how well each model fits

Table 6
Performance comparison for the LR, RF, KNN, ANN, SVR, LSTM, DNN, and XGBoost models.

Model	Training time (s)*	Training		Validation		Test	
		RMSE (t/ha)	R^2	RMSE (t/ha)	R^2	RMSE (t/ha)	R^2
LR	0.35	0.212	0.88	0.245	0.85	0.385	0.61
RF	68.76	0.088	0.98	0.243	0.85	0.313	0.74
KNN	0.04	0.000	1.00	0.228	0.86	0.398	0.59
ANN	11.79	0.146	0.94	0.219	0.87	0.374	0.64
SVR	49.32	0.114	0.97	0.192	0.90	0.339	0.70
LSTM	172.35	0.202	0.91	0.258	0.84	0.292	0.78
DNN	125.02	0.193	0.91	0.224	0.88	0.304	0.76
XGBoost	57.34	0.037	1.00	0.197	0.90	**0.246**	**0.82**

* Runtime on a computer with 8 CPU cores and 64 GB of RAM.

the known yield data, the validation metrics are the model's full yield data prediction performance, and the test metrics are the model's target year prediction performance.

The training times varied significantly for different algorithms. With the same running environment and input matrix (10418 × 1129), the training time of the RF model was 68.76 s, while the training times of the LR, KNN, ANN, SVR, and XGBoost models were <60 s. We trained the LSTM and DNN models under the same conditions; their training times were 172.35 s and 125.02 s, respectively, significantly exceeding the time consumed by the general machine learning algorithms. The training R^2 values of all algorithms exceeded 0.9, and the validation R^2 values of all algorithms were above 0.84. However, the performance changed significantly during testing. LR had the poorest performance, with an R^2 of 0.61 and an RMSE of 0.385 t/ha, while XGBoost had the best performance, with an R^2 of 0.82 and an RMSE of 0.246 t/ha. Despite being the most widely used machine learning model for yield prediction, RF achieved a performance that was not as good as that of XGBoost; its test R^2 of 0.08 was lower than that of XGBoost, and the test RMSE of 0.059 t/ha was higher than that of XGBoost. SVR achieved the best performance in the validation phase; however, it performed worse than XGBoost in the test phase, where the test R^2 was 0.12 lower than that of XGBoost and the test RMSE was 0.058 t/ha higher than that of XGBoost. In addition, the widely used LSTM deep learning model, considered promising, had a lower performance than XGBoost, with a test R^2 and an RMSE of 0.78 and 0.292 t/ha, respectively. In conclusion, XGBoost outperformed the other algorithms, with better prediction and generalization ability, combined with multidimensional feature engineering.

3.4. Stability assessment of XGBoost

To further explore the stability of the constructed model, we used the yield data for one year between 2003 and 2020 as the test set, and the remaining yield data for other years were divided into training and validation sets (0.85:0.15). The model hyperparameters were consistent with the yield prediction in 2020. Three different pseudorandom number seeds were fixed to ensure the reproducibility of the trials and to avoid the uncertainty associated with random separation during training and validation. The obtained performance statistics are summarized in Tables 7 and 8; in this instance, the R^2 and RMSE were obtained from the means of three trials ± standard deviation (STD). Our results show that XGBoost outperformed the other models in all years except 2008, 2012, and 2016. SVR outperformed XGBoost in 2008 and 2012, and LSTM only outperformed XGBoost in 2016.

3.4.1. Xgboost vs. RF

Many previous studies have confirmed that RF remains a simple and effective algorithm with relatively satisfactory yield prediction accuracy due to its bagging ensemble-based strategy (Breiman, 2001; Sakamoto, 2020). However, in the average results of our three trials, XGBoost attained higher accuracy than RF on the test set for all years (Table 7). RF and XGBoost are ensemble models based on decision trees. XGBoost differs from RF mainly in its gradient tree boosting concept and efficient computation process (Chen and Guestrin, 2016). XGBoost has been validated in many fields, performing better than RF (Oughali et al., 2019; Zamani Joharestani et al., 2019; Zhang et al., 2021). XGBoost is currently the first choice of algorithm for most practitioners and data science competitions (Grinsztajn et al., 2022). However, only a few studies have used XGBoost to predict crop yield (Herrero-Huerta et al., 2020; Kang et al., 2020; Obsie et al., 2020). Several papers that have applied XGBoost in yield prediction scenarios also support our conclusions (Kang et al., 2020; Shahhosseini et al., 2019).

3.4.2. Xgboost vs. SVR and LSTM

In our study, XGBoost, SVR, and LSTM outperformed each other in different years. Therefore, we further compared the performance differences among the SVR, LSTM, and XGBoost models in 2008, 2012, 2015, 2016, and 2020. The five years were representative, among which 2008 and 2012 were low-yield years; 2016 was a high-yield year; 2015 was the best year for yield based on the predictions of SVR, LSTM, and XGBoost; and 2020 was the year used to match the true yield prediction. Scatter plots (Fig. 13) and absolute error maps (Fig. 14) were adopted to illustrate the differences among the four models.

(1) XGBoost vs. SVR

In the 2008 and 2012 tests, SVR outperformed XGBoost, attributed mainly to SVR sacrificing prediction accuracy in other years for high accuracy in low-yield years. In regression, the support vectors are the samples farthest from the hyperplane, and the aim is to minimize the distance between the support vectors and the hyperplane. In this study, the average soybean yield was 2.75 t/ha, the minimum yield was 0.03 t/ha, and the maximum yield was 4.30 t/ha. The distance between the minimum yield and the average yield was greater than that between the maximum yield and the average yield. Thus, the hyperplane of the SVR was closer to the low-yield sample and ignored the other yields that made up the majority of the samples. These operating principles of the SVR regressor make it predict yields better in low-yield years but worse in normal- and high-yield years. This conclusion is also confirmed by Figs. 13 b1–d1 and 14 b1–d1. In the low-yield year of 2012, the scatterplot of SVR was more concentrated than the scatterplots of the other three models, while in the high-yield years of 2015 and 2016, the scatterplot of SVR was skewed toward the lower delta. Unlike this

Table 7
Test R^2 for LR, RF, KNN, ANN, SVR, LSTM, DNN and XGBoost between 2003 and 2020.

