Downscaling Administrative-Level Crop Yield Statistics to 1 km Grids Using Multisource Remote Sensing Data and Ensemble Machine Learning

Jie Pei, Yaopeng Zou, Yibo Liu, Yinan He, Shaofeng Tan, Tianxing Wang, *Member, IEEE*, and Jianxi Huang, *Senior Member, IEEE*

***Abstract*—The United States (U.S.) is a global leader in the production and exportation of soybeans and corn. Accurate monitoring and estimation of soybean and corn yields in the U.S. is essential for improving global food security. However, there is currently a lack of publicly available spatial distribution datasets with high temporal and spatial resolution for U.S. corn and soybean yields, which hampers related research and policy-making. Therefore, in this study, we proposed a statistical downscaling framework to produce spatially explicit crop yield estimates by utilizing multisource environmental covariates and ensemble machine learning methods. We produced distribution maps of soybean and corn yields in the U.S. from 2006 to 2021 at a 1-km resolution through the optimal Cubist model, resulting in the USASoy&CornYield1km dataset. The results demonstrated stable accuracy, with R^2 values for corn ranging from 0.70 to 0.89 (average of 0.80) and for soybeans ranging from 0.74 to 0.90 (average of 0.81) during the period 2006–2021. Comparison with the spatial production allocation model (SPAM) dataset further confirmed the reliability of this dataset, with correlations of 0.84 for soybeans and 0.78 for corn when compared to SPAM2010. Spatial uncertainty analysis showed that the yield estimation uncertainty was 14.04% for soybeans and 20.49% for corn, indicating a generally low level of uncertainty. Overall, the USASoy&CornYield1km dataset offers higher spatial and temporal resolution, captures yield variations within counties, and covers a long time span. This study provides significant insights for analyzing U.S. soybean and corn yields and improving agricultural production.**

***Index Terms*—Corn, crop yield estimation, cubist model, multisource data, soybeans.**

Manuscript received 29 May 2024; revised 24 July 2024; accepted 6 August 2024. Date of publication 9 August 2024; date of current version 26 August 2024. This work was supported in part by the "Unveiling the List of Hanging" Science and Technology Project of Jinggangshan Agricultural High-tech Industrial Demonstration Zone under Grant 20222-051244, and in part by the Guangdong Basic and Applied Basic Research Foundation under Grant 2021A1515110442. *(Corresponding author: Jianxi Huang.)*

Jie Pei, Yaopeng Zou, Yibo Liu, Shaofeng Tan, and Tianxing Wang are with the School of Geospatial Engineering and Science, Sun Yat-sen University, Zhuhai 519082, China (e-mail: peij5@mail.sysu.edu.cn; zouyp7@mail2.sysu.edu.cn; liuyb59@mail2.sysu.edu.cn; tanshf6@mail2.sysu.edu.cn; wangtx23@mail.sysu.edu.cn).

Yinan He is with the Earth and Environmental Sciences Area, Lawrence Berkeley National Laboratory, Berkeley, CA 94720-8126 USA (e-mail: yinan.he@lbl.gov).

Jianxi Huang is with the College of Land Science and Technology, China Agricultural University, Beijing 100107, China, and also with the Key Laboratory of Remote Sensing for Agri-Hazards, Ministry of Agriculture and Rural Affairs, Beijing 100083, China (e-mail: jxhuang@cau.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2024.3441252

I. Introduction

CORN, as a crop used for food, feed, and industrial purposes, ranks as the third-largest cereal crop globally, following rice and wheat. Soybeans, on the other hand, represent the world's largest source of feed protein and the second-largest source of vegetable oil [1], [2]. The United States (U.S.) is the leading producer and exporter of both corn and soybeans. In 2020, it produced 115 million tons of soybeans and 358 million tons of corn, accounting for over 30% of the world's soybean and corn output [3], [4], [5]. The production of soybeans and corn in the U.S. has a significant impact on the global agricultural market. However, the growth of these crops is influenced by several factors, including global warming and extreme weather conditions, which significantly affect their development [6], [7], [8]. Therefore, accurately and timely estimating corn and soybean yield in the U.S. is of great importance for agricultural production, food security, and international food trade [9], [10].

