Science of the Total Environment 815 (2022) 152880

Contents lists available at ScienceDirect

Science of the Total Environment

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

Exploring the potential role of environmental and multi-source satellite data in crop yield prediction across Northeast China

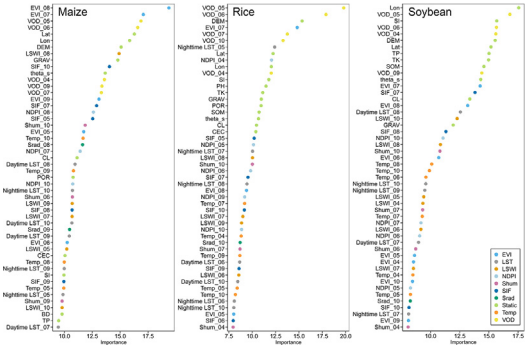
Zhenwang Li [a,*], Lei Ding [b], Dawei Xu [c]

[a] State Key Laboratory of Soil and Sustainable Agriculture, Institute of Soil Science, Chinese Academy of Sciences, Nanjing 210008, China
[b] College of Environmental and Resource Sciences, Zhejiang University, Hangzhou 310058, China
[c] National Field Scientific Observation and Research Station of Hulunbuir Grassland Ecosystem in Inner Mongolia, Institute of Agricultural Resources and Regional Planning, Chinese Academy of Agricultural Sciences, Beijing 100081, China

HIGHLIGHTS

- Stacked ensemble model outperformed single machine learning models.
- Environmental and satellite data are complementary and useful to estimate crop yield.
- SIF, LST and VOD provided extra information beyond EVI for crop yield prediction.
- Crop yield can be satisfactorily forecasted at two to three months before harvest.
- Geography, DEM, VOD, EVI, soil hydraulic and nutrient properties are important predictors.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 25 October 2021
Received in revised form 20 December 2021
Accepted 30 December 2021
Available online 6 January 2022

Editor: Martin Drews

Keywords:
Crop yield
Multi-source satellite data
Environmental data
Yield prediction
Machine learning

ABSTRACT

Developing an accurate crop yield predicting system at a large scale is of paramount importance for agricultural resource management and global food security. Earth observation provides a unique source of information to monitor crops from a diversity of spectral ranges. However, the integrated use of these data and their values in crop yield prediction is still understudied. Here we proposed the combination of environmental data (climate, soil, geography, and topography) with multiple satellite data (optical-based vegetation indices, solar-induced fluorescence (SIF), land surface temperature (LST), and microwave vegetation optical depth (VOD)) into the framework to estimate crop yield for maize, rice, and soybean in northeast China, and their unique value and relative influence on yield prediction was assessed. Two linear regression methods, three machine learning (ML) methods, and one ML ensemble model were adopted to build yield prediction models. Results showed that the individual ML methods outperformed the linear regression methods, the ML ensemble model further improved the single ML models. Moreover, models with more inputs achieved better performance, the combination of satellite data with environmental data, which explained 72%, 69%, and 57% of maize, rice, and soybean yield variability, respectively, demonstrated higher yield prediction performance than individual inputs. While satellite data contributed to crop yield prediction mainly at the early-peak of the growing season, climate data offered extra information mainly at the peak-late season. We also found that the combined use of EVI, LST and SIF has improved the model accuracy compared to the benchmark EVI model. However, the optical-based vegetation indices shared similar information and did not provide much extra information beyond EVI. The within-season yield forecasting showed that crop yields can be satisfactorily forecasted at two to three months prior to harvest. Geography, topography, VOD, EVI, soil hydraulic and nutrient parameters are more important for crop yield prediction.

* Corresponding author at: State Key Laboratory of Soil and Sustainable Agriculture, Institute of Soil Science, Chinese Academy of Sciences, Nanjing 210008, China.
E-mail address: zwli@issas.ac.cn (Z. Li).

http://dx.doi.org/10.1016/j.scitotenv.2021.152880

1. Introduction

Understanding the spatiotemporal patterns of crop yield, along with accurately predicting those patterns are a challenging issue and a key research area in agricultural studies (Franz et al., 2020; Li et al., 2019b). Such estimates can facilitate better assessments of yield response to environmental stresses, help to better understand the gaps between actual and potential yields, and thus provide better information for farm resource management (Guan et al., 2017; Ma et al., 2021). Moreover, information about crop yield at the regional and national scales can provide important information to food security, agricultural commodity markets, and to guide policy decision-making (Hoffman et al., 2015; Sherrick et al., 2014).

In agriculture, crop yield is strongly influenced by various variables including environment (e.g. climate, soil properties), genetics, and management (Mathieu and Aires, 2018), all these factors need to be generally considered in monitoring and forecasting crop yield through statistical or physical simulation models. Climate data and soil properties describe the environmental information that constrains the growing condition of the crop, they are extensively used in crop predicting systems. However, crop growing status is not only affected by abiotic factors, but also by biotic factors (Cai et al., 2019; Lichtenthaler, 1996; Mahlein et al., 2012). Thus, simply using environment data may not be sufficient.

Satellite remote sensing (RS) data has been widely used for crop yield estimation across a wide range of scales and geographic locations (Guan et al., 2017; Sakamoto et al., 2013; Sibley et al., 2014). Previous studies have also shown better yield estimations by using satellite data or combining satellite data with environmental information than using climate data only (Cai et al., 2019; Li et al., 2019b). In particular, the Normalized Difference Vegetation Index (NDVI) and Enhanced Vegetation Index (EVI) derived from visible and near-infrared (NIR) satellite data, which provide a general indicator of photosynthetic canopy cover or aboveground biomass, were the most commonly used predictors for their long-time records and recognized value in monitoring crop conditions. However, these vegetation indices (VIs) only utilize information from a small portion of the electromagnetic spectrum within optical wavelengths, the crop information they provided is also limited. In fact, current earth observation satellites can capture crop growing conditions from a diversity of spectral ranges (Guan et al., 2017), including visible, infrared, thermal, and microwave, satellite observations from these data platforms provide unique information that can describe crop growth condition from both biotic and abiotic stresses. Solar-induced chlorophyll fluorescence (SIF), derived from a specific narrow range of the near-infrared band, has emerged as a proxy of plant photosynthesis (Guanter et al., 2014; Porcar-Castell et al., 2014). Some other VIs like normalized difference phenology index (NDPI) and land surface water index (LSWI) that use broader spectral wavelengths from visible to shortwave infrared (SWIR) are found to have better performances in detecting crop biomass and canopy water content (Dong et al., 2015; Xu et al., 2021). The thermal RS data, a direct measurement of land-surface temperature (LST), can capture heat stress and drought impact on yield variations (Johnson, 2014; Khanal et al., 2017), it was also considered to be an alternative to air temperature in data-limited regions (Heft-Neal et al., 2017). Microwave data with longer wavelength bands, either passive or active, is commonly referred to as vegetation optical depth (VOD) using microwave radiative transfer models (Jackson and Schmugge, 1991; Vreugdenhil et al., 2016), this indicator provides frequency-dependent information related to the crop canopy density, biomass, and water content of vegetation (Liu et al., 2015; Momen et al., 2017). VOD estimates from long wavelengths (e.g. C, L or P-band) are generally more sensitive to deeper vegetation layers while VOD estimates from short wavelengths (e.g. Ku-, X-band) are more sensitive to leaf moisture content (Chaparro et al., 2018; Konings et al., 2019; Tian et al., 2018). Satellite data from a single platform or combinations of several platforms were extensively used in crop monitoring, however, crop yield estimation using the whole set of the available spectral bands has been comparatively less studied (Guan et al., 2017).

Generally, two yield prediction methods have been widely used: the physical simulation models and the statistical models. Physical-based crop models estimate yield by dynamically simulating crop growth and yield formation processes (Jeong et al., 2022; Jones et al., 2017; Rosenzweig et al., 2014), even powerful, these models require extensive locally crop specific biotic and abiotic inputs, limiting their applicability in large-scale yield modeling (Kang and Özdoğan, 2019; Lecerf et al., 2019). Statistical models are widely used in operational large-scale crop yield forecasting systems due to their simplicity, fewer inputs required, and relatively high predictive power when sufficient training data are available (Chipanshi et al., 2015; Johnson, 2014; Li et al., 2019b). Comparatively, statistical machine learning (ML) models have complex functions and abilities to handle complicated relationships between the predictors and the target variable (Johnson et al., 2016; Ma et al., 2021), thus the approaches have been increasingly employed in the research fields of agriculture in recent years (Cao et al., 2021; Schauberger et al., 2020).

Given the public availability of global environmental data and various remote sensing products across a diverse spectral range, each of them can provide unique information and offer new opportunities for agricultural monitoring. Several questions regarding the integration and use of these data remain: (1) How much information can the environmental and satellite data provide to the crop yield prediction and what combinations of these data will achieve the best performance? (2) What is the performance of various satellite data in predicting crop yield and how to combine them for crop yield predictions in northeast China? (3) How does the within-season crop yield forecasting perform with the progression of growing season and more data available? In this study, we used environmental data including climate, soil, geography, and topography and a diverse set of satellite data introduced above to build two linear regression models (partial least-square regression (PLSR) and least absolute shrinkage and selection operator (LASSO)), three machine learning models (stochastic gradient boosting (SGB), support vector regression (SVR), and random forest (RF)), and one ML ensemble model for yield prediction of three major crops (maize, rice, and soybean) across northeast China. The end-of-season yield modeling using different methods and different combinations of inputs was conducted to quantify the contributions of the environmental and satellite data in determining crop yield, the within-season yield forecasting was conducted to analyze the model performance with crop growth progression and more input information. The results of this study will facilitate the synergistic use and development of new generation satellite RS products, and provide valuable information for crop yield forecasting system development.

2. Materials

2.1. Study region

Our study was conducted in the Northeastern region of China, which includes the Heilongjiang province, the Jilin province, and the Liaoning province, the total area is about 0.79 million km^2 (Fig. 1). Northeast China is the leading grain production region in China with a crop planting area of 0.26 million km^2, which occupies more than 15% of the total crop planting area in China, about one-fifth of the national grain is produced here. The major crops are maize, soybean, and rice, the sum of the planting area of these three crops exceeded 90% of the total crop planting areas in Northeast China (Hu et al., 2021; You et al., 2021). On the other hand, Northeast China is one of the most susceptive areas to climate change of China. It spans the warm temperate zone, the mid-temperate zone, and the cold temperate zone from south to north, with annual accumulated air temperatures above 0 °C range from 2000 to 4200 °C·day and annual precipitation ranging from 500 to 800 mm. The meteorological disasters, such as droughts, floods, and cold damage, have become more frequent with the increase in climatic variability, which makes northeast China to be one of the areas with the greatest fluctuation in grain yields in China.