Year	LR	RF	KNN	ANN	SVR	LSTM	DNN	XGBoost
2003	−0.26 ± 0.20	0.36 ± 0.02	0.03 ± 0.01	0.52 ± 0.06	0.46 ± 0.10	0.37 ± 0.08	0.27 ± 0.10	**0.52** ± 0.02
2004	0.45 ± 0.05	0.67 ± 0.01	0.63 ± 0.01	0.50 ± 0.06	0.72 ± 0.01	0.57 ± 0.03	0.59 ± 0.04	**0.73** ± 0.01
2005	−0.16 ± 0.06	0.66 ± 0.01	0.61 ± 0.01	0.25 ± 0.07	0.70 ± 0.02	0.55 ± 0.06	0.56 ± 0.04	**0.74** ± 0.00
2006	0.47 ± 0.06	0.79 ± 0.01	0.71 ± 0.01	0.73 ± 0.01	0.79 ± 0.02	0.72 ± 0.03	0.78 ± 0.03	**0.82** ± 0.01
2007	0.56 ± 0.01	0.66 ± 0.01	0.54 ± 0.02	0.59 ± 0.07	0.73 ± 0.01	0.65 ± 0.04	0.64 ± 0.10	**0.75** ± 0.02
2008	0.47 ± 0.05	0.64 ± 0.01	0.35 ± 0.01	0.41 ± 0.10	**0.74** ± 0.01	0.57 ± 0.06	0.68 ± 0.09	0.73 ± 0.01
2009	0.48 ± 0.24	0.70 ± 0.01	0.57 ± 0.02	0.57 ± 0.03	0.71 ± 0.00	0.68 ± 0.02	0.69 ± 0.02	**0.74** ± 0.01
2010	0.62 ± 0.03	0.74 ± 0.01	0.53 ± 0.01	0.44 ± 0.03	0.72 ± 0.01	0.69 ± 0.03	0.61 ± 0.06	**0.78** ± 0.01
2011	0.61 ± 0.02	0.72 ± 0.01	0.65 ± 0.00	0.69 ± 0.03	0.76 ± 0.01	0.76 ± 0.03	0.74 ± 0.02	**0.82** ± 0.01
2012	0.19 ± 0.10	0.71 ± 0.01	0.66 ± 0.01	0.66 ± 0.06	**0.73** ± 0.02	0.62 ± 0.12	0.69 ± 0.06	0.68 ± 0.04
2013	0.63 ± 0.03	0.70 ± 0.01	0.60 ± 0.02	0.35 ± 0.12	0.69 ± 0.02	0.64 ± 0.03	0.68 ± 0.04	**0.74** ± 0.01
2014	0.56 ± 0.20	0.76 ± 0.00	0.64 ± 0.01	0.63 ± 0.04	0.75 ± 0.02	0.71 ± 0.03	0.57 ± 0.16	**0.79** ± 0.02
2015	0.34 ± 0.07	0.80 ± 0.01	0.76 ± 0.01	0.68 ± 0.04	0.79 ± 0.01	0.74 ± 0.04	0.77 ± 0.04	**0.84** ± 0.01
2016	0.50 ± 0.07	0.60 ± 0.01	0.52 ± 0.02	0.62 ± 0.06	0.63 ± 0.01	**0.72** ± 0.01	0.63 ± 0.01	0.69 ± 0.02
2017	0.72 ± 0.01	0.76 ± 0.01	0.63 ± 0.02	0.68 ± 0.05	0.78 ± 0.01	0.78 ± 0.00	0.79 ± 0.02	**0.80** ± 0.01
2018	0.70 ± 0.04	0.74 ± 0.00	0.60 ± 0.00	0.66 ± 0.04	0.66 ± 0.03	0.73 ± 0.03	0.73 ± 0.09	**0.79** ± 0.01
2019	0.30 ± 0.15	0.69 ± 0.01	0.65 ± 0.01	0.58 ± 0.02	0.70 ± 0.04	0.62 ± 0.03	0.56 ± 0.16	**0.74** ± 0.01
2020	0.61 ± 0.03	0.74 ± 0.01	0.59 ± 0.00	0.64 ± 0.06	0.70 ± 0.02	0.78 ± 0.03	0.76 ± 0.02	**0.82** ± 0.01

Table 8
Test RMSE (t/ha) for the LR, RF, KNN, ANN, SVR, LSTM, DNN and XGBoost models between 2003 and 2020.

Year	LR	RF	KNN	ANN	SVR	LSTM	DNN	XGBoost
2003	0.627 ± 0.050	0.449 ± 0.009	0.552 ± 0.003	0.390 ± 0.024	0.413 ± 0.039	0.445 ± 0.030	0.479 ± 0.033	**0.389** ± 0.006
2004	0.530 ± 0.026	0.412 ± 0.006	0.434 ± 0.004	0.503 ± 0.032	0.376 ± 0.008	0.471 ± 0.015	0.457 ± 0.020	**0.369** ± 0.005
2005	0.641 ± 0.017	0.350 ± 0.002	0.374 ± 0.005	0.516 ± 0.024	0.329 ± 0.007	0.398 ± 0.029	0.395 ± 0.017	**0.304** ± 0.002
2006	0.469 ± 0.025	0.299 ± 0.001	0.349 ± 0.008	0.335 ± 0.006	0.294 ± 0.012	0.342 ± 0.020	0.300 ± 0.024	**0.273** ± 0.006
2007	0.409 ± 0.006	0.363 ± 0.004	0.420 ± 0.007	0.397 ± 0.031	0.319 ± 0.001	0.369 ± 0.021	0.371 ± 0.049	**0.310** ± 0.012
2008	0.376 ± 0.020	0.311 ± 0.004	0.419 ± 0.004	0.397 ± 0.033	**0.265** ± 0.003	0.336 ± 0.021	0.289 ± 0.039	0.272 ± 0.004
2009	0.354 ± 0.081	0.276 ± 0.004	0.327 ± 0.006	0.327 ± 0.012	0.270 ± 0.001	0.285 ± 0.010	0.279 ± 0.010	**0.254** ± 0.003
2010	0.327 ± 0.013	0.270 ± 0.001	0.364 ± 0.002	0.398 ± 0.009	0.282 ± 0.005	0.298 ± 0.014	0.332 ± 0.025	**0.248** ± 0.006
2011	0.440 ± 0.012	0.372 ± 0.006	0.419 ± 0.002	0.395 ± 0.019	0.343 ± 0.005	0.349 ± 0.021	0.365 ± 0.013	**0.297** ± 0.002
2012	0.635 ± 0.037	0.379 ± 0.003	0.411 ± 0.004	0.408 ± 0.032	**0.367** ± 0.010	0.429 ± 0.067	0.391 ± 0.040	0.398 ± 0.026
2013	0.378 ± 0.013	0.343 ± 0.003	0.394 ± 0.008	0.499 ± 0.046	0.344 ± 0.012	0.374 ± 0.015	0.349 ± 0.022	**0.314** ± 0.005
2014	0.377 ± 0.085	0.283 ± 0.003	0.346 ± 0.003	0.353 ± 0.018	0.292 ± 0.007	0.309 ± 0.014	0.377 ± 0.071	**0.263** ± 0.008
2015	0.479 ± 0.026	0.259 ± 0.004	0.291 ± 0.002	0.333 ± 0.022	0.275 ± 0.003	0.299 ± 0.021	0.280 ± 0.021	**0.240** ± 0.006
2016	0.358 ± 0.024	0.322 ± 0.003	0.348 ± 0.004	0.311 ± 0.024	0.308 ± 0.001	**0.266** ± 0.004	0.307 ± 0.005	0.281 ± 0.011
2017	0.330 ± 0.006	0.306 ± 0.006	0.378 ± 0.010	0.351 ± 0.031	0.293 ± 0.005	0.292 ± 0.003	0.285 ± 0.009	**0.283** ± 0.008
2018	0.363 ± 0.021	0.335 ± 0.002	0.416 ± 0.002	0.385 ± 0.025	0.381 ± 0.019	0.343 ± 0.016	0.339 ± 0.058	**0.302** ± 0.007
2019	0.456 ± 0.049	0.303 ± 0.002	0.324 ± 0.002	0.357 ± 0.007	0.299 ± 0.018	0.340 ± 0.011	0.358 ± 0.064	**0.281** ± 0.007
2020	0.385 ± 0.013	0.313 ± 0.002	0.398 ± 0.001	0.374 ± 0.027	0.339 ± 0.013	0.292 ± 0.018	0.304 ± 0.012	**0.268** ± 0.009

"unfairness" in SVR, XGBoost worked well to ensure prediction accuracy for most years, including years with abnormal yields.