Numerous scholars have conducted extensive research on crop yields in the United States. For instance, Jiang et al. [11] developed a long short-term memory model, incorporating multisource data to predict corn yield in the U.S. Corn Belt, achieving robust county-level crop yield estimates. Johnson et al. [12] compared the effectiveness of remote sensing vegetation index-based models and simple trend analysis in estimating county-level crop yields. They found that remote sensing models have advantages in predicting corn yields. Li et al. [3] used the extreme gradient boosting (XGBoost) model and multidimensional feature engineering to develop a framework for predicting county-level soybean yields, achieving higher prediction accuracy compared to other models. However, these methods are limited to county-level regional scales for yield predictions, lacking the capability to conduct spatial analysis of crop yields within counties or observe the continuous geographical variation of crop yields. Recent advancements have demonstrated the feasibility to conduct more detailed spatial analyses. For example, Zhang et al. [13] presented a machine learning framework that successfully predicted the spatial distribution of crop planting at a detailed (30 m) resolution using historical cropland data layer (CDL) maps. However, their study only focused on the spatial distribution of crops and did not involve yield prediction. As a result, important intracounty variations in crop productivity are missed, leading to less precise agricultural management and

suboptimal resource allocation. This limitation hinders efforts to identify specific areas of high or low productivity, ultimately affecting decisions related to fertilizer application, irrigation, and other critical farming practices.

There are some publicly available global crop yield datasets currently. For example, the global gridded crop model intercomparison (GGCMI) phase 1, completed by a collaboration of multiple crop model groups globally [14], offers a dataset at a spatial resolution of 0.5 arc-degree longitude and latitude. The global dataset of historical yields of major crops (GDHY), produced based on agricultural census statistics and satellite remote sensing data [15], presents another resource. The spatial production allocation model (SPAM), which enhances and improves existing crop downscaling models [16], and the global agro-ecological zones model version 4 (GAEZ 2010), created by integrating global climate, soil, terrain, and land cover data based on the United Nations Food and Agriculture Organization crop yield data [17], are also notable contributions. However, due to different research objectives and technical method limitations, these datasets generally have a coarse spatial resolution and limited time span. For instance, the spatial resolution of GAEZ and SPAM is approximately 10 km, while that of GDHY and GGCMI is about 55 km. GDHY and GGCMI offer annual time resolution, SPAM is updated every five years, and GAEZ was produced only around 2010. Such spatial resolutions and time spans limit the application of these data in small-scale spatial analysis of crop yields and the ability to analyze the dynamic changes in crop yield and distribution. To the best of our knowledge, there are currently very few publicly available crop yield datasets that combine high spatial resolution, a long time span, and dynamic spatial distribution data of crop yields.

The rapid development of remote sensing technology has made it feasible to produce high spatial resolution crop yield datasets. Satellite-based remote sensing observations, offering long-term and wide-ranging information on crop growth, have been used by numerous studies for crop yield estimation over the past decades. Lobell et al. [18] developed a satellite-based scalable crop yield mapper approach to provide gridded yield data at a regional scale. Cheng et al. [19] generated and evaluated a 1 km resolution, long-time span dataset of corn and wheat crop yields in China using multiple remote sensing indices and the random forest (RF) model. Zhao et al. [20] developed the ChinaWheatYield30m, a high-resolution dataset of annual winter wheat yields in China, by integrating satellite observations with meteorological data using a hierarchical linear model. Wu et al. [21] created the AsiaRiceYield4 km dataset by inputting satellite-derived soil and climate data, along with vegetation parameters, into a RF model, estimating seasonal rice yields across Asia. Additionally, the application of various machine learning models in the field of crop yield estimation has achieved satisfactory results [22], [23], [24], [25]. Compared to traditional process-based crop models and statistical regression methods, machine learning models, despite relying on large amounts of training data and often lacking interpretability [26], allow for the free selection of input variables without complex parameters [27] and offer higher efficiency and spatial generalization capabilities. Therefore, the combination of multisource remote sensing data with machine learning algorithms presents a promising opportunity for large-scale, high-precision agricultural crop yield prediction.