2.2. Crop yield

We obtained harvested yield for maize, rice, and soybean from the agricultural statistical yearbook of each statistical division (http: //

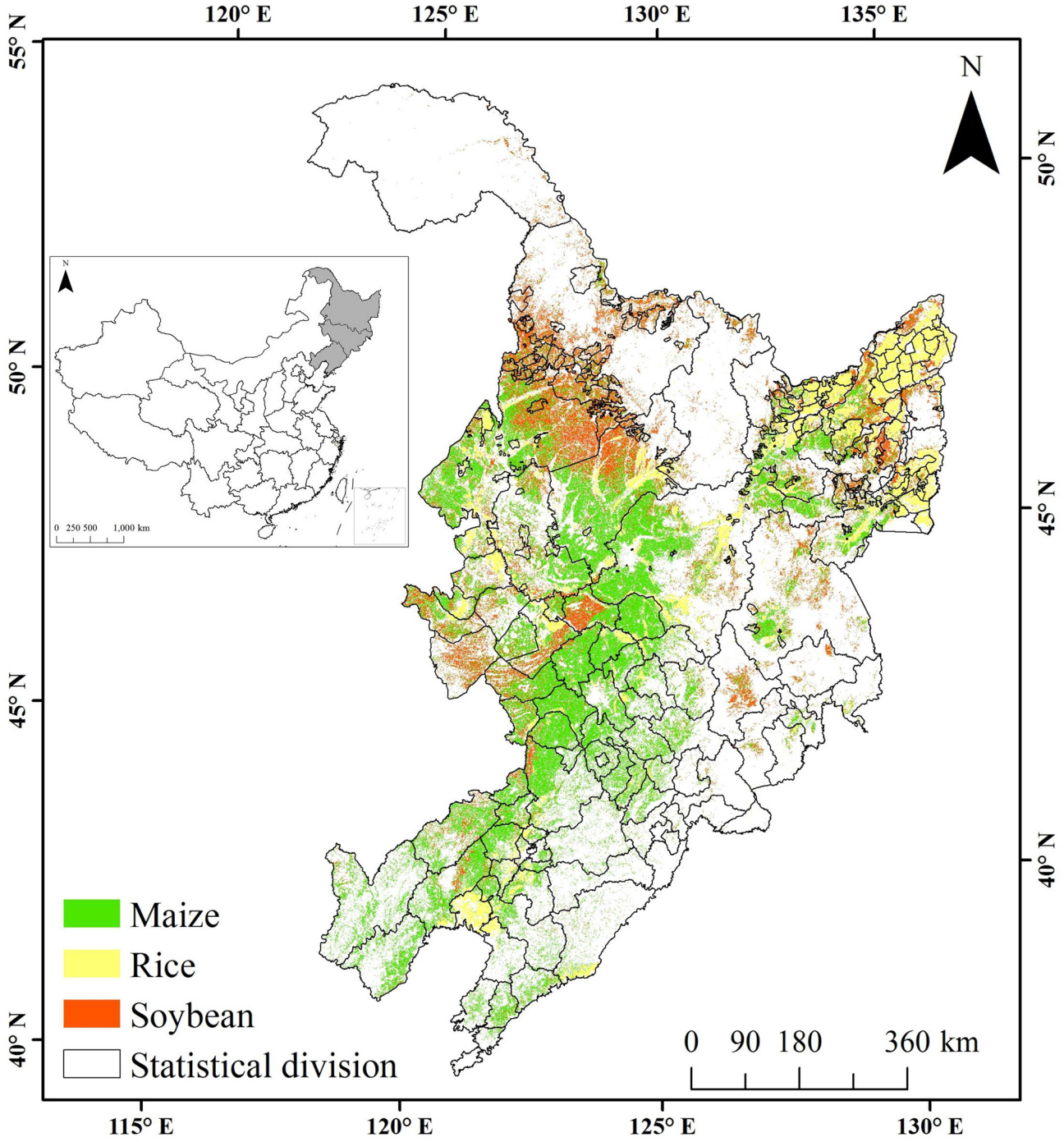


Fig. 1. The statistical division, spatial distribution of maize, rice and soybean cultivation area in northeast China.

www.stats.gov.cn) from 2003 to 2018 (unit: kg/ha). According to the data availability, the crop yield was collected at the county-level in Jilin province and Heilongjiang land reclamation system, and the crop yield of Liaoning province and Heilongjiang province was collected at the city-level. To reduce potential uncertainty in our following models, a quality check was firstly conducted to identify and filter the outliers, the yield records that fell outside the reasonable range (the mean plus or minus two times the variance) were discarded, the statistical division where the total crop fraction below 10% were also abandoned. Overall, a totally of 2265, 1688, and 2117 yield records for maize, rice, and soybean respectively were selected at the research areas of this study (Fig. 1).

2.3. Satellite remote sensing data

2.3.1. MODIS VIs

Three satellite VIs, including Enhanced Vegetation Index (EVI), Normalized Difference Phenology Index (NDPI), and Land Surface Water Index (LSWI), were used in this study. EVI is a widely-used VI based on the leaf red-edge spectral feature in the red and NIR spectral bands. The EVI is similar to the NDVI, but sensitive to higher canopy leaf area index and less affected by atmospheric aerosol impacts (Huete et al., 2002). Here we use EVI from the NASA Terra Moderate Resolution Imaging Spectroradiometer (MODIS) MOD13A3 (Collection 6) record, with monthly repeat and global 1 km spatial resolution. NDPI uses a weighted

red-SWIR combination to replace the red band in the NDVI, it has proven to have good performance in alleviating the adverse impact of soil background while keeping high sensitivity to canopy water content and above-ground biomass (Wang et al., 2017; Xu et al., 2021). LSWI is calculated as normalized ratios between NIR and SWIR bands, due to the sensitive of SWIR bands to soil/leaf water content, LSWI is widely employed to monitor the water content changes in crop leaves (Xiao et al., 2006). NDPI and LSWI used in this study were extracted from the daily MODIS MCD43A4 product at 500 m spatial resolution. All these MODIS data preprocessing was implemented on the GEE (Google Earth Engine) platform (https://earthengine.google.com/).

$$EVI = 2.5 \times \frac{(NIR - Red)}{NIR + 6 \times Red - 7.5 \times Blue + 1} \quad (1)$$

$$NDPI = \frac{NIR - (0.74 \times Red + 0.26 \times SWIR)}{NIR + (0.74 \times Red + 0.26 \times SWIR)} \quad (2)$$

$$LSWI = \frac{NIR - SWIR}{NIR + SWIR} \quad (3)$$

where Blue, Red, NIR, and SWIR represent the atmospherically-corrected surface reflectance in blue, red, NIR, and SWIR bands, respectively.

2.3.2. Satellite-based SIF data

A newly developed monthly 0.05° global SIF product (collected from https://cornell.box.com/s/gkp4moy4grvqsus1q5oz7u5lc30i7o41) retrieved near 740 nm spectral window was used in this study (Wen et al., 2020). The fine-resolution and long-term SIF product was downscaled and harmonized from the 1° SCanning Imaging Absorption spectroMeter for Atmospheric CHartographY (SCIAMACHY) SIF dataset and the 0.5° Global Ozone Monitoring Experiment 2 (GOME-2) onboard MetOp-A developed at German Research Center for Geosciences (GFZ) SIF dataset using machine learning techniques, the harmonized SIF dataset was then validated using ground-measured SIF, and found good consistent with the tower SIF in terms of both magnitude and timing. Moreover, the dataset also demonstrated the capability in characterizing plant stress during droughts and heatwaves while preserving the spatial and temporal variability contained in the original SIF dataset.

2.3.3. MODIS LST

Satellite daytime and nighttime land surface temperature (LST) data were collected and extracted from the thermal infrared bands-based MODIS Aqua product (MYD11B3), which provides monthly data with a 5600 m spatial resolution (Wan et al., 2021). The daytime and nighttime MODIS LST from Aqua has an overpass time of 1:30 PM and 1:30 AM (local time) respectively which approximates the maximum and minimum temperature of a day, and each LST pixel value in the MYD11B3 is a simple average of all the corresponding values from the daily LST collected during the month period. The product was collected from the NASA's Earthdata search client website (https://search.earthdata.nasa.gov/).

2.3.4. Vegetation optical depth (VOD)

We used a new series of satellite passive-microwave-based VOD product (VODCA) with 0.25° spatial resolution and daily frequency (Moesinger et al., 2020). The product combines VOD retrievals that have been derived from multiple sensors (Special Sensor Microwave/Imager (SSM/I), Microwave Imager on board the Tropical Rainfall Measuring Mission (TMI), Advanced Microwave Scanning Radiometer - Earth Observing System (AMSR-E), WindSat, and AMSR2) using the Land Parameter Retrieval Model (Owe et al., 2008), and produces three separate VOD products for microwave observations in different spectral bands, namely the K-band (18.70 or 19.35 GHz), X-band (10.65 GHz), and C-band (6.93–7.30 GHz).

2.4. Environmental data

2.4.1. Climate data

The climate variables were extracted from the China regional surface meteorological feature dataset (He et al., 2020). The dataset was produced by merging a variety of data sources, including Princeton meteorological forcing data, Global Land Data Assimilation System (GLDAS) data, The Global Energy and Water Cycle Experiment-Surface Radiation Budget (GEWEX-SRB) shortwave radiation dataset, Tropical Rainfall Measuring Mission (TRMM) satellite precipitation analysis data and China Meteorological Administration (CMA) station data and has a temporal and spatial resolution of 3 h and 0.1°, respectively. The dataset incorporated CMA station data, therefore, it is more accurate in China compared with other datasets and is generally preferable for modeling studies in China (Chen et al., 2011; Liu and Xie, 2013). The primary climate variables utilized in this study included mean temperature (Temp), precipitation (Pre), specific humidity (Shum), and shortwave radiations (Srad) and were acquired from the National Tibetan Plateau/Third Pole Environment Data Center (https://data.tpdc.ac.cn).

2.4.2. Soil data

Soil properties are also critical for plant growth and have significant impacts on crop yield. The soil physical and chemical attributes used in this study were extracted from the 1 km resolution China soil properties and soil hydraulic parameters dataset (http://globalechange.bnu.edu.cn), which was derived from 8979 soil profiles and the Soil Map of China using the polygon linkage method (Dai et al., 2013; Shangguan et al., 2013). We considered nine topsoil (0–4.5 cm) soil properties and two soil hydraulic parameters in our study, including bulk density (BD), cation exchange capacity (CEC), soil texture, organic carbon content (SOC), pH, rock fragment (GRAV), porosity (POR), total phosphorus (TP), total potassium (TK) saturated water content (theta_s) and saturate hydraulic conductivity (k_s).

2.4.3. Geography and topography data

In addition to the climate and soil data, we also used the geography and topography data in our yield prediction models, the former is the longitude and latitude of the center of each statistical division, and the latter is the elevation extracted from the 1 km SRTM GDEM dataset released from the NASA Land Processes Distributed Active Archive Center (https://lpdaac.usgs.gov/).

2.5. Crop type maps

The 10 m crop type maps in Northeast China of 2018 were used to aggregate the satellite and environmental variables to the statistical division for maize, rice and, soybean, respectively. The crop type maps are a yearly time-series Sentinel-2 images based crop type classification product using the random forest (RF) algorithm and a sophisticated feature selection procedure (You et al., 2021). The classification accuracy for maize, rice, and soybean is 82%, 96%, and 79%, respectively. Details on the data for this study are provided in Table 1.

3. Methodology

3.1. Exploratory data analysis

Before model construction, an exploratory data analysis was conducted to reduce input dimensionality and select appropriate inputs. For this purpose, simple correlation analysis was firstly performed between each variable and crop yield. Variables that have insignificant correlation ($P > 0.01$) with crop yield were discarded to avoid bias and exclude the impractical variables. For the remaining variables that have significant correlation with yield variations, the correlation coefficients among the variables were calculated to identify the highly correlated variables with correlation coefficients large than 0.85. In order to avoid bias resulting from the co-

Table 1
Summary of the datasets used in this study.