(2) XGBoost vs. LSTM

In recent years, LSTM has been widely used in yield prediction, and many studies have concluded that deep learning is better than machine learning in this field (Jiang et al., 2020; Kim et al., 2019; Srivastava et al., 2022; Sun et al., 2020). Our study found that XGBoost was better than LSTM except in the particularly high-yield year of 2016. LSTM tended to consume more training time due to its deeper structure, which led to slower and more difficult backward propagation (BP) of the derivatives. LSTM tends to extract features automatically in an end-to-end manner, and the lack of feature engineering and other preprocessing steps can significantly affect its prediction performance. LSTM produced a higher standard deviation, implying a high level of uncertainty in yield prediction with a higher standard deviation (Tables 7 and 8), and the good results yielded by LSTM were based more on chance. This uncertainty stems from the unique training process of LSTM. Backpropagation and stochastic gradient descent (SGD) made it possible for deep learning models to stop at different local optima. While they may have been very close to the global optimum on the validation set, they exhibited large fluctuations on the test set. For example, the average R^2 of the three LSTM results for 2012 was 0.61, while the standard deviation of the R^2 of LSTM for 2012 reached 0.12. The performance of LSTM relies on the input sample size and the time series length. LSTM was effective in the long-time series information extraction problem, while the sample size for county-wide yield prediction was still small for LSTM. The most important variable was the vegetation index. However, the vegetation index itself was the final indication of the cumulative plant growth process at a given moment, so the advantage of the cumulative temporal effect of LSTM disappeared. Yield data are more heterogeneous tabular data than image or verbal data. This is because each yield sample produces dense numerical and sparse categorical features after completing feature engineering.

While few studies have compared XGBoost with deep learning models in crop yield prediction, XGBoost currently outperforms many deep learning algorithms in other fields according to comparisons on real datasets of different sizes and with different learning objectives (Borisov et al., 2021). XGBoost combined with detailed feature engineering outperformed various advanced deep learning models on the same yield dataset. Kang et al. (2020) similarly concluded that the XGBoost algorithm outperforms deep learning models such as LSTM and CNNs in terms of accuracy and stability.

4. Discussion

The framework that combined XGBoost, multidimensional feature engineering, and yield detrending successfully predicted soybean yield at the county scale in the midwestern United States. Spatially, the predicted yield distributions were very similar to those derived from NASS statistics. Temporally, we could accurately predict yields at the end of the soybean pod-setting period in late August. Our framework outperformed LR, RF, KNN, ANN, SVR, LSTM, and DNN on the same soybean yield dataset.

4.1. Comparison with previous studies

We reviewed and compared the differences between our prediction results and the previous U.S. soybean yield forecasts. To ensure comparability, we developed strict filtering criteria:

i. The studies all concerned soybeans;
ii. The study regions and scales were comparable, all of which focused on county-scale yield forecasting rather than state-scale forecasting;
iii. The historical data used in the studies were all USDA historical yield data;
iv. The model evaluation metrics included at least one of the RMSE and R^2 measures.

The above criteria ensured consistency in the sources and distributions of the yield data and model evaluations. The performance differences between our method and other approaches are listed in Table 9.

Our method achieved the best performance compared to the other approaches. Note that the test results of Sun et al. (2019) produced an R^2 of 0.78 for five years, but the average R^2 over five years was calculated to be 0.74. Sakamoto (2020) studied only five states and constructed an RF–PR (polynomial regression) model for each state. Fan et al. (2022) produced yield predictions for 41 states. Our study achieved the lowest RMSE (0.246 t/ha) and the highest R^2 (0.82) in the test year, regardless of whether the best results of other studies or their average results were compared and regardless of the size of the study area. Most significantly, our newly proposed framework provides higher accuracy than that obtained in previous similar studies. Moreover, we quantified the importance of features in model training and how changes in features affect yield prediction using the feature importance of XGBoost and the SHAP method, which provide more in-depth model interpretation. Based on our approach, accurate predictions can be achieved at the end of the soybean pod-setting stage, which is the most critical period for yield prediction.

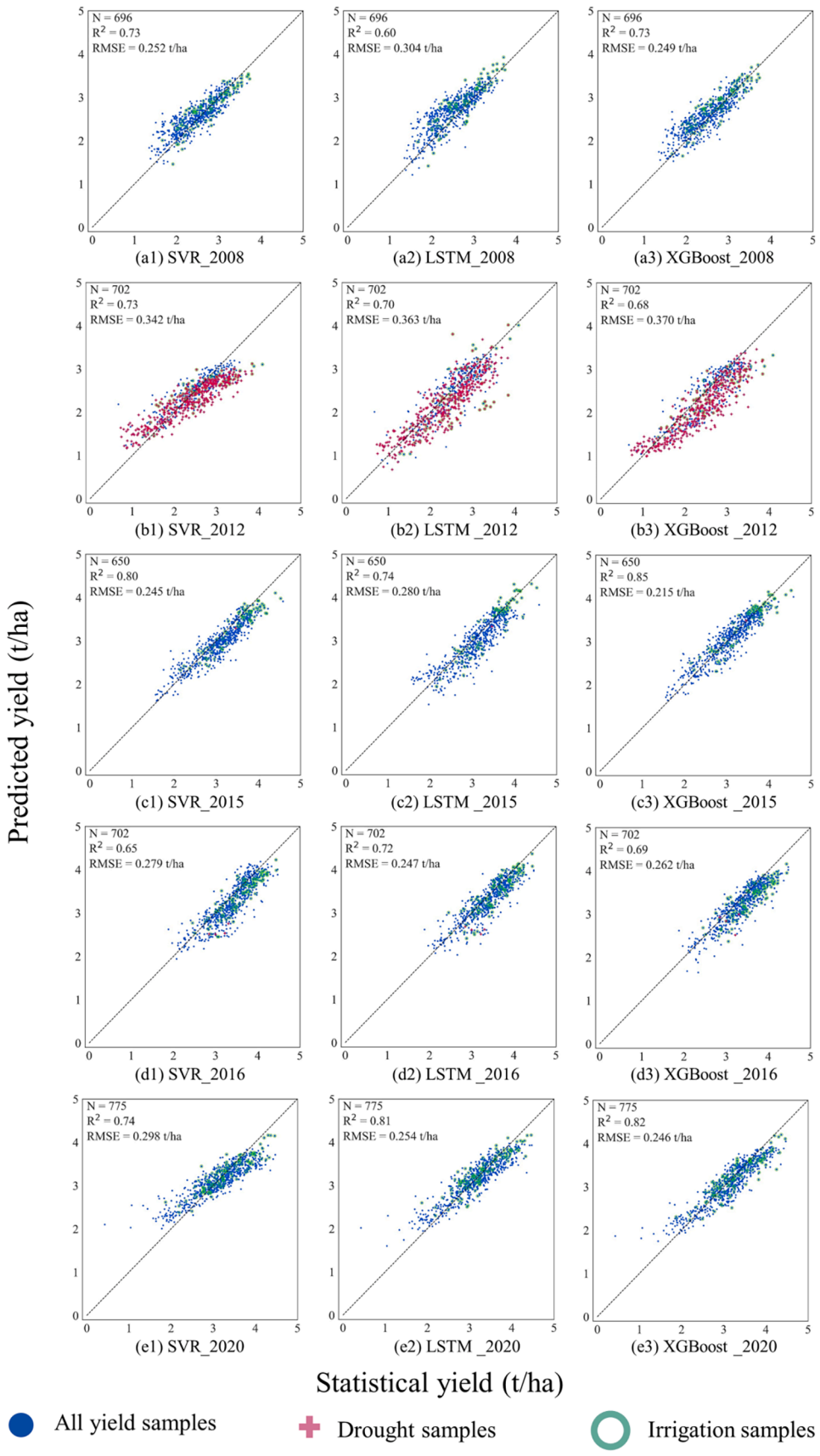


Fig. 13. Scatter plots of the statistical yield vs. the yields predicted by the (a) SVR, (b) LSTM, and (c) XGBoost models in 2008, 2012, 2015, 2016, and 2020.