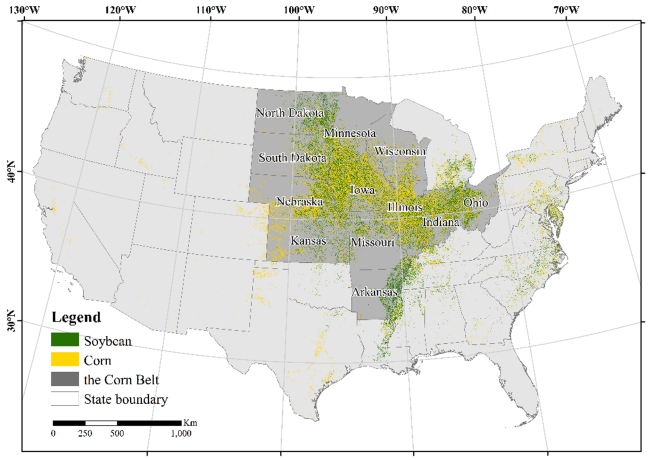

Fig. 1. Distribution of soybean and corn within the study area.

In this study, we proposed a methodological framework for generating gridded crop yield estimates for soybean and corn by leveraging multisource remote sensing data and ensemble machine learning models. The U.S. Corn Belt region was selected as the study area due to its fundamental importance in food production. We developed a spatial distribution dataset of crop yields at 1 km resolution, with annual intervals from 2006 to 2021 (USASoy&CornYield1km). This dataset offers more precise regional yield information and captures dynamic changes in annual yields and the geographical distribution of these crops, which is essential for understanding agricultural production patterns, guiding agricultural policy-making, and enhancing food security.

II. Data and Methods

A. Study Area

Our study area was the Corn Belt of the U.S. (see Fig. 1), which is located in the central and northern U.S. and includes 12 states, namely Arkansas, Illinois, Indiana, Iowa, Kansas, Minnesota, Missouri, Nebraska, North Dakota, Ohio, South Dakota, and Wisconsin. In more recent years, these states have averaged more than 84% and 85% of U.S. soybean and corn production, respectively [28]. Most of the region is characterized by a temperate continental climate, with significant spatial heterogeneity in crop yields among the states.

B. Data

In this study, county-level yield data from official statistics and annual crop area maps were collected. In addition, we gathered four categories of remote sensing variables for soybean and corn yield estimations, including vegetation indices, climate data, soil data, and elevation. All of these variables are accessible through the Google Earth Engine (GEE) to enhance the generalizability and transferability of the research method. All variables were standardized to the WGS84 coordinate system and resampled

TABLE I
SUMMARY OF THE DATASETS USED IN THIS STUDY

Data type	Content	Spatial resolution	Temporal resolution	Data usage	Data sources
Cropland data layer	Classification of corn and soybean	30 m	Yearly	Crop mask	USDA NASS
Census yield data	Statistical data	County-level	Yearly	Model construction and assessment	USDA NASS
Vegetation index	EVI NDVI	1 km	16-day	Input variables	MYD13A2
Climate data	ET PET	500 m	8-day	Input variables	MOD16A2
	LST_Day LST_Night	1 km	8-day	Input variables	MOD11A2
	PPT Tmean VPDmin VPDmax	4 km	Monthly	Input variables	PRISM
	PDSI	4 km	5-day	Input variables	GRIDMET
Soil data	volumetric_soil_water_layer soil_temperature_level	0.1°	Monthly	Input variables	ECMWF
Elevation	DEM	30 m	–	Input variables	NASADEM
Yield spatial dataset	SPAM	5 arcmin	–	Dataset comparison	Yu et al. [16]
Crop productivity	GOSIF GPP	0.05°	8-day	Dataset comparison	Li and Xiao [29]

to 1 km spatial resolution using bilinear interpolation before subsequent analysis. The detailed information of the data used in this study have been summarized in Table I.