Category	Variable (abbreviation)	Spatial Resolution	Temporal Resolution	Time Coverage	Source/Reference
Crop yield	Crop yield	Statistical division	Yearly	2003–2018	Agricultural Statistical Yearbook (http://www.stats.gov.cn)
Satellite data	Enhanced vegetation index (EVI)	1000 m	Monthly	2003–2018	MODIS MOD13A3 EVI product
	Normalized difference phenology index (NDPI) Land surface water index (LSWI)	500 m	Daily	2003–2018	MODIS MCD43A4 NBAR product
	Solar-induced fluorescence (SIF)	0.05°	Monthly	2003–2018	Wen et al. (2020)
	Land surface temperature (LST)	5600 m	Monthly	2003–2018	MODIS MYD11B3 LST product
	Vegetation optical depth (VOD)	0.25°	Daily	C-band (2003–2018), K-band (2003–2016), X-band (2003–2018)	Moesinger et al. (2020)
Climate data	Mean temperature (Temp), precipitation (Prec), specific humidity (Shum), shortwave radiation (Srad)	0.1°	3 h	2003–2018	He et al. (2020)
Soil property	Bbulk density (BD), cation exchange capacity (CEC), soil texture, organic carbon content (SOC), pH, rock fragment (GRAV), porosity (POR), total phosphorus (TP), total potassium (TK), saturated water content (theta_s), saturate hydraulic conductivity (k_s)	1000 m	–	–	Dai et al. (2013), Shangguan et al. (2013)
Geography	Longitude (Lon), latitude (Lat)	Statistical division	–	–	–
Topography	Elevation	1000 m	–	–	NASA LPDAAC (https://lpdaac.usgs.gov/)
Crop type maps	Maize, rice, soybean map	10 m	Yearly	2018	You et al. (2021)

linearity of the environmental variables, the variable that has a smaller absolute correlation coefficient value with crop yield in each high correlated variable pair was discarded.

3.2. Model configuration

We used two linear regression methods (PLSR and LASSO), three ML methods (SGB, SVR, and RF), and one ML ensemble model to develop the crop yield models. A general description of the six models is provided in the Supplementary Material. Prior to yield prediction, all the satellite and environmental data were spatially resampled to 1 km using nearest neighbor method and temporally aggregated to the monthly steps. Then, the crop type maps of the three crops were resampled to 1 km and were used on the image layers to mask out the non-target crop pixels from 2003 to 2018, and all the crop masked satellite and environmental layers were processed to extract statistical division-level data based on the statistical division boundaries. To evaluate the performance of the models, the ten-fold-cross-validation method was used, the method randomly splits the whole dataset into 10 folds, and uses 9 folds for training and 1 fold for testing. For each model, the crucial hyper-parameters were optimized based on the highest R^2 and the lowest RMSE by ten-fold cross-validated using the training data. Finally, the optimized models were applied to the testing dataset and calculated the predicted R^2s and RMSEs.

3.3. Yield prediction using various combinations of environmental and satellite data

To answer the research questions proposed in this paper, three groups of input variables were designed and applied with the two linear regression methods and four machine learning methods. The first group was designed to explore the contributions of either environmental data or satellite data for predicting crop yield. We divided the potential input data into three types of feature sets, including climate data, satellite data, and static variables (soil properties, geography, and topography data), the following six combinations of inputs were set, they are (1) Climate only; (2) Satellite only; (3) Climate plus satellite; (4) Static plus climate; (5) Static plus satellite; (6) Static plus climate and satellite.

The second group was designed to identify the unique and overlapping contributions of satellite data from a diversity of spectral bands to crop yield prediction. Due to the recognized value and widely utilization of EVI in crop yield estimation, the combined variables of static plus EVI was set as the benchmark input, and the following ten combinations of inputs were set, they are (1) Static plus EVI; (2) Static plus NDPI; (3) Static plus LSWI; (4) Static plus SIF; (5) Static plus LST; (6) Static plus VOD; (7) Static plus EVI, NDPI and LSWI; (8) Static plus EVI and SIF; (9) Static plus EVI and LST; (10) Static plus EVI and VOD. Since the three VOD products performed almost the same in predicting crop yield (Fig. S1), the VOD data used in the above combinations were extracted from the C-band VOD product. The above two groups were run on an end-of-season mode, the predictions were conducted with the full knowledge of each combination, and the estimation of the crop yield can be performed only once the growing season is concluded. This mode aims at analyzing the sensitivity of environmental data and satellite data from a diversity of spectral bands on the final crop yield, obtaining an independent crop yield assessment at the end of the season.

The third group was conducted on a within-season forecasting mode which intends to estimate the final crop yield during the growing season, before the harvest. All predictor variables after exploratory data analysis and the ML ensemble method were used to develop the yield forecasting system. The multi-source environmental and satellite variables were firstly aggregated into six groups by different month (from April to October) for each crop type. The forecasting events were then triggered successively at each month, and the predictors were added with crop growth progression. The leave-one-year-out cross validation method was used to evaluate the model performance.

3.4. Variable importance

To demonstrate the most important predictors, the relative importance of each input variable was calculated using the Boruta algorithm (Kursa and Rudnicki, 2010). Boruta is a wrapper algorithm in which several runs of random forest regression are performed. Before each run, a shadow feature, which is derived by shuffling the values of the original feature across data items, is created for each feature. After each random forest run, the features

that have significantly lower importance than the shadow feature with the highest importance are classified as unimportant and removed from the following runs, the features that have significantly higher importance are classified as important and are included in the following runs. Boruta ends when all features are classified either as unimportant or important or when a specified limit of random forest runs is reached. In this study, we run the Boruta algorithm with a 0.99 confidence level, z-scores of mean decrease accuracy measure to gather permutation importance, and 100 maximum runs of random forest.

4. Results

4.1. Correlation of crop yield with climate and remote sensing variables

Before exploratory data analysis, the seasonal cycles of all the input covariates were examined (Fig. 2). Generally, all the covariates showed a similar pattern of mid-summer peak during the crop growing season (July–August) but differed with onset and peak timing. EVI, NDPI, LSWI, and VOD reached their peak in August, while EVI, NDPI, LSWI demonstrated similar seasonal cycles, the seasonal cycle of VOD lags behind the other covariates both at the beginning and the end of the growing season. SIF and nighttime LST increased faster than EVI in the green-up stage and reached their peak in July, which was consistent with climate variables including Temp, Prec, and Shum. The two covariates also showed similar cycles and have earlier drops than EVI during the latter portion of the growing season, indicating photosynthesis is susceptible to climate variations. However, compare to SIF, the nighttime LST showed a faster increasing rate in the green-up stage and a slower decreasing rate after the peak timing. The seasonal cycles of daytime LST and Srad differed with the other covariates, these two covariates increased earlier than the other covariates and reached their peak in June, after that, the two covariates start to decrease with a slow rate. The three VOD products showed similar seasonal cycles (Fig. S2), only small difference of signal magnitude was found in July and October.

We then conducted the exploratory data analysis by looking at correlations among all the satellite/climate variables and crop yield from 2003 to 2016 (Fig. 3). The maize yield was better correlated with remote sensing variables (EVI, NDPI, LSWI, SIF, LST, and VOD) than climate variables

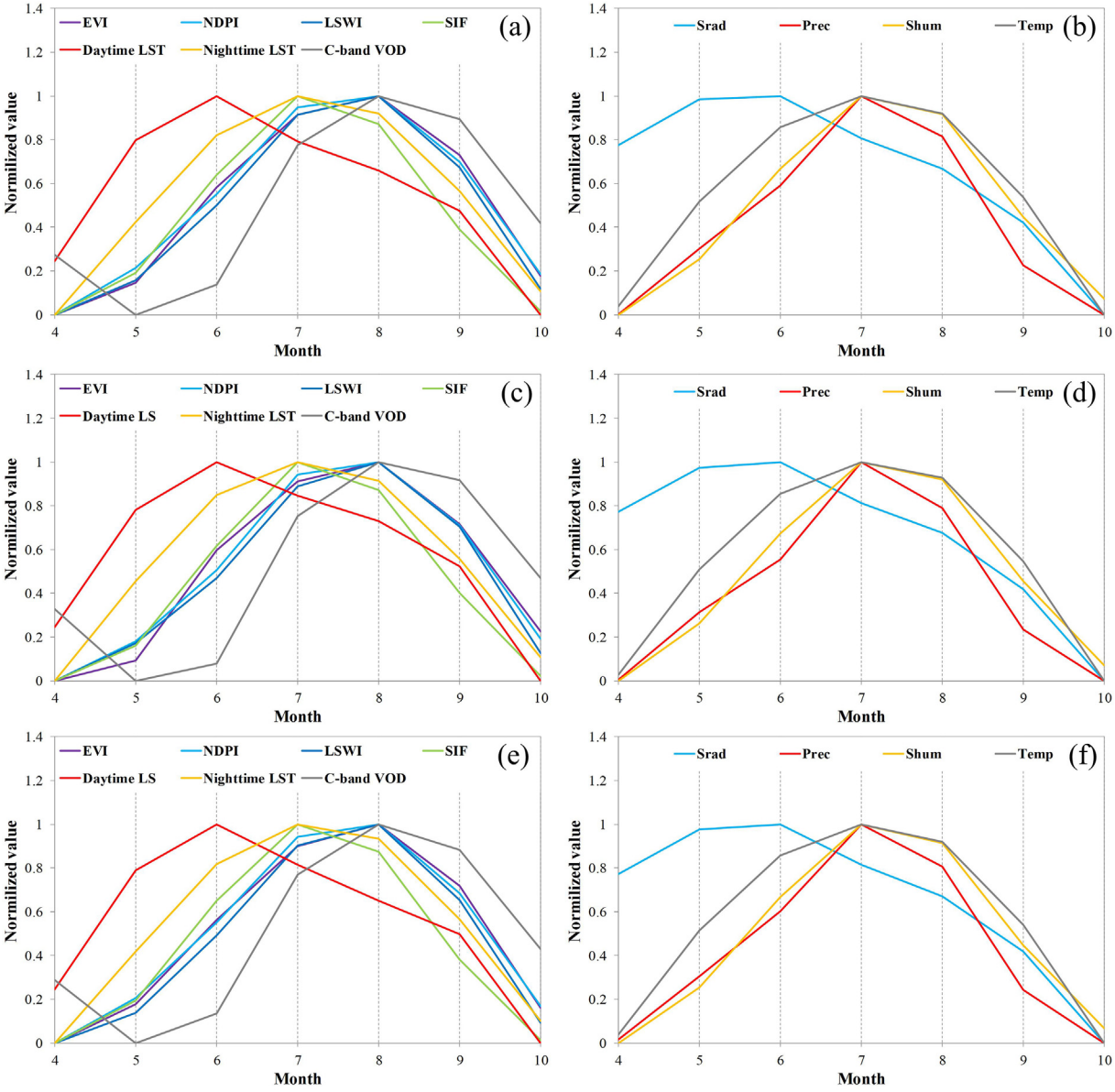


Fig. 2. Normalized average monthly values of satellite (left column) and climate (right column) covariates averaged for the study region of maize (top), rice (middle) and soybean (bottom), the original values were normalized to 0–1 to match their minimum and maximum values.