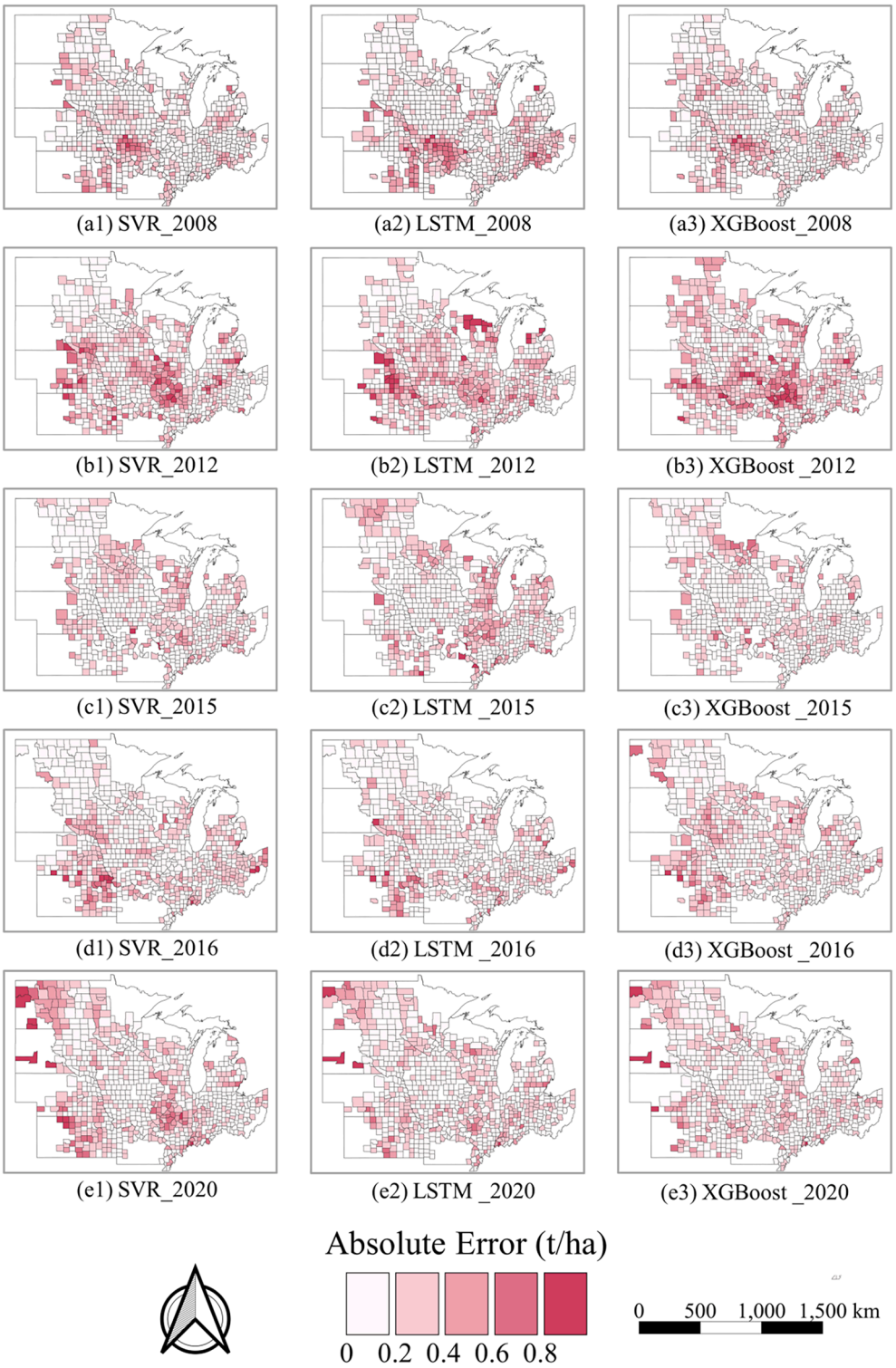


Fig. 14. Absolute error of the (a) SVR, (b) LSTM, and (c) XGBoost models in 2008, 2012, 2015, 2016, and 2020.

4.2. Sample filtering and validation assessment

Machine learning and deep learning algorithms depend on samples. Yield sample processing directly affects the performance and generalization abilities of models. Excessive yield sample quality control would lead to abnormally good but unreliable yield prediction results. Unfortunately, many studies tend to overfilter yield samples and only keep the samples that are within the "reasonable" range for model training and validation (Cao et al., 2021a; Chen et al., 2018; Kouadio et al., 2014), such as keeping the samples that fall within the interval of (average yield - N × STD, average yield + N × STD). This approach possesses two weaknesses. First, this data filtering strategy assumes that yield data conform to a normal distribution, but this assumption is unsuitable for yield data. Second, the coefficient N depends on experience and the importance of "outliers." In fact, yield is affected by many factors, such as extreme weather events, and pests lead to low yield. Therefore, excluding yield samples beyond the preset range is unreasonable and could significantly reduce the predictive performance of the utilized

Table 9
Comparison of the results derived from the latest U.S. county-level soybean yield prediction studies.

No.	RMSE	R^2	Model	Area	Time span	Model strategy	Test year	Source
1	0.348	–	CNN–GP	11 states	2003–2015	Single for all	2011, 2012, 2013, 2014, 2015	(You et al., 2017)
2	0.330	0.74	CNN–LSTM	15 states	2003–2015	Single for all	2011, 2012, 2013, 2014, 2015	(Sun et al., 2019)
3	0.285		DNN	5 states	2006–2015	Single for all	One-year-out	(Kim et al., 2019)
4	0.316	–	RF–PR	**5 states**	2000–2018	Single per state	2018	(Sakamoto, 2020)
5	0.280	–	CNN–RNN	13 states	1980–2018	Single for all	2016, 2017, 2018	(Khaki et al., 2019)
6	0.475	0.73	GNN–RNN	**41 states**	1981–2019	Single for all	2018, 2019	(Fan et al., 2022)
Ours	**0.246**	**0.82**	XGBoost	12 states	2003–2020	Single for all	2020	–

model for target years with abnormal agroclimatic or agronomic conditions. In this study, we used the public yield samples of NASS, and we know little about data collection approaches; therefore, we kept all yield samples to avoid the occurrence of overfiltering. Some good-sounding results are attributed to incorrect validation strategies. For example, some studies divide all samples into training and validation groups using a random method (Gao et al., 2018; Han et al., 2020) and then use the validation groups to assess the resulting prediction performance. This method has a weak generalization capability and cannot be applied to predict the yield of a target year. Our experiment indicated that the validation R^2 based on this method was spuriously 8% to 21% higher than the leave-one-year-out test for a particular year (Table 6). In this study, we used a train–validate–test approach to objectively evaluate the performance of our designed framework for soybean yield prediction.

4.3. Detrending vs. No detrending

Yield trends are essential for predictions involving long-time series of yield data (Wu et al., 2007). A clear increasing trend was observed for the soybean yield data, as shown in Fig. 2. This significant linear increase is attributed to technological improvements, crop genotypes, or better agricultural management. However, it is difficult to directly collect these data because they are difficult to measure. To overcome the shortcomings of missing data, many statisticians have used detrending methods when conducting yield forecasting (Hansen et al., 2004; Lu et al., 2017; Wang et al., 2020b), but detrending seems to have been neglected in nascent machine learning methods. To reveal whether detrending improves the effect of yield prediction, we used the same detrending method to predict soybean yield at the county level. The R^2 obtained on the test set without detrending decreased from 0.82 to 0.58, while the RMSE increased from 0.246 t/ha to 0.374 t/ha (Fig. 15). Due to the linear growth trend of soybean yield data, the model without detrending yielded data that greatly underestimated the actual yield of soybeans. This is because, in the absence of detrending, the model did not learn about the trend of increasing yields from year to year and instead relied on historical yield levels from previous years to predict yields for the coming year; thus, the test resulted in an overall underprediction.