1) Crop Yield and Distribution: County-level soybean and corn yields were obtained from the National Agricultural Statistics Service (NASS) of the United States Department of Agriculture (USDA).[1] The original unit of soybean and corn yields, bushels/acre, was converted to kilogram/hectare by multiplying it with a conversion factor of 67.2 and 62.7, respectively. Data for the cultivation distribution of soybean and corn were obtained annually from the cropland data layer[2] [30], [31] with a 30 m resolution. We selected the period from 2006 to 2021 for our study due to gaps in CDL for the Corn Belt prior to 2006. The CDL data were used to extract remote sensing variable data for the corresponding crop planting area.

2) Vegetation Index: Vegetation indices play an important role in monitoring the state of crop growth and have been widely used for crop yield estimation. Normalized difference vegetation index (NDVI) and enhanced vegetation index (EVI), obtained from the aqua moderate resolution imaging spectroradiometer (MODIS) dataset provided by the GEE,[3] were used in our study. Both indices were a composite of 16 days, with a spatial resolution of 1 km.

3) Climate Data: Climate predictors used in our study were extracted from four gridded data sources: the parameter-elevation regressions on independent slopes model (PRISM) dataset, MODIS Evapotranspiration/Latent Heat Flux MOD16A2 product, MODIS Land Surface Temperature (LST) product and gridded surface meteorological (GRIDMET) dataset. Specifically, we collected monthly total precipitation (PPT), monthly mean temperature (Tmean), monthly minimum and maximum vapor pressure deficit (VPDmin, VPDmax) from the 4-km PRISM dataset.[4] The average 8-day daytime and

[1]Online. [Available]: https://quickstats.nass.usda.gov/
[2]Online. [Available]: https://nassgeodata.gmu.edu/CropScape/
[3]Online. [Available]: https://developers.google.cn/earth-engine/datasets/catalog/MODIS_061_MYD13A2
[4]Online. [Available]: https://developers.google.cn/earth-engine/datasets/catalog/OREGONSTATE_PRISM_AN81m