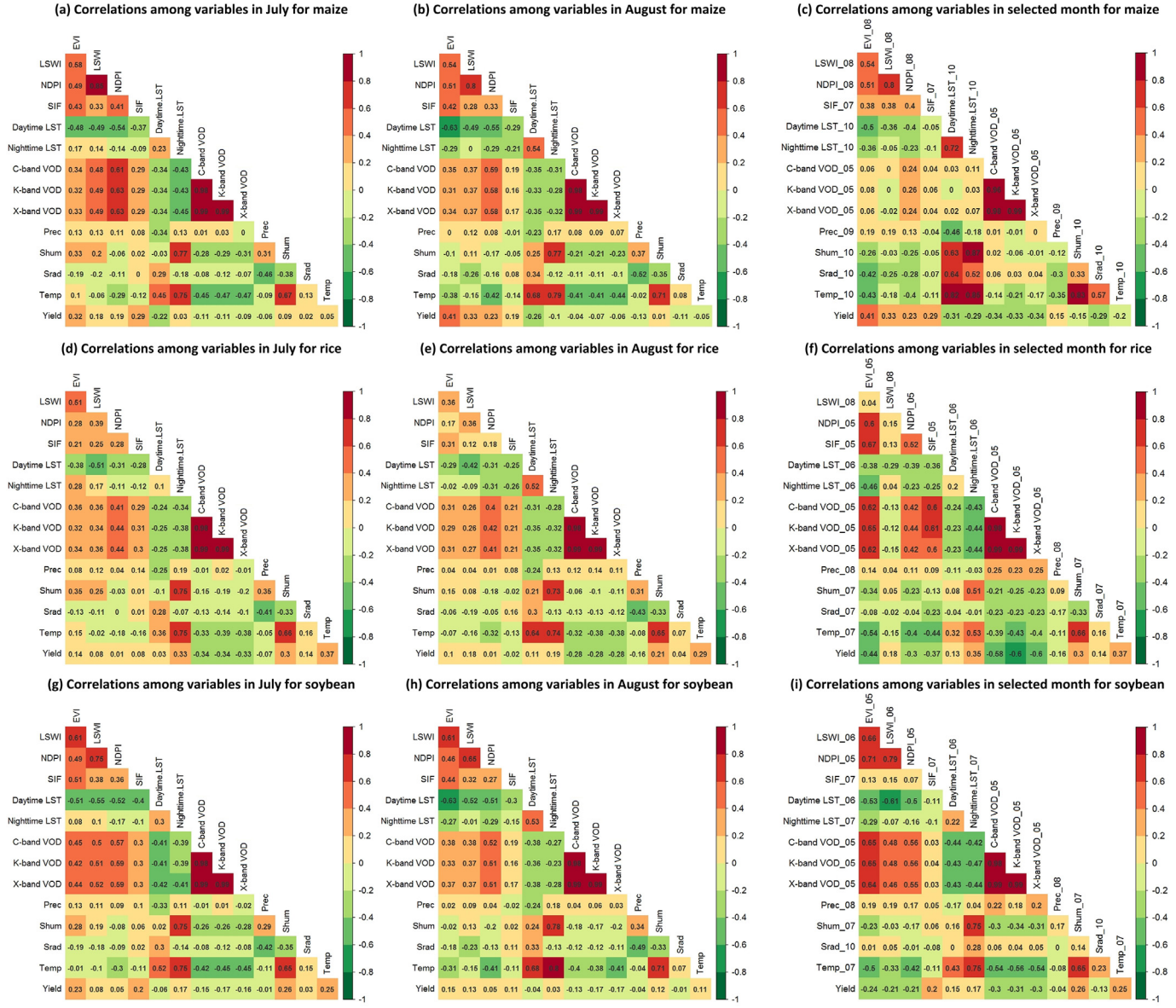


Fig. 3. Pearson correlation coefficients among statistical division-level climate and satellite variables in July (left column), August (middle column) and selected month (month that have the highest correlation coefficient with yield, right column) as well as final harvested yield for maize (top), rice (middle) and soybean (bottom).

(Temp, Prec, Shum, and Srad), EVI in July and August, LSWI in August, daytime LST in October and VOD in May showed correlation coefficients larger than 0.3 with maize yield. The climate variables were less correlated with maize yield with correlation coefficients smaller than 0.2, except for the Srad in October. Among all the climate variables, maize yield was more correlated with Prec in September, Temp, Shum, and Srad in October than in the other months (Fig. 3c). Different patterns were observed for rice and soybean, during the peak growing season (July and August), the yield of these two crops was better correlated with climate variables (Temp, Prec, and Shum) than the remote sensing variables (EVI, NDPI, LSWI, SIF, and LST) (Fig. 3d, e, g, & h), while remote sensing variables in the early growing season showed better correlations with rice and soybean yield. EVI in May, nighttime LST in June and July, NDPI, SIF, and VOD in May showed significant correlations with rice yield (correlation coefficients larger than 0.3), Shum and Temp in July also positively correlated with rice yield with correlation coefficients larger than 0.3 (Fig. 3f). For soybean, EVI, and SIF in July, LSWI in June, NDPI, and VOD in May showed better correlations with soybean yield than the other variables, Shum and Temp in July also positively correlated with soybean yield (Fig. 3i). The three VOD products in May, July and August all showed significant negative correlations with crop yield, while VOD in May has the best correlation with crop yield relative to VOD in the other months. Maize yield was slightly more correlated with X-band VOD, while rice and soybean yield was slightly more correlated with K-band VOD.

There were also strong correlations among climate and remote sensing variables themselves. Except for the strong correlations between the three VOD products, significant positive correlation coefficients were observed between NDPI and LSWI and between NDPI and VOD for both maize and soybean, respectively, Temp positively correlated with nighttime LST, daytime LST and Shum for all the three crops. Compared with Temp and Shum, Prec and Srad are less correlated with remote sensing variables.

4.2. Performance of environmental and satellite data in crop yield prediction

The first group experiment was conducted by applying the six methods using the six combinations of environmental and satellite data from 2003 to 2018, the performance of yield prediction is shown in Fig. 4. Generally, better performance is achieved with more input data. The ML ensemble model

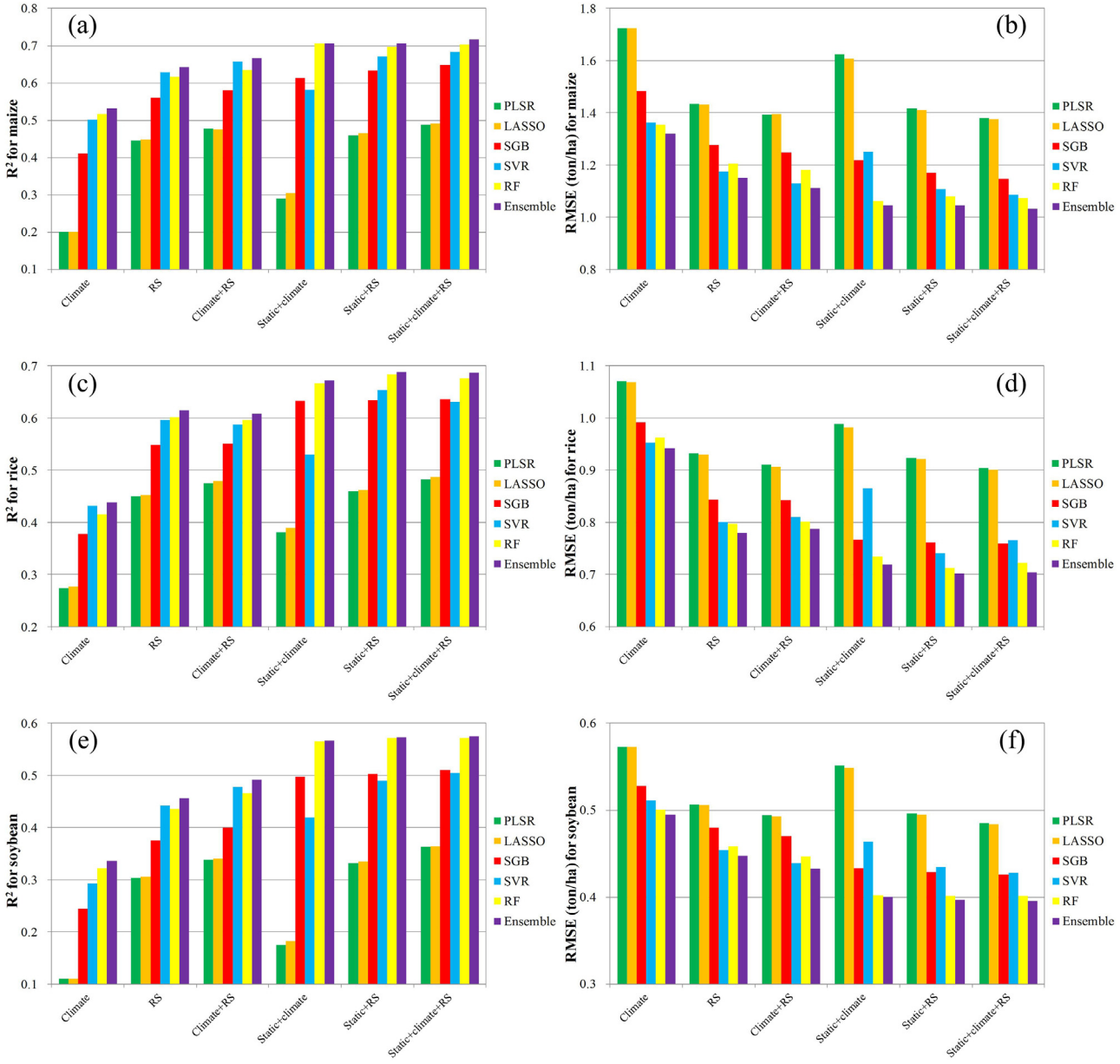


Fig. 4. The model performance (R^2 and RMSE) of maize (top), rice (middle) and soybean (bottom) yield prediction using the six methods and different combinations of environmental and satellite data for the whole growing season.

outperformed the other models, followed by the three ML models (RF, SVR, and SGB), the linear model based on PLSR performed the worst, indicating the non-linear relationship between yield and different variables. Static, satellite, and climate data together as inputs achieved the best performance, which explains 72%, 69%, and 57% of maize, rice, and soybean yield variability, respectively. For all the experiments, climate data only as input had the worst performance, explaining 53%, 44%, and 36% of maize, rice, and soybean yield variability, respectively. Satellite data only as input achieved much better performance than climate data, improved more than 10% of yield prediction ability compared to climate data, indicating satellite data with expanded spectral ranges can provide more information than climate variables to reflect crop growth. For maize and soybean yield prediction, the combined use of satellite and climate data as inputs achieved higher performance than the individual data, demonstrating unique value was provided from both the two datasets. Adding static variables to the satellite or climate data as inputs also showed better performance than individual input, while models with static plus satellite data as input outperformed models with static plus climate data, but slightly underperformed the model with all the variables, indicating that the static variables, climate data, and satellite data contain complementary information that is worth exploiting jointly, satellite data used in this study cannot replace climate data for maize and soybean yield prediction in northeast China. However, for rice yield prediction, even static variables improved the prediction performance by adding them to the individual input, we notice that almost equivalently performance was achieved between satellite data and climate + satellite data and between static + satellite data and static + climate + satellite data, indicating that climate data does not add extra contributions beyond satellite data for rice yield prediction.