4.4. Considering phenology vs. without phenology in feature engineering

Feature engineering is crucial for machine learning. The performance of most machine learning methods depends heavily on their representations of feature vectors (Bengio et al., 2013). Feature engineering has been used extensively in yield prediction, and one of the most significant contributions to yield prediction is the creation and application of vegetation indices (Panda et al., 2010; Teal et al., 2006; Zhou et al., 2017). In our study, in addition to the extensive use of vegetation indices and water indices, we established phenology-based dynamic features for soybean yield prediction, which have been neglected by many studies. Regarding yield prediction at a large scale, crop phenology varies significantly from county to county, and different values of the same index obtained at the same time cannot reflect the quality levels of crop situations in different counties. The presented phenology-based dynamic features successfully solved this problem by reorganizing the vegetation and water stress indices according to the development of phenology so that the features could reflect the crop at different growth stages. In addition to dynamic features based on phenology, we applied agricultural knowledge to create new features, such as cumulative precipitation, physiological cumulative temperature, temperature difference, cumulative radiation, irrigation category, and drought category, to help XGBoost better understand stress factors for predicting crop yields. Fig. 16 shows the difference in soybean yield prediction with or without including phenology-based dynamic features. When including the phenology-based features, R^2 increased by 0.03 and RMSE decreased by 0.024 t/ha compared to those obtained

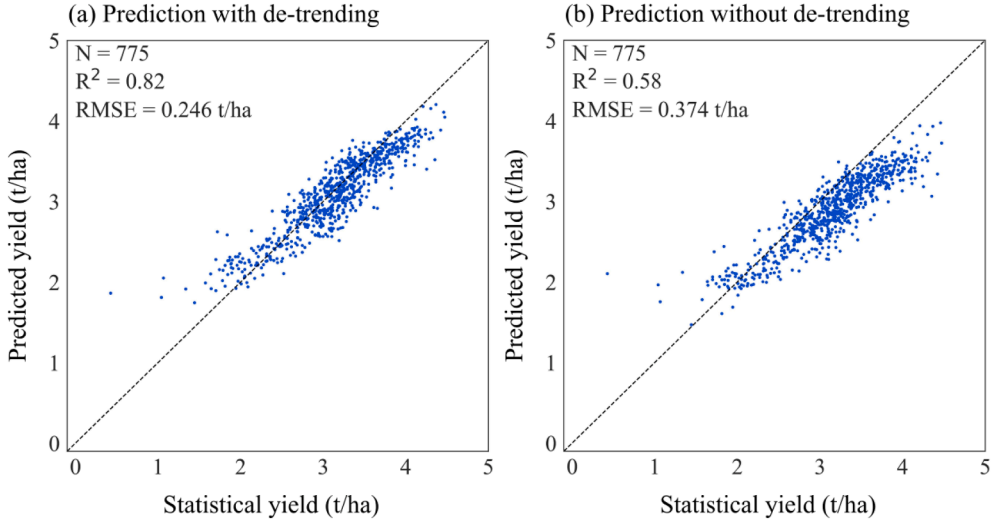


Fig. 15. Comparison of soybean yield prediction performances with (a) detrending and (b) no detrending.

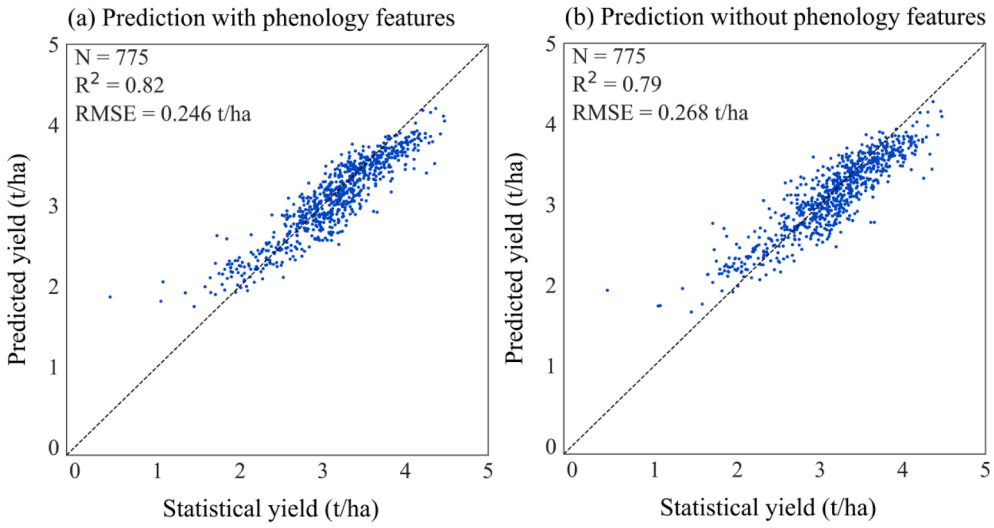


Fig. 16. Comparison of soybean yield prediction performances: (a) including phenology-based dynamic features and (b) without phenology-based dynamic features.

when not including them.

4.5. Limitations

The timeliness of the soybean mask layer limited this study. Herein, this study used a CDL to extract the soybean mask layer, but these data are usually released in the following year with a time lag. Near-real-time yield prediction requires timelier soybean mask data. Some studies have used the cropland layer instead of the crop layer (You et al., 2017) or performed early crop identification to address the missing soybean mask layer (Yaramasu et al., 2020; Zhang et al., 2019a). In terms of improving the timeliness of prediction, coupling the proposed approach with climate models is one way to potentially improve the accuracy of the yield predictions obtained for different years and climates.

Although some literature has claimed that XGBoost is not affected by multicollinearity (Guo et al., 2021), considering that the multicollinearity problem affects the interpretation of the predictors (Garg and Tai, 2013), we attempted to reduce the dimensionality using the principal component analysis (PCA) method while constructing the yield prediction framework based on the XGBoost model. There was no correlation problem among the obtained principal components. However, we found that the application of PCA significantly reduced the accuracy of the yield prediction when other conditions were kept consistent. The related experimental results of Otchere et al. (2021) and Wan et al. (2021) suggest that PCA reduces the prediction accuracy of XGBoost, probably because the factors obtained by PCA are more challenging to distinguish. In the future, more studies are needed to further validate yield prediction.

5. Conclusions

In this paper, a novel county-level soybean yield prediction framework coupled with XGBoost and multidimensional feature engineering was proposed. Our method outperformed LR, RF, KNN, ANN, SVR, LSTM, and DNN on the same soybean yield dataset, with the lowest RMSE (0.246 t/ha) and the highest R^2 (0.82) in the test year. The sequential structure, early stopping approach and L1 regularization term in the loss function reduce the risk of feature multicollinearity and overfitting problems. Multidimensional feature engineering extends the variable space and gives the model more windows to identify which variables are important from multiple dimensions. When evaluated individually, the application of multidimensional feature engineering and XGBoost significantly improved the test R^2 values by 3% and 4%–23%, respectively. The results indicate that accurate in-season forecasts of soybean yield at the county level in the United States can be made as early as the end of pod-setting and can provide timely information for soybean production prediction and trade decisions supporting. The combination of the feature importance of the XGBoost model and SHAP algorithm indicated that the EVI at pod-setting was the most crucial factor for soybean yield prediction, but yield prediction was not dependent on a few key features.