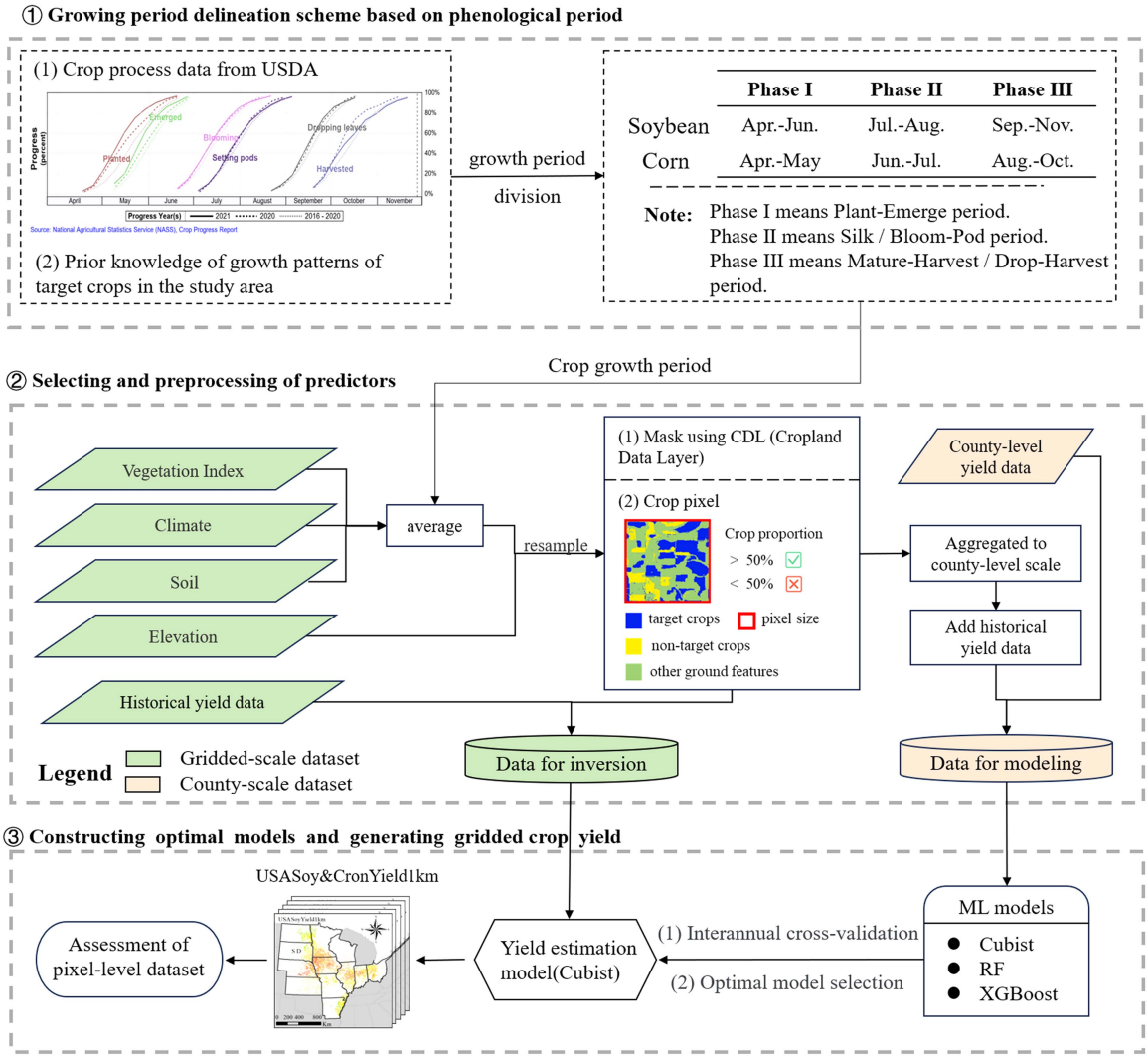

Fig. 2. Flowchart of the gridded crop yield generation framework.

nighttime land surface temperatures (LST_Day, LST_Night) were obtained from the 1 km MODIS LST product,[5] while evaporation (ET) and potential evaporation (PET) were collected from MOD16A2 product,[6] an 8-day composite dataset with a 500 m resolution. Additionally, to characterize regional drought conditions, the palmer drought severity index (PDSI) was derived from GRIDMET dataset,[7] which has a 4-km spatial resolution and 5-day temporal resolution.

In summary, the variables can be divided into two categories: (1) Temperature-related variables, including Tmean, LST_Day and LST_Night. (2) Water-related variables, including PPT, ET, PET, VPDmin, VPDmax and PDSI.

4) Soil Data: Soil conditions also have a significant effect on crop growth and final yield. In our study, volumetric soil water (VSW) and soil temperature (ST) were derived from ERA5-Land monthly data with a spatial resolution of 0.1°.[8] This dataset is produced by replaying the land component of the ECMWF ERA5 climate reanalysis. Considering the depth of the main root distribution of the studied crop, two depth groups, 0–7 cm and 7–28 cm, were selected for these soil attributes, comprising a total of four variables for yield estimation (ST1, ST2, VSW1 and VSW2, with 1 representing 0–7 cm and 2 representing 7–28 cm).

5) Elevation: The relationship between terrain attributes and the spatial patterns of crop yield has been demonstrated by previous studies, and the digital elevation model (DEM) data contribute to estimating crop yields [32], [33]. Elevation from NASADEM with a spatial resolution of 30 m,[9] which is a reprocessing of the shuttle radar topography mission digital elevation dataset, was used to predict crop yield in this study.