4.3. Performance of multi-source satellite data in crop yield prediction

We further conducted the second group experiment using static plus satellite data with various spectrum bands from 2003 to 2018 (Fig. 5). The benchmark model with static plus EVI as input has a lower R^2 performance

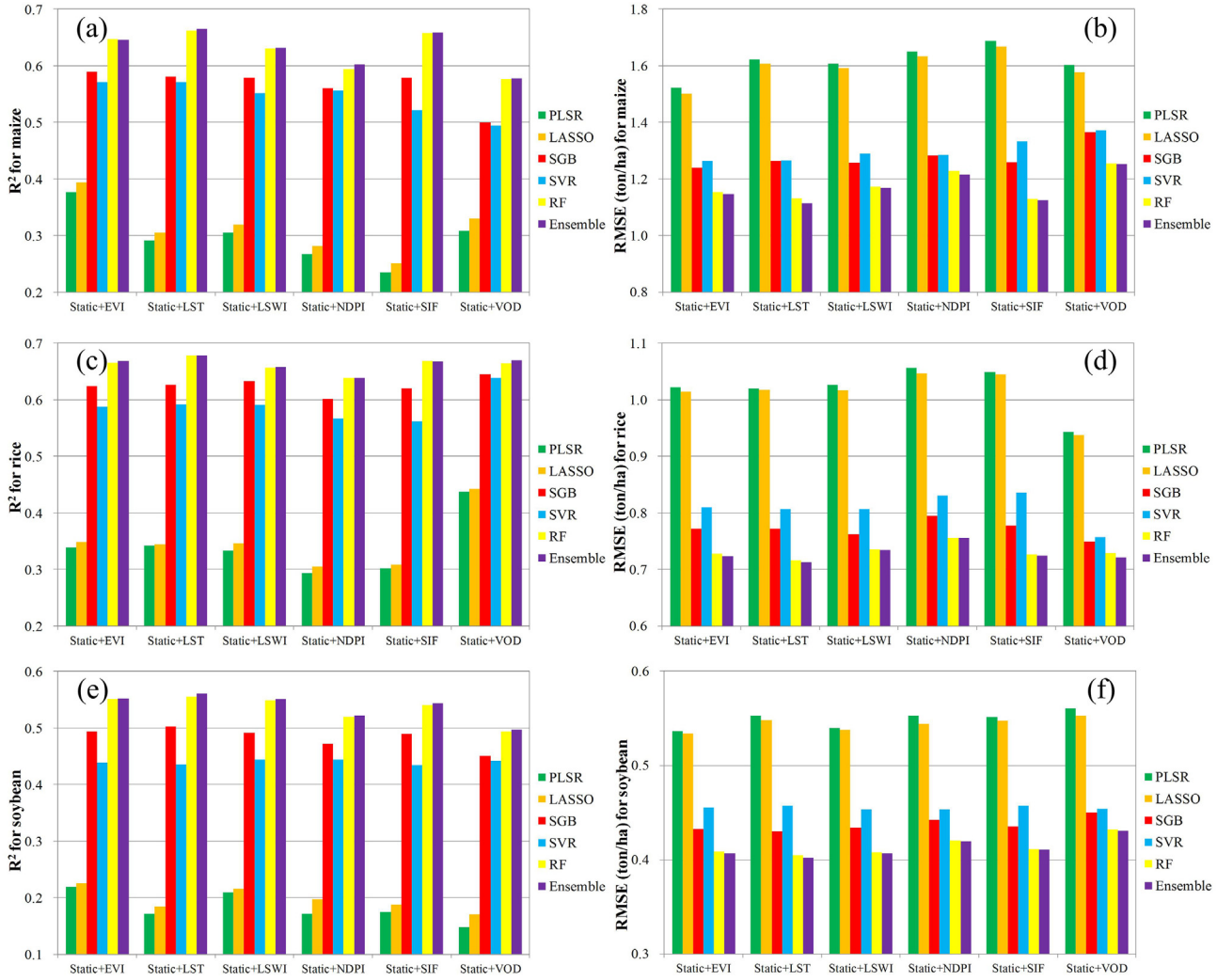


Fig. 5. The model performance (R^2 and RMSE) of maize (top), rice (middle) and soybean (bottom) yield prediction using the six methods and different satellite data for the whole growing season.

than that of static plus all the satellite variables, indicating that other satellite data beyond EVI provide added value in explaining crop yield variability. For maize yield prediction, static plus LST performed better than the other RS variables with the highest R^2, followed by static plus SIF, models with these two combination inputs all showed higher R^2s than the benchmark models (Fig. 5a). Static plus LSWI or NDPI or VOD achieved a worse performance than the benchmark model, static plus VOD has the worst performance with the lowest R^2. For rice yield prediction, static plus SIF, static plus LST, and static plus EVI showed equivalent performance, while static plus NDPI performed the worst with the lowest R^2 (Fig. 5b). For soybean yield prediction, static plus LST performed best than the other RS variables, followed by static plus LSWI, which performed better than static plus EVI, static plus VOD performed the worst (Fig. 5c).

To further explore which satellite variables beyond EVI provided unique value for crop yield prediction, we rerun the models by adding one or two satellite variables into the benchmark model, with results shown in Fig. 6. For maize, all the combinations with more satellite variables achieved better performance than the benchmark model (Fig. 6a), indicating each satellite variable beyond EVI can provide added value in explaining crop yield variability. Adding SIF to the benchmark model improved the prediction performance the most, the R^2 was improved from 0.65 to 0.70 for the ensemble model, the use of LST also improved the benchmark model obviously, with R^2 of 0.69 for the ensemble model. The combined use of EVI with NDPI and LSWI or VOD only improved the performance slightly, varying from 0.01 to 0.02 increases in the predicted R^2. For rice, the yield prediction performance was only improved by adding SIF or VOD into the benchmark model (Fig. 6b), although the improvement is slight with a 0.01 to 0.02 increase in the predicted R^2. Adding NDPI, LSWI, and LST into the benchmark model did not improve the prediction performance, indicating no unique information was provided to the benchmark model. For soybean, improvement of yield prediction was observed in the case of adding each satellite variable except for VOD into the benchmark model (Fig. 6c), adding LST to the benchmark model improved the prediction performance the most, the R^2 was improved from 0.54 to 0.57 for the ensemble model, no added value of VOD for soybean yield prediction was observed relative to the benchmark model.

4.4. Within-season forecasting performance

The third group experiment was conducted to analysis the within-season forecasting performance for the three crops at different months, the time series R^2 achieved by the ML ensemble model from April to October are shown in Fig. 7. In general, the model performed poorly during the early growing season. Along with crop growth and development, the prediction accuracy gradually increased as more information became available, and model performance became stable since July (for rice) and August (for maize and soybean) when crops transit from the vegetative stage to the reproductive stage. Moreover, crop yields can be satisfactorily

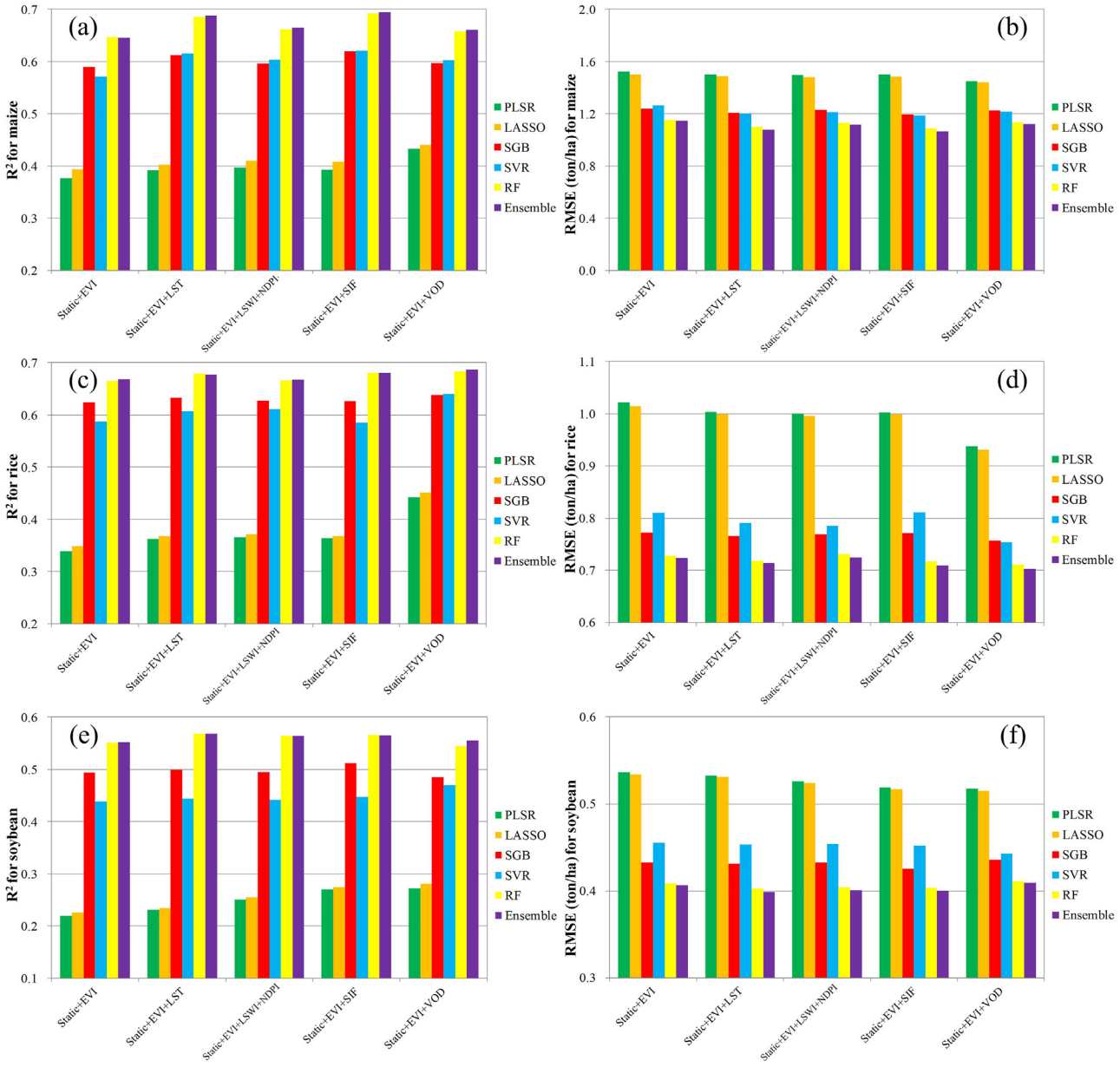


Fig. 6. The model performance (R^2 and RMSE) of maize (top), rice (middle) and soybean (bottom) yield prediction using the six methods and different combinations of satellite data for the whole growing season.

forecasted at around two to three months prior to harvest for maize (R^2 = 0.70, RMSE = 1.05 ton/ha), rice (R^2 = 0.69, RMSE = 0.70 ton/ha), and soybean (R^2 = 0.57, RMSE = 0.40 ton/ha).