Our results demonstrate that accurate predictions can be made as early as the end of pod-setting; however, our prediction framework depends on an accurate soybean layer mask. With the availability of high-resolution satellites with global coverage (e.g., Landsat, Sentinel, and Planet satellites) and advances in cloud computing, in-season soybean mapping has become feasible. In the future, integrating in-season mapping of soybeans into our framework could provide operational and timely yield prediction of soybeans, which would support soybean production assessment and marketing policy development.

Funding

This research was supported by the National Key Research and Development Project of China (No. 2019YFE0126900), Natural Science Foundation of China (No. 41861144019), strategic consulting project of the Alliance of International Science Organizations (ANSO-SBA-2022-02), and the Youth Innovation Promotion Association of Chinese Academy of Sciences.

Credit authorship contribution statement

Yuanchao Li: Software, Visualization, Writing – original draft. **Hongwei Zeng:** Conceptualization, Writing – review & editing. **Miao Zhang:** Writing – review & editing. **Bingfang Wu:** Conceptualization, Writing – review & editing. **Yan Zhao:** Writing – review & editing. **Xia Yao:** Writing – review & editing. **Tao Cheng:** . **Xingli Qin:** Writing – review & editing. **Fangming Wu:** Writing – review & editing.

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

Data will be made available on request.

Acknowledgments

We are very grateful for the data provided by the Quick Stats Database of the United States Department of Agriculture, National Agricultural Statistics Service; the SSURGO database of the Natural Resources Conservation Service Soils, United States Department of Agriculture; and the MODIS data from NASA.

References

Abatzoglou, J.T., 2013. Development of gridded surface meteorological data for ecological applications and modelling. Int. J. Climatol. 33, 121–131.
Archontoulis, S.V., Castellano, M.J., Licht, M.A., Nichols, V., Baum, M., Huber, I., Martinez-Feria, R., Puntel, L., Ordonez, R.A., Iqbal, J., Wright, E.E., Dietzel, R.N., Helmers, M., Vanloocke, A., Liebman, M., Hatfield, J.L., Herzmann, D., Cordova, S. C., Edmonds, P., Togliatti, K., Kessler, A., Danalatos, G., Pasley, H., Pederson, C., Lamkey, K.R., 2020. Predicting crop yields and soil-plant nitrogen dynamics in the US Corn Belt. Crop. Sci. 60, 721–738.
Bengio, Y., Courville, A., Vincent, P., 2013. Representation learning: a review and new perspectives. IEEE Trans. Pattern Anal. Mach. Intell. 35, 1798–1828.
Blair, D.L., 1999. Intellectual property protection and its impact on the US seed industry. Drake J. Agric. L. 4, 297.
Bocca, F.F., Rodrigues, L.H.A., 2016. The effect of tuning, feature engineering, and feature selection in data mining applied to rainfed sugarcane yield modelling. Comput. Electron. Agr. 128, 67–76.
Bolton, D.K., Friedl, M.A., 2013. Forecasting crop yield using remotely sensed vegetation indices and crop phenology metrics. Agr. Forest Meteorol. 173, 74–84.
Borisov, V., Leemann, T., Seßler, K., Haug, J., Pawelczyk, M., Kasneci, G.J.A.P.A., 2021. Deep neural networks and tabular data: A survey.
Boryan, C., Yang, Z.W., 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.
Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.
Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., Niculae, V., Prettenhofer, P., Gramfort, A., Grobler, J.J.A.P.A., 2013. API design for machine learning software: experiences from the scikit-learn project.
Cai, Y.P., Guan, K.Y., Lobell, D., Potgieter, A.B., Wang, S.W., Peng, J., Xu, T.F., Asseng, S., Zhang, Y.G., You, L.Z., Peng, B., 2019. Integrating satellite and climate data to predict wheat yield in Australia using machine learning approaches. Agr. Forest Meteorol. 274, 144–159.
Cao, J., Zhang, Z., Luo, Y.C., Zhang, L.L., Zhang, J., Li, Z.Y., Tao, F.L., 2021a. Wheat yield predictions at a county and field scale with deep learning, machine learning, and google earth engine. Eur. J. Agron. 123, 126204.
Cao, J., Zhang, Z., Tao, F.L., Zhang, L.L., Luo, Y.C., Zhang, J., Han, J.C., Xie, J., 2021b. Integrating multi-source data for rice yield prediction across China using machine learning and deep learning approaches. Agr. Forest Meteorol. 297, 108275.
Castelvecchi, D., 2016. Can we open the black box of AI? Nature 538, 20–23.
Chen, T., Guestrin, C., 2016. Xgboost: a scalable tree boosting system. In: Proceedings of the 22nd ACM sigkdd International Conference on Knowledge Discovery and Data Mining, pp. 785–794.
Chen, Y., Zhang, Z., Tao, F.L., 2018. Improving regional winter wheat yield estimation through assimilation of phenology and leaf area index from remote sensing data. Eur. J. Agron. 101, 163–173.
Cosgrove, B.A., Lohmann, D., Mitchell, K.E., Houser, P.R., Wood, E.F., Schaake, J.C., Robock, A., Marshall, C., Sheffield, J., Duan, Q.Y., Luo, L.F., Higgins, R.W., Pinker, R. T., Tarpley, J.D., Meng, J., 2003. Real-time and retrospective forcing in the North American Land Data Assimilation System (NLDAS) project. J. Geophys. Res.-Atmos. 108.
Elavarasan, D., Vincent, P.M.D.R., Srinivasan, K., Chang, C.Y., 2020. A Hybrid CFS filter and RF-RFE wrapper-based feature extraction for enhanced agricultural crop yield prediction modeling. Agric.-Basel 10, 400.
Fan, J., Bai, J., Li, Z., Ortiz-Bobea, A., Gomes, C.P., 2022. A GNN-RNN approach for harnessing geospatial and temporal information: application to crop yield prediction. Proc. AAAI Conf. Artif. Intelli. 36, 11873–11881.
Feng, P.Y., Wang, B., Liu, D.L., Waters, C., Xiao, D.P., Shi, L.J., Yu, Q., 2020. Dynamic wheat yield forecasts are improved by a hybrid approach using a biophysical model and machine learning technique. Agr. For. Meteorol. 285, 107922.
Fritz, S., See, L., Bayas, J.C.L., Waldner, F., Jacques, D., Becker-Reshef, I., Whitcraft, A., Baruth, B., Bonifacio, R., Crutchfield, J., Rembold, F., Rojas, O., Schucknecht, A., Van der Velde, M., Verdin, J., Wu, B.F., Yan, N.N., You, L.Z., Gilliams, S., Mucher, S., Tetrault, R., Moorthy, I., McCallum, I., 2019. A comparison of global agricultural monitoring systems and current gaps. Agr. Syst. 168, 258–272.
Fuglie, K.O., 2007. Productivity Growth in US Agriculture. US Department of Agriculture, Economic Research Service.
Gao, F., Anderson, M., Daughtry, C., Johnson, D., 2018. Assessing the variability of corn and soybean yields in central iowa using high spatiotemporal resolution multi-satellite imagery. Rem. Sens.-Basel 10, 1489.
Garg, A., Tai, K., 2013. Comparison of statistical and machine learning methods in modelling of data with multicollinearity. Int. J. Model. Ident. Control 18, 295–312.
Gavahi, K., Abbaszadeh, P., Moradkhani, H., 2021. DeepYield: a combined convolutional neural network with long short-term memory for crop yield forecasting. Exp. Syst. Appl. 184, 115511.
Grinsztajn, L., Oyallon, E., Varoquaux, G., 2022. Why do tree-based models still outperform deep learning on tabular data?.
Guo, M., Yuan, Z., Janson, B., Peng, Y., Yang, Y., Wang, W., 2021. Older pedestrian traffic crashes severity analysis based on an emerging machine learning XGBoost. Sustainability 13, 926.
Han, J.C., Zhang, Z., Cao, J., Luo, Y.C., Zhang, L.L., Li, Z.Y., Zhang, J., 2020. Prediction of winter wheat yield based on multi-source data and machine learning in China. Rem. Sens.-Basel 12, 236.
Hansen, J.W., Potgieter, A., Tippett, M.K., 2004. Using a general circulation model to forecast regional wheat yields in northeast Australia. Agr. Forest Meteorol. 127, 77–92.
Heaton, J., 2016. An empirical analysis of feature engineering for predictive modeling. In: SoutheastCon 2016. IEEE, pp. 1–6.
Herrero-Huerta, M., Rodriguez-Gonzalvez, P., Rainey, K.M., 2020. Yield prediction by machine learning from UAS-based mulit-sensor data fusion in soybean. Plant Methods 16, 78.
Hlavinka, P., Trnka, M., Semerádová, D., Dubrovský, M., Žalud, Z., Možný, M., 2009. Effect of drought on yield variability of key crops in Czech Republic. Agr. Forest Meteorol. 149, 431–442.
Hunt, M.L., Blackburn, G.A., Carrasco, L., Redhead, J.W., Rowland, C.S., 2019. High resolution wheat yield mapping using Sentinel-2. Remote Sens. Environ. 233, 111410.
Hussain, A., Thapa, G.B., 2012. Smallholders' access to agricultural credit in Pakistan. Food Secur. 4, 73–85.
Jaafar, H.H., Ahmad, F.A., 2015. Crop yield prediction from remotely sensed vegetation indices and primary productivity in arid and semi-arid lands. Int. J. Remote Sens. 36, 4570–4589.
Jagtap, S.S., Jones, J.W., 2002. Adaptation and evaluation of the CROPGRO-soybean model to predict regional yield and production. Agr. Ecosyst. Environ. 93, 73–85.
Jain, A., Nandakumar, K., Ross, A., 2005. Score normalization in multimodal biometric systems. Pattern Recogn. 38, 2270–2285.
Jiang, H., Hu, H., Zhong, R., Xu, J., Xu, J., Huang, J., Wang, S., Ying, Y., Lin, T., 2020. A deep learning approach to conflating heterogeneous geospatial data for corn yield estimation: a case study of the US Corn Belt at the county level. Glob. Chang. Biol. 26, 1754–1766.
Johnson, D.M., 2014. An assessment of pre- and within-season remotely sensed variables for forecasting corn and soybean yields in the United States. Remote Sens. Environ. 141, 116–128.
Kang, Y.H., Ozdogan, M., 2019. Field-level crop yield mapping with Landsat using a hierarchical data assimilation approach. Remote Sens. Environ. 228, 144–163.
Kang, Y.H., Ozdogan, M., Zhu, X.J., Ye, Z.W., Hain, C., Anderson, M., 2020. Comparative assessment of environmental variables and machine learning algorithms for maize yield prediction in the US Midwest. Environ. Res. Lett. 15, 064005.
Keller, J.M., Gray, M.R., Givens, J.A., 1985. A fuzzy K-nearest neighbor algorithm. IEEE Trans. Syst. Man Cybern. 15, 580–585.
Khaki, S., Wang, L., Archontoulis, S.V., 2019. A CNN-RNN framework for crop yield prediction. Front. Plant. Sci. 10, 1750.
Kim, N., Ha, K.-J., Park, N.-W., Cho, J., Hong, S., Lee, Y.-W., 2019. A Comparison between major artificial intelligence models for crop yield prediction: case study of the Midwestern United States, 2006–2015. ISPRS Int. J. Geo Inf. 8, 240.
Klompenburg, T.V., Kassahun, A., Catal, C., 2020. Crop yield prediction using machine learning: a systematic literature review. Comput. Electron. Agr. 177, 105709.
Kouadio, L., Newlands, N.K., Davidson, A., Zhang, Y.S., Chipanshi, A., 2014. Assessing the performance of MODIS NDVI and EVI for seasonal crop yield forecasting at the ecodistrict scale. Rem. Sens.-Basel 6, 10193–10214.
Lepot, M., Aubin, J.B., Clemens, F.H.L.R., 2017. Interpolation in time series: an introductive overview of existing methods, their performance criteria and uncertainty assessment. Water 9, 796.
Lesk, C., Rowhani, P., Ramankutty, N., 2016. Influence of extreme weather disasters on global crop production. Nature 529, 84–87.
Liu, W.B., Wang, Z.D., Liu, X.H., Zengb, N.Y., Liu, Y.R., Alsaadi, F.E., 2017. A survey of deep neural network architectures and their applications. Neurocomputing 234, 11–26.
Lu, J., Carbone, G.J., Gao, P., 2017. Detrending crop yield data for spatial visualization of drought impacts in the United States, 1895–2014. Agr. Forest. Meteorol. 237–238, 196–208.
Lundberg, S.M., Lee, S.-I., 2017. A unified approach to interpreting model predictions. In: 31st Conference on Neural Information Processing Systems (NIPS 2017). Curran Associates, Inc., pp. 4765–4774.
Ma, Y.C., Zhang, Z., Kang, Y.H., Ozdogan, M., 2021. Corn yield prediction and uncertainty analysis based on remotely sensed variables using a Bayesian neural network approach. Rem. Sens. Environ. 259, 112408.
Maimaitijiang, M., Sagan, V., Sidike, P., Hartling, S., Esposito, F., Fritschi, F.B., 2020. Soybean yield prediction from UAV using multimodal data fusion and deep learning. Rem. Sens. Environ. 237, 111599.
Malik, W., Hoogemboom, G., Dechmi, F., Boote, K.J., Cavero, J., 2018. Adapting the CROPGRO model to simulate alfalfa growth and yield. Agron J 110, 1777–1790.
Mkhabela, M.S., Bullock, P., Raj, S., Wang, S., Yang, Y., 2011. Crop yield forecasting on the Canadian Prairies using MODIS NDVI data. Agr. Forest Meteorol. 151, 385–393.
Montgomery, D.C., Peck, E.A., Vining, G.G., 2021. Introduction to linear regression analysis. John Wiley & Sons.
Mushtaq, R., 2011. Augmented dickey fuller test. SSRN Electron. J.
Myneni, R., Knyazikhin, Y., Park, T., 2015. MCD15A3H MODIS/Terra+Aqua Leaf Area Index/FPAR 4-day L4 Global 500m SIN Grid V006, 2015 ed, NASA EOSDIS Land Processes DAAC.
Noble, W.S., 2006. What is a support vector machine? Nat. Biotechnol. 24, 1565–1567.