4.5. Feature importance of predictor variables

The feature importance of each predictor variables was used to reflect the contribution of different predictors to predict yield. According to the Boruta algorithm, all predictor variables after exploratory data analysis were confirmed important to predict crop yield. The relative importance of the top 50 predictor variables for yield prediction of the three crops was shown in Fig. 8. In general, the longitude, latitude, DEM, EVI, and VOD in several months were ranked the top and identified more important than other variables, soil properties like GRAV and theta_s were also highly ranked, indicating their strong ability to explain crop yield variations. Yield prediction accuracy was less affected by precipitation and shortwave radiation in northeast China for their lowly-rank among all the variables. However, other variables explain yield varied for different crops. For maize, EVI in July and August, VOD in May and June were ranked the most important, followed by longitude, latitude, DEM and LSWI in August, climate variables were identified less important and ranked in the middle-latter portion of all the variables, Temp in September and October, Shum in October, and Srad in August were ranked more important than the other climate variables. SOM, CEC, TK, BD, and nighttime LST were also identified less important (Fig. 8a). For rice, VOD in May, June, July, and October, DEM, EVI in July were top-ranked, nighttime LST in May, NDPI in April, and soil nutrition indicators (TK, SOM) were also identified as important variables, climate variables were lowly ranked with Shum in October, Temp in July were more important than the other climate variables (Fig. 8b). For soybean, VOD and static variables were dominant in the top important variables, specifically, VOD in April, May, and June, longitude, latitude, DEM, soil texture (SI and CL), and soil nutrition indicators (TK, TP, and SOM) were highly ranked. Beyond that, EVI in July and August, SIF in July and daytime LST in August were also highly ranked. For climate

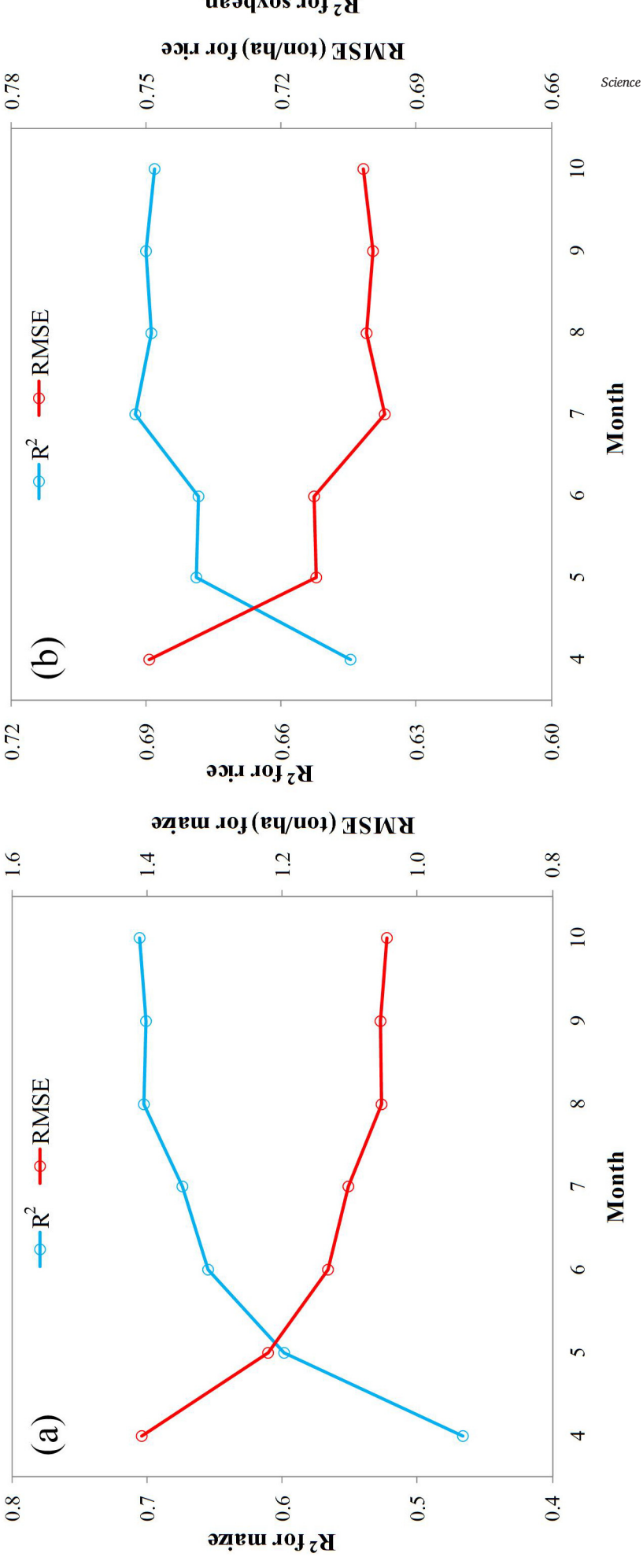


Fig. 7. Within-season forecasting performance for maize (a), rice (b), and soybean (c).

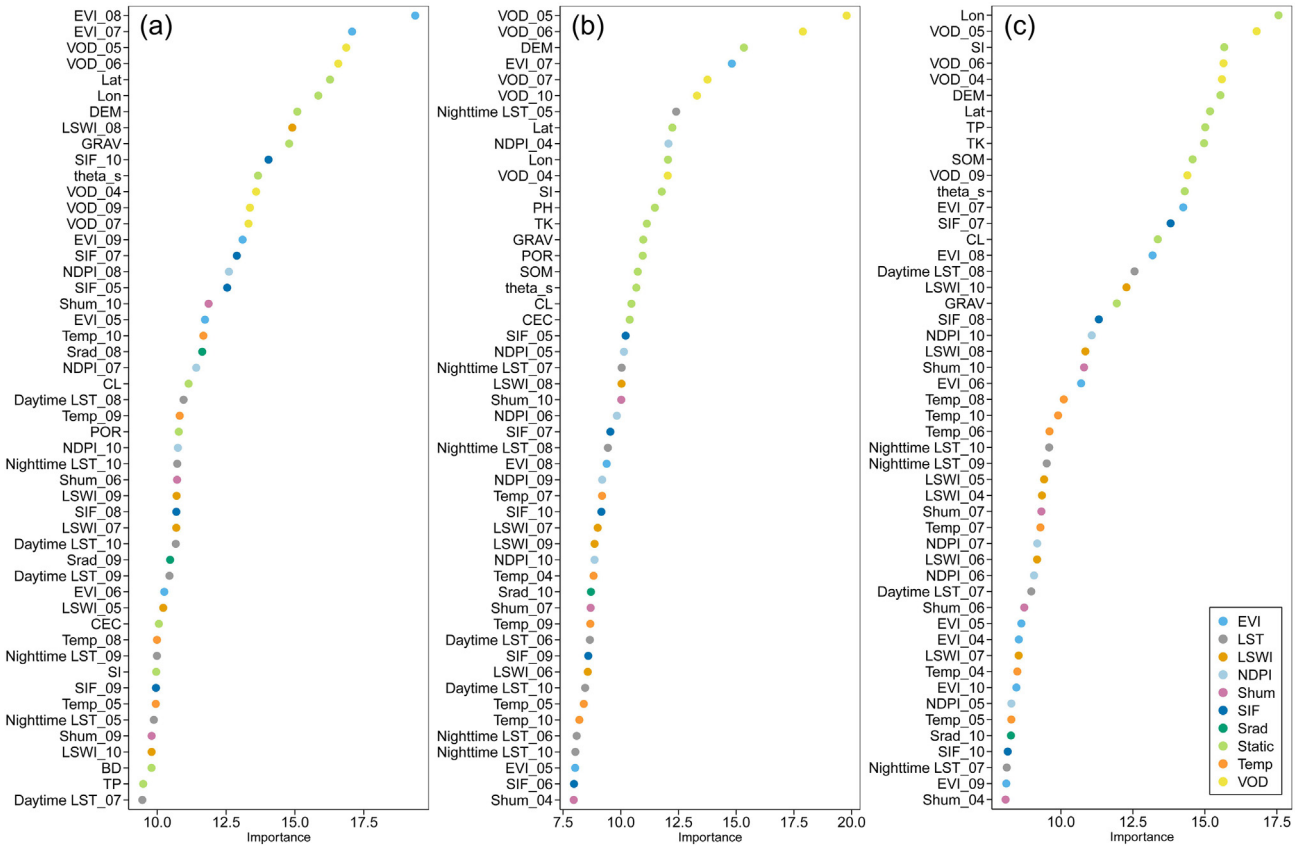


Fig. 8. The relative importance of top 50 predictor variables for yield prediction of maize (a), rice (b) and soybean (c). The importance of each variable was based on the z-scores of mean decrease accuracy from the Boruta algorithm.

variables, Shum in July and October, Temp in August, September, and October were more important than the other climate variables (Fig. 8c).

5. Discussion

5.1. Combining environmental and satellite data achieves best yield prediction

Our results showed that the combination of environmental (climate, soil, geography, and topography) and satellite data has achieved the best performance for predicting crop yield in northeast China, the combined use of satellite and climate data outperformed the individual input, indicating all the three types of data provided unique information to the final crop yield prediction. In the situation of using individual data sources, satellite data can achieve better predictive performance than the climate data alone, which is different from the previous studies (Cai et al., 2019). In our study, more satellite products covering a variety of spectrum were used, these data can provide more information beyond AGB that traditional visible-NIR based VIs provides, such as LST can reflect the water and heat stress of crop canopy (Johnson, 2014; Khanal et al., 2017), SIF is a good proxy of plant photosynthesis, VOD can reflect the soil and canopy water conditions in deeper layer (Chaparro et al., 2018; Konings et al., 2019; Tian et al., 2018). The Pearson's correlation and feature importance analysis showed that satellite data in the early-peak of the growing season were more important and better correlated with crop yield, while crop yield was better correlated with climate factors in the peak-late season (Figs. 3; 8). The information provided by satellite data in the early-peak season is an important indicator for crop photosynthesis and aboveground biomass, which reflects the accumulated historical climate effects as well as other biotic effects before and during the peak season, the newly carbon accumulated during these periods mostly goes to the grain, which makes the satellite data before and during the peak time highly correlated to final yield (Cai et al., 2019; Guan et al., 2017). After that, satellite data is less correlated with the yield due to the grain formation process and canopy senescence (Cai et al., 2019). In fact, crop yield ultimately depends on grain weight, a product of aboveground biomass and "Harvest Index" (HI), with the latter mainly determined by the middle-late growing season, the crop growth processes in these periods are highly sensitive to temperature and water stresses, which can be better reflected by climate factors than satellite data (Guan et al., 2017). This can explain the highly negative correlations between October shortwave radiation and temperature with maize yield, and the highly positive correlation between July specific humidity and July temperature with rice and soybean yield. Some other critical factors, such as biotic stress information, can be captured by soil property, geography and topography variables, which also contributed to explaining more yield variability. Soil properties can provide unique information which directly influence crop water uptake and nutrient uptake, such as saturated water content (theta_s) decides the water storage in the soil column and thus influences crop's resistance to drought stress (Li et al., 2019a), soil with high SOM, TP and TK provides nutrient-rich conditions which benefit crop growth (Kravchenko and Bullock, 2000; Ma et al., 2021). Other predictors like longitude, latitude, and DEM are good indicators of crop growth environmental conditions like photoperiod, climate spatial variations (Xie et al., 2015). Our results demonstrate the necessity of combining climate, satellite, and other environmental data for estimating crop yield.

5.2. The unique value of multi-sensor satellite data in crop yield prediction

Our study found that additional and unique information about crop growth can be revealed by other satellite data beyond widely used EVI. In the spectral ranges from visible to SWIR, VIs like LSWI and NDPI have similar seasonality congruent with the crop growth cycle of EVI (Fig. 2), the three VIs all shared information related to aboveground canopy biomass, combining the three VIs together as inputs just slightly improve the model performance compared with EVI only, thus little unique information beyond EVI was

provided for crop yield. Comparatively, SIF achieved a better performance than EVI for maize yield prediction and comparative predictive performance with EVI for soybean and rice yield prediction, combining the two sources has obviously improved the model performance compared with individual input for the three crops. SIF is the active emission from plant chlorophyll in its photo-machinery and has been used as a proxy of plant photosynthesis, it was proven to be more sensitive to short-term crop stresses induced from biotic and abiotic factors (Hao et al., 2021; Porcar-Castell et al., 2021), the fine resolution SIF product used in this study has provided additional information beyond EVI for crop yield prediction.

Among all the tested satellite data, LST had an overall best performance in predicting maize and soybean yield than the other satellite variables, combining EVI and LST as inputs also achieved a much better predictive performance than individual input. MODIS LST product is a direct measurement of canopy temperature, it is widely accepted as an indicator of field-level crop water stress. Healthy crops with sufficient water were expected to have a lower canopy temperature than the air temperature, as water becomes limited, evapotranspiration decreases and canopy temperature approaches air temperature, thus LST is superior to reflect canopy water stress and evaporative cooling (Pede et al., 2019). LST provided unique information beyond EVI for crop yield prediction.