Obsie, E.Y., Qu, H.C., Drummond, F., 2020. Wild blueberry yield prediction using a combination of computer simulation and machine learning algorithms. Comput. Electron. Agr. 178, 105778.
Otchere, D.A., Ganat, T.O.A., Gholami, R., Lawal, M., 2021. A novel custom ensemble learning model for an improved reservoir permeability and water saturation prediction. J. Nat. Gas Sci. Eng. 91, 103962.
Oughali, M.S., Bahloul, M., El Rahman, S.A., 2019. Analysis of NBA players and shot prediction using random forest and XGBoost models. In: 2019 International Conference on Computer and Information Sciences (ICCIS). IEEE, pp. 1–5.
Panda, S.S., Ames, D.P., Panigrahi, S., 2010. Application of vegetation indices for agricultural crop yield prediction using neural network techniques. Rem. Sens.-Basel 2, 673–696.
Parsa, A.B., Movahedi, A., Taghipour, H., Derrible, S., Mohammadian, A.K., 2020. Toward safer highways, application of XGBoost and SHAP for real-time accident detection and feature analysis. Accid. Anal. Prev. 136, 105405.
Patel, N.R., Parida, B.R., Venus, V., Saha, S.K., Dadhwal, V.K., 2012. Analysis of agricultural drought using vegetation temperature condition index (VTCI) from Terra/MODIS satellite data. Environ. Monit. Assess. 184, 7153–7163.
Quiring, S.M., Papakryiakou, T.N., 2003. An evaluation of agricultural drought indices for the Canadian prairies. Agr. Forest Meteorol. 118, 49–62.
Raju, V.G., Lakshmi, K.P., Jain, V.M., Kalidindi, A., Padma, V., 2020. Study the influence of normalization/transformation process on the accuracy of supervised classification. In: 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT). IEEE, pp. 729–735.
Running, S., Mu, Q., Zhao, M.J.N.E.L.P.D., 2017. Mod16a2 modis/terra net evapotranspiration 8-day l4 global 500m sin grid v006. 6.
Sakamoto, T., 2020. Incorporating environmental variables into a MODIS-based crop yield estimation method for United States corn and soybeans through the use of a random forest regression algorithm. ISPRS J. Photogramm. Remote Sens. 160, 208–228.
Sakamoto, T., Gitelson, A.A., Arkebauer, T.J., 2014. Near real-time prediction of U.S. corn yields based on time-series MODIS data. Rem. Sens. Environ. 147, 219–231.
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. Agr. Forest Meteorol. 284, 107886.
Shahhosseini, M., Hu, G., Archontoulis, S.V., 2020. Forecasting corn yield with machine learning ensembles. Front. Plant Sci. 11, 1120.
Shahhosseini, M., Martinez-Feria, R.A., Hui, G.P., Archontoulis, S.V., 2019. Maize yield and nitrate loss prediction with machine learning algorithms. Environ. Res. Lett. 14, 124026.
Smith, A.B., Matthews, J.L., 2015. Quantifying uncertainty and variable sensitivity within the US billion-dollar weather and climate disaster cost estimates. Nat. Hazards 77, 1829–1851.
Song, X.P., Hansen, M.C., Potapov, P., Adusei, B., Pickering, J., Adami, M., Lima, A., Zalles, V., Stehman, S.V., Di Bella, C.M., Conde, M.C., Copati, E.J., Fernandes, L.B., Hernandez-Serna, A., Jantz, S.M., Pickens, A.H., Turubanova, S., Tyukavina, A., 2021. Massive soybean expansion in South America since 2000 and implications for conservation. Nat. Sust. 2021, 784–792.
Srivastava, A.K., Safaei, N., Khaki, S., Lopez, G., Zeng, W., Ewert, F., Gaiser, T., Rahimi, J., 2022. Winter wheat yield prediction using convolutional neural networks from environmental and phenological data. Sci Rep 12, 3215.
Sun, J., Di, L., Sun, Z., Shen, Y., Lai, Z., 2019. County-level soybean yield prediction using deep CNN-LSTM model. Sens. (Basel) 19, 4363.
Sun, J., Lai, Z.L., Di, L.P., Sun, Z.H., Tao, J.B., Shen, Y.L., 2020. Multilevel deep learning network for county-level corn yield estimation in the US corn belt. IEEE J.-Stars 13, 5048–5060.
Svoboda, M., LeComte, D., Hayes, M., Heim, R., Gleason, K., Angel, J., Rippey, B., Tinker, R., Palecki, M., Stooksbury, D., 2002. The drought monitor. Bull. Am. Meteorol. Soc. 83, 1181–1190.
Teal, R.K., Tubana, B., Girma, K., Freeman, K.W., Arnall, D.B., Walsh, O., Raun, W.R., 2006. In-season prediction of corn grain yield potential using normalized difference vegetation index. Agron. J. 98, 1488–1494.
Tianqi Chen, T.H., Michaël Benesty, Yuan Tang, 2021. Understand your dataset with XGBoost.
United States Census Bureau, 2013. Census Regions and Divisions of the United States, United States Census Bureau.
USDA/NASS, 2021. United States Department of Agriculture National Agricultural Statistics Service.
Vermote, E.J.N.E.L.P.D., 2015. MOD09A1 MODIS/terra surface reflectance 8-day L3 global 500m SIN grid V006. 10.
Walkinshaw, M., A.T. O'Geen, D.E. Beaudette, 2021. Soil Properties, California Soil Resource Lab.
Wan, Z., Hook, S., Hulley, G.J.N.E.L.P.D., 2015. MOD11A2 MODIS/Terra land surface temperature/emissivity 8-day L3 global 1km SIN grid V006. 10.
Wan, Z., Xu, Y., Savija, B., 2021. On the Use of Machine Learning Models for Prediction of Compressive Strength of Concrete: Influence of Dimensionality Reduction on the Model Performance. Materials (Basel) 14, 713.
Wang, S.C., 2003. Artificial neural network. In: Interdisciplinary Computing in Java Programming. Springer, pp. 81–100.
Wang, S., Di Tommaso, S., Deines, J.M., Lobell, D.B., 2020a. Mapping twenty years of corn and soybean across the US Midwest using the Landsat archive. Sci. Data 7, 307.
Wang, X.L., Huang, J.X., Feng, Q.L., Yin, D.Q., 2020b. Winter wheat yield prediction at county level and uncertainty analysis in main wheat-producing regions of china with deep learning approaches. Rem. Sens.-Basel. 12, 1744.
Wang, Y.M., Zhang, Z., Feng, L.W., Du, Q.Y., Runge, T., 2020c. Combining multi-source data and machine learning approaches to predict winter wheat yield in the conterminous United States. Rem. Sens.-Basel. 12, 1232.
Wu, Z., Huang, N.E., Long, S.R., Peng, C.K., 2007. On the trend, detrending, and variability of nonlinear and nonstationary time series. Proc. Natl. Acad. Sci. U.S.A. 104, 14889–14894.
Yaramasu, R., Bandaru, V., Pnvr, K., 2020. Pre-season crop type mapping using deep neural networks. Comput. Electron. Agr. 176, 105664.
You, J., Li, X., Low, M., Lobell, D., Ermon, S., 2017. Deep gaussian process for crop yield prediction based on remote sensing data. Thirty-First AAAI Conference on Artificial Intelligence.
Zamani Joharestani, M., Cao, C., Ni, X., Bashir, B., Talebiesfandarani, S., 2019. PM2.5 prediction based on random forest, XGBoost, and deep learning using multisource remote sensing data. Atmos. 10, 373.
Zhang, C., Di, L.P., Lin, L., Guo, L.Y., 2019a. Machine-learned prediction of annual crop planting in the US Corn Belt based on historical crop planting maps. Comput. Electron. Agr. 166, 104989.
Zhang, W.G., Wu, C.Z., Zhong, H.Y., Li, Y.Q., Wang, L., 2021. Prediction of undrained shear strength using extreme gradient boosting and random forest based on Bayesian optimization. Geosci. Front. 12, 469–477.
Zhang, Z., Jin, Y., Chen, B., Brown, P., 2019b. California almond yield prediction at the orchard level with a machine learning approach. Front Plant Sci 10, 809.
Zhou, X., Zheng, H.B., Xu, X.Q., He, J.Y., Ge, X.K., Yao, X., Cheng, T., Zhu, Y., Cao, W.X., Tian, Y.C., 2017. Predicting grain yield in rice using multi-temporal vegetation indices from UAV-based multispectral and digital imagery. ISPRS J. Photogramm. Remote Sens. 130, 246–255.