Unlike optical VIs, which are related to the aboveground biomass, the microwave RS data-derived VOD is sensitive to the canopy structure, the living aboveground biomass and other plant hydraulic properties (volumetric water content) (Grant et al., 2016). In our case, VOD had a worse performance than EVI for yield prediction of the three crops, which might be caused by the coarse spatial resolution (0.25°) that could not capture small and localized crop features in space. However, VOD was highly ranked among all the RS data in the crop prediction models, combining EVI and VOD has improved the maize and rice yield prediction performance, indicating VOD revealed other unique information beyond EVI. Pearson's correlation showed that crop yield was highly correlated with VOD in the early growing season (May and June), the microwave VOD also showed a lagged seasonal cycle compared with the other RS metrics (Fig. 2), the C-band VOD product used in this study has a lower microwave frequency, it may contain more information from deeper canopy layers and surface soil moisture, especially prior to canopy closure (Mateo-Sanchis et al., 2019), indicating VOD provided useful independent information for crop yield prediction in our study.

5.3. Ensemble model improved the crop yield prediction performance

The use of a single model to predict crop yield has become popular, our results showed that ML methods (SGB, SVR, and RF) achieved better performance than the linear regression (PLSR and LASSO), which may be attributed to the non-linear relationships between climate, satellite variables and crop yield that documented by previous researches (Azzari et al., 2017; Lobell, 2013), ML methods are well performed in capturing the potential complex and nonlinear relationships between input variables and crop yields rather than the linear regression models (Cao et al., 2021). However, the single model will inevitable be under-fitted or over-fitted (Fang et al., 2021). Our study showed that the single ML model can be outperformed by the ML ensemble model. Ensemble models are proved to be effective in reducing the uncertainty of model fitting and improving the generality and robustness compared to single models by combining the prediction results of several single models (Pham and Olafsson, 2019; Schwenker, 2013; Shahhosseini et al., 2020a; Shahhosseini et al., 2020b). Thus compared to traditional yield prediction methods, the ensemble method provides new opportunities for yield predictions over a large area, future studies are still needed to explore more robust ensemble models by integrating more ML or deep learning models and using robust base learners.

5.4. Study limitations and future directions

This study highlighted the values of environmental and multi-source satellite data for crop yield estimation in northeast China using linear regression and ML models. Nevertheless, there are some limitations and prospects which should be addressed in the future study. Firstly, the current crop maps of maize, rice, and soybean are static for all the years from 2003 to 2018. In fact, the crop planting area changes from year to year, and a static crop map can lead to errors when we extract the area for the climate and satellite data in modeling yield. Future studies should use crop type maps updated annually to reduce the potential errors. Secondly, it would be useful in the future to add more climate and satellite variables to the yield prediction model as they may provide complementary information to further improve the crop yield prediction, such as agro-climatic indices obtained from direct weather data (Jiang et al., 2020; Mathieu and Aires, 2018), newly developed satellite products like near-infrared reflectance of vegetation (NIRV), which will provide complementary information than satellite variables that are used here (Badgley et al., 2017; Guan et al., 2017; Peng et al., 2020), and microwave-based soil moisture products (e.g. European Space Agency Climate Change Initiative (CCI), AMSR-E/2, Soil Moisture and Ocean Salinity (SMOS), Soil Moisture Active Passive (SMAP), and Chinese FengYun-3) were also found to be important for crop yield prediction throughout the growing season (Li et al., 2021; Zeng et al., 2020; Zeng et al., 2015). Thirdly, most of the environmental and satellite datasets used in this study are only available at relatively coarse spatial resolution ($\geq$ 5 km), which might bring errors to the yield forecasting models due to the inability of these sensors to resolve biomass and yield characteristics of the heterogeneous croplands. Future studies may use satellite sensors with a higher spatiotemporal resolution including satellite sensors in optical (e.g. Landsat, Sentinel-2, Fluorescence Explorer (Drusch et al., 2017)) and microwave synthetic aperture radar (e.g. Sentinel-1), which may provide more effective information in local areas. Moreover, fusing fine temporal resolution but coarse spatial resolution satellite data (e.g. MODIS, Advanced Very High Resolution Radiometer (AVHRR)) with fine spatial resolution but coarse temporal resolution satellite data (e.g. Landsat, Sentinel) into high spatial-temporal satellite data is also an alternative strategy for fine-scale crop monitoring and yield estimation (Gao et al., 2017; Li et al., 2018).

6. Conclusions

In this study, we investigated the relative performances of environmental and multi-source satellite data for yield forecasting of three major crops in northeast China, two linear regression approaches, three ML approaches, and one ML ensemble method were used to build yield prediction models with different combinations of input variables. Our study showed that, overall, the ensemble mode outperformed the regression and ML models, and better performance is achieved with more input data, integrating climate, satellite, soil, geography, and topography data achieved the best performance for maize and soybean yield prediction, while combining satellite, soil, geography and topography data as input performed equivalently with all data included and performed better than the other combinations for rice yield prediction. Specifically, satellite data mainly contribute to yield prediction in the early-peak growing season, while climate data contribute more to yield prediction at peak-late growing season. For the multi-source satellite data, the VIs that cover spectral ranges from visible to SWIR share similar information related to aboveground biomass, the combined use of these VIs did not obviously improve yield prediction performance compared to the EVI only model. Instead, the SIF, LST, and VOD products can provide unique information rather than aboveground biomass, improvements of model predictive skills for crop yield at the county level were observed compared to the EVI only model, thus we suggest to continue taking advantage of the visible-NIR VI record, while incorporating other satellite data from a wide spectral range for cropland monitoring. The within-season yield forecasting showed that, with crop growth progression and more information became available, crop yields can be satisfactorily forecasted at two to three months prior to harvest. Geography, topography, VOD, EVI, and soil hydraulic and nutrient parameters were identified as important

covariates. This study provided key information to identify useful predictor variables for capturing the spatiotemporal variability of crop yield, which provides a path to fully integrating these data to develop robust yield forecasting systems.

CRediT authorship contribution statement

Zhenwang Li: Conceptualization, Methodology, Writing – original draft. **Lei Ding:** Data curation, Writing – original draft. **Dawei Xu:** Data curation, Resources.

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.

Acknowledgements

This work was supported by the National Key Research and Development Program of China (No. 2018YFE0107000) and Key Research and Development Program of Shandong Province (2019JZZY010713).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.scitotenv.2021.152880.

References

Azzari, G., Jain, M., Lobell, D.B., 2017. Towards fine resolution global maps of crop yields: testing multiple methods and satellites in three countries. Remote Sens. Environ. 202, 129–141.
Badgley, G., Field, C.B., Berry, J.A., 2017. Canopy near-infrared reflectance and terrestrial photosynthesis. Sci. Adv. 3, e1602244.
Cai, Y., Guan, K., Lobell, D., Potgieter, A.B., Wang, S., Peng, J., et al., 2019. Integrating satellite and climate data to predict wheat yield in Australia using machine learning approaches. Agric. For. Meteorol. 274, 144–159.
Cao, J., Zhang, Z., Tao, F., Zhang, L., Luo, Y., Zhang, J., et al., 2021. Integrating multi-source data for rice yield prediction across China using machine learning and deep learning approaches. Agric. For. Meteorol. 297, 108275.
Chaparro, D., Piles, M., Vall-llossera, M., Camps, A., Konings, A.G., Entekhabi, D., 2018. L-band vegetation optical depth seasonal metrics for crop yield assessment. Remote Sens. Environ. 212, 249–259.
Chen, Y., Yang, K., He, J., Qin, J., Shi, J., Du, J., et al., 2011. Improving land surface temperature modeling for dry land of China. J. Geophys. Res. Atmos. 116.
Chipanshi, A., Zhang, Y., Kouadio, L., Newlands, N., Davidson, A., Hill, H., et al., 2015. Evaluation of the integrated Canadian crop yield forecaster (ICCYF) model for in-season prediction of crop yield across the Canadian agricultural landscape. Agric. For. Meteorol. 206, 137–150.
Dai, Y., Shangguan, W., Duan, Q., Liu, B., Fu, S., Niu, G., 2013. Development of a China dataset of soil hydraulic parameters using pedotransfer functions for land surface modeling. J. Hydrometeorol. 14, 869–887.
Dong, J., Xiao, X., Wagle, P., Zhang, G., Zhou, Y., Jin, C., et al., 2015. Comparison of four EVI-based models for estimating gross primary production of maize and soybean croplands and tallgrass prairie under severe drought. Remote Sens. Environ. 162, 154–168.
Drusch, M., Moreno, J., Bello, U.D., Franco, R., Goulas, Y., Huth, A., et al., 2017. The FLuorescence EXplorer Mission Concept—ESA's Earth Explorer 8. IEEE Transactions on Geoscience and Remote Sensing 55, 1273–1284.
Fang, Y., Zhang, X., Wei, H., Wang, D., Chen, R., Wang, L., et al., 2021. Predicting the invasive trend of exotic plants in China based on the ensemble model under climate change: a case for three invasive plants of Asteraceae. Sci. Total Environ. 756, 143841.
Franz, T.E., Pokal, S., Gibson, J.P., Zhou, Y., Gholizadeh, H., Tenorio, F.A., et al., 2020. The role of topography, soil, and remotely sensed vegetation condition towards predicting crop yield. Field Crop Res. 252, 107788.
Gao, F., Anderson, M.C., Zhang, X., Yang, Z., Alfieri, J.G., Kustas, W.P., et al., 2017. Toward mapping crop progress at field scales through fusion of Landsat and MODIS imagery. Remote Sens. Environ. 188, 9–25.
Grant, J.P., Wigneron, J.P., De Jeu, R.A.M., Lawrence, H., Mialon, A., Richaume, P., et al., 2016. Comparison of SMOS and AMSR-E vegetation optical depth to four MODIS-based vegetation indices. Remote Sens. Environ. 172, 87–100.
Guan, K., Wu, J., Kimball, J.S., Anderson, M.C., Frolking, S., Li, B., et al., 2017. The shared and unique values of optical, fluorescence, thermal and microwave satellite data for estimating large-scale crop yields. Remote Sens. Environ. 199, 333–349.
Guanter, L., Zhang, Y., Jung, M., Joiner, J., Voigt, M., Berry, J.A., et al., 2014. Global and time-resolved monitoring of crop photosynthesis with chlorophyll fluorescence. Proc. Natl. Acad. Sci. 111, E1327–E1333.
Hao, D., Asrar, G.R., Zeng, Y., Yang, X., Li, X., Xiao, J., et al., 2021. Potential of hotspot solar-induced chlorophyll fluorescence for better tracking terrestrial photosynthesis. Glob. Chang. Biol. 27, 2144–2158.
He, J., Yang, K., Tang, W., Lu, H., Qin, J., Chen, Y., et al., 2020. The first high-resolution meteorological forcing dataset for land process studies over China. Sci. Data 7, 25.
Heft-Neal, S., Lobell, D.B., Burke, M., 2017. Using remotely sensed temperature to estimate climate response functions. Environ. Res. Lett. 12, 014013.
Hoffman, L.A., Etienne, X.L., Irwin, S.H., Colino, E.V., Toasa, J.I., 2015. Forecast performance of WASDE price projections for U.S. corn. Agric. Econ. 46, 157–171.
Hu, Q., Yin, H., Friedl, M.A., You, L., Li, Z., Tang, H., et al., 2021. Integrating coarse-resolution images and agricultural statistics to generate sub-pixel crop type maps and reconciled area estimates. Remote Sens. Environ. 258, 112365.
Huete, A., Didan, K., Miura, T., Rodriguez, E.P., Gao, X., Ferreira, L.G., 2002. Overview of the radiometric and biophysical performance of the MODIS vegetation indices. Remote Sens. Environ. 83, 195–213.
Jackson, T.J., Schmugge, T.J., 1991. Vegetation effects on the microwave emission of soils. Remote Sens. Environ. 36, 203–212.
Jeong, S., Ko, J., Yeom, J.-M., 2022. Predicting rice yield at pixel scale through synthetic use of crop and deep learning models with satellite data in South and North Korea. Sci. Total Environ. 802, 149726.
Jiang, H., Hu, H., Zhong, R., Xu, J., Xu, J., Huang, J., et al., 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.
Johnson, M.D., Hsieh, W.W., Cannon, A.J., Davidson, A., Bédard, F., 2016. Crop yield forecasting on the Canadian prairies by remotely sensed vegetation indices and machine learning methods. Agric. For. Meteorol. 218–219, 74–84.
Jones, J.W., Antle, J.M., Basso, B., Boote, K.J., Conant, R.T., Foster, I., et al., 2017. Brief history of agricultural systems modeling. Agric. Syst. 155, 240–254.
Kang, Y., Özdoğan, M., 2019. Field-level crop yield mapping with Landsat using a hierarchical data assimilation approach. Remote Sens. Environ. 228, 144–163.
Khanal, S., Fulton, J., Shearer, S., 2017. An overview of current and potential applications of thermal remote sensing in precision agriculture. Comput. Electron. Agric. 139, 22–32.
Konings, A.G., Rao, K., Steele-Dunne, S.C., 2019. Macro to micro: microwave remote sensing of plant water content for physiology and ecology. New Phytol. 223, 1166–1172.
Kravchenko, A.N., Bullock, D.G., 2000. Correlation of corn and soybean grain yield with topography and soil properties. Agron. J. 92, 75–83.
Kursa, M.B., Rudnicki, W.R., 2010. Feature selection with the Boruta package. J. Stat. Softw. 36, 13.
Lecerf, R., Ceglar, A., López-Lozano, R., Van Der Velde, M., Baruth, B., 2019. Assessing the information in crop model and meteorological indicators to forecast crop yield over Europe. Agric. Syst. 168, 191–202.
Li, Z., Huang, C., Zhu, Z., Gao, F., Tang, H., Xin, X., et al., 2018. Mapping daily leaf area index at 30 m resolution over a meadow steppe area by fusing Landsat, Sentinel-2A and MODIS data. Int. J. Remote Sens. 39, 9025–9053.
Li, Y., Guan, K., Schnitkey, G.D., DeLucia, E., Peng, B., 2019a. Excessive rainfall leads to maize yield loss of a comparable magnitude to extreme drought in the United States. Glob. Chang. Biol. 25, 2325–2337.
Li, Y., Guan, K., Yu, A., Peng, B., Zhao, L., Li, B., et al., 2019b. Toward building a transparent statistical model for improving crop yield prediction: modeling rainfed corn in the U.S. Field Crop Res. 234, 55–65.
Li, L., Wang, B., Feng, P., Wang, H., He, Q., Wang, Y., et al., 2021. Crop yield forecasting and associated optimum lead time analysis based on multi-source environmental data across China. Agric. For. Meteorol. 308–309, 108558.
Lichtenthaler, H.K., 1996. Vegetation stress: an introduction to the stress concept in plants. J. Plant Physiol. 148, 4–14.
Liu, J.G., Xie, Z.H., 2013. Improving simulation of soil moisture in China using a multiple meteorological forcing ensemble approach. Hydrol. Earth Syst. Sci. 17, 3355–3369.
Liu, Y.Y., van Dijk, A.I.J.M., de Jeu, R.A.M., Canadell, J.G., McCabe, M.F., Evans, J.P., et al., 2015. Recent reversal in loss of global terrestrial biomass. Nat. Clim. Chang. 5, 470–474.
Lobell, D.B., 2013. The use of satellite data for crop yield gap analysis. Field Crop Res. 143, 56–64.
Ma, Y., Zhang, Z., Kang, Y., Özdoğan, M., 2021. Corn yield prediction and uncertainty analysis based on remotely sensed variables using a Bayesian neural network approach. Remote Sens. Environ. 259, 112408.
Mahlein, A.-K., Oerke, E.-C., Steiner, U., Dehne, H.-W., 2012. Recent advances in sensing plant diseases for precision crop protection. Eur. J. Plant Pathol. 133, 197–209.
Mateo-Sanchis, A., Piles, M., Muñoz-Marí, J., Adsuara, J.E., Pérez-Suay, A., Camps-Valls, G., 2019. Synergistic integration of optical and microwave satellite data for crop yield estimation. Remote Sens. Environ. 234, 111460.
Mathieu, J.A., Aires, F., 2018. Assessment of the agro-climatic indices to improve crop yield forecasting. Agric. For. Meteorol. 253–254, 15–30.
Moesinger, L., Dorigo, W., de Jeu, R., van der Schalie, R., Scanlon, T., Teubner, I., et al., 2020. The global long-term microwave vegetation optical depth climate archive (VODCA). Earth Syst. Sci. Data 12, 177–196.
Momen, M., Wood, J.D., Novick, K.A., Pangle, R., Pockman, W.T., McDowell, N.G., et al., 2017. Interacting effects of leaf water potential and biomass on vegetation optical depth. J. Geophys. Res. Biogeosci. 122, 3031–3046.
Owe, M., de Jeu, R., Holmes, T., 2008. Multisensor historical climatology of satellite-derived global land surface moisture. J. Geophys. Res. Earth Surf. 113.

Pede, T., Mountrakis, G., Shaw, S.B., 2019. Improving corn yield prediction across the US Corn Belt by replacing air temperature with daily MODIS land surface temperature. Agric. For. Meteorol. 276–277, 107615.
Peng, B., Guan, K., Zhou, W., Jiang, C., Frankenberg, C., Sun, Y., et al., 2020. Assessing the benefit of satellite-based solar-induced chlorophyll fluorescence in crop yield prediction. Int. J. Appl. Earth Obs. Geoinf. 90, 102126.
Pham, H., Olafsson, S., 2019. Bagged ensembles with tunable parameters. Comput. Intell. 35, 184–203.
Porcar-Castell, A., Tyystjärvi, E., Atherton, J., van der Tol, C., Flexas, J., Pfündel, E.E., et al., 2014. Linking chlorophyll a fluorescence to photosynthesis for remote sensing applications: mechanisms and challenges. J. Exp. Bot. 65, 4065–4095.
Porcar-Castell, A., Malenovský, Z., Magney, T., Van Wittenberghe, S., Fernández-Marín, B., Maignan, F., et al., 2021. Chlorophyll a fluorescence illuminates a path connecting plant molecular biology to earth-system science. Nat. Plants 7, 998–1009.
Rosenzweig, C., Elliott, J., Deryng, D., Ruane, A.C., Müller, C., Arneth, A., et al., 2014. Assessing agricultural risks of climate change in the 21st century in a global gridded crop model intercomparison. Proc. Natl. Acad. Sci. 111, 3268–3273.
Sakamoto, T., Gitelson, A.A., Arkebauer, T.J., 2013. MODIS-based corn grain yield estimation model incorporating crop phenology information. Remote Sens. Environ. 131, 215–231.
Schauberger, B., Jägermeyr, J., Gornott, C., 2020. A systematic review of local to regional yield forecasting approaches and frequently used data resources. Eur. J. Agron. 120, 126153.
Schwenker, F., 2013. Ensemble Methods: Foundations and Algorithms [Book Review]. IEEE Computational Intelligence Magazine 8, 77–79.
Shahhosseini, M., Hu, G., Archontoulis, S.V., 2020;. Forecasting corn yield with machine learning ensembles. FrontiersPlant Sci. 11.
Shahhosseini, M., Hu, G., Pham, H., 2020. Optimizing Ensemble Weights for Machine Learning Models: A Case Study for Housing Price Prediction. Springer International Publishing, Cham, pp. 87–97.
Shangguan, W., Dai, Y., Liu, B., Zhu, A., Duan, Q., Wu, L., et al., 2013. A China data set of soil properties for land surface modeling. J. Adv. Model. Earth Syst. 5, 212–224.
Sherrick, B.J., Lanoue, C.A., Woodard, J., Schnitkey, G.D., Paulson, N.D., 2014. Crop yield distributions: fit, efficiency, and performance. Agricultural Finance Review 74, 348–363.
Sibley, A.M., Grassini, P., Thomas, N.E., Cassman, K.G., Lobell, D.B., 2014. Testing remote sensing approaches for assessing yield variability among maize fields. Agron. J. 106, 24–32.
Tian, F., Wigneron, J.-P., Ciais, P., Chave, J., Ogée, J., Peñuelas, J., et al., 2018. Coupling of ecosystem-scale plant water storage and leaf phenology observed by satellite. Nat. Ecol. Evol. 2, 1428–1435.
Vreugdenhil, M., Dorigo, W.A., Wagner, W., RAMd, Jeu, Hahn, S., MJEv, Marle, 2016. Analyzing the vegetation parameterization in the TU-Wien ASCAT soil moisture retrieval. IEEE Transactions on Geoscience and Remote Sensing 54, 3513–3531.
Wan, Z., Hook, S., Hulley, G., 2021. MODIS/Aqua Land Surface Temperature/Emissivity Monthly L3 Global 6km SIN Grid V061.
Wang, C., Chen, J., Wu, J., Tang, Y., Shi, P., Black, T.A., et al., 2017. A snow-free vegetation index for improved monitoring of vegetation spring green-up date in deciduous ecosystems. Remote Sens. Environ. 196, 1–12.
Wen, J., Köhler, P., Duveiller, G., Parazoo, N.C., Magney, T.S., Hooker, G., et al., 2020. A framework for harmonizing multiple satellite instruments to generate a long-term global high spatial-resolution solar-induced chlorophyll fluorescence (SIF). Remote Sens. Environ. 239, 111644.
Xiao, X., Boles, S., Frolking, S., Li, C., Babu, J.Y., Salas, W., et al., 2006. Mapping paddy rice agriculture in South and Southeast Asia using multi-temporal MODIS images. Remote Sens. Environ. 100, 95–113.
Xie, Y., Wang, X., Silander Jr., J.A., 2015. Deciduous forest responses to temperature, precipitation, and drought imply complex climate change impacts. Proc. Natl. Acad. Sci. U. S. A. 112, 13585–13590.
Xu, D., Wang, C., Chen, J., Shen, M., Shen, B., Yan, R., et al., 2021. The superiority of the normalized difference phenology index (NDPI) for estimating grassland aboveground fresh biomass. Remote Sens. Environ. 264, 112578.
You, N., Dong, J., Huang, J., Du, G., Zhang, G., He, Y., et al., 2021. The 10-m crop type maps in Northeast China during 2017–2019. Sci. Data 8, 41.
Zeng, J., Li, Z., Chen, Q., Bi, H., Qiu, J., Zou, P., 2015. Evaluation of remotely sensed and reanalysis soil moisture products over the Tibetan Plateau using in-situ observations. Remote Sens. Environ. 163, 91–110.
Zeng, J., Chen, K., Cui, C., Bai, X., 2020. A physically based soil moisture index from passive microwave brightness temperatures for soil moisture variation monitoring. IEEE Trans. Geosci. Remote Sens. 58, 2782–2795.