[5]Online. [Available]: https://developers.google.cn/earth-engine/datasets/catalog/MODIS_061_MOD11A2

[6]Online. [Available]: https://developers.google.cn/earth-engine/datasets/catalog/MODIS_006_MOD16A2

[7]Online. [Available]: https://developers.google.cn/earth-engine/datasets/catalog/GRIDMET_DROUGHT

[8]Online. [Available]: https://developers.google.cn/earth-engine/datasets/catalog/ECMWF_ERA5_LAND_MONTHLY_AGGR

[9]Online. [Available]: https://developers.google.cn/earth-engine/datasets/catalog/NASA_NASADEM_HGT_001

C. Methods

1) Model Selection and Evaluation: In this study, we selected three widely used ensemble machine learning models for soybean and corn yield estimation: RF, XGBoost, and Cubist. RF is an ensemble learning algorithm that improves accuracy and robustness by constructing multiple decision trees and aggregating their predictions [34]. The RF model can efficiently process remote sensing datasets with numerous features due to its low sensitivity to outliers and resistance to overfitting [19], [35]. The XGBoost model is an efficient implementation of the gradient boosting algorithm. It enhances model performance by sequentially adding weak learners to gradually reduce residuals [36]. XGBoost is not affected by highly correlated features, thereby reducing issues of multicollinearity among features [3], [37]. Cubist is a machine learning algorithm that combines decision trees with linear regression. It initially uses rules similar to decision trees to split the data into different subsets and then applies a linear regression model to each subset for prediction [38]. This approach enables Cubist to effectively handle datasets containing both linear and nonlinear relationships [39].

Interannual cross-validation was employed to assess the performance of models and select the optimal model in this study. Specifically, we adopted a "leave-one-year-out" test for model evaluation [40], [41]. For instance, data from 2007 to 2021 were used to train the model, which was then used to predict the yield for 2006. Then the data from 2006 and 2008–2021 were used to predict for 2007, and so on for each year. The 16 years' predictions were then compared with the actual annual statistical yield data year by year. The performance of models, unaffected by the length of training data, benefits from this testing method, which provides a comprehensive evaluation of the predictive capability under various conditions of models.

In this study, we used the coefficient of determination (R^2), root mean squared error (RMSE) and mean absolute error (MAE) to quantify the model prediction performance. These metrics were individually computed for each predicted year. The model with the highest R^2 and the lowest RMSE and MAE was deemed as the optimal model.

2) Gridded Crop Yield Generation: Based on the obtained optimal yield prediction model, the USASoy&CornYield1km was generated according to the following steps (see Fig. 2).

a) Data preparation: As stated in Section II-B, we collected a total of 16 remote sensing variables in four categories as independent variables for the model: vegetation indices (NDVI and EVI), climate variables (ET, PET, LST_Day, LST_Night, PPT, Tmean, VPDmin, VPDmax, and PDSI), soil attributes (ST and VSW at 0–7 cm and 7–28 cm depths), and elevation, were used as independent variables in the model. Crop statistical yield for each year provided by NASS was the dependent variable of the model. Previous studies have demonstrated that using a priori crop yield information from the previous five years in the same region will facilitate the construction of estimation models [4], [42]. This is because crop yields in the same region tend to remain relatively stable in the absence of major environmental changes. Therefore, we also input the average yield of the previous five years as the independent variable of the model. For example, to estimate the crop yield for 2021, we calculated the average yield for each county from 2016 to 2020 as the historical average yield. Therefore, the model includes a total of 17 independent variables.

b) Growth stage aggregation: The 17 variables selected for the study include many images taken throughout the crop growing season, with excessive images much likely to lead to data redundancy and hamper model calculations. To reduce the number of predictors, we aggregated 15 remote sensing variables (excluding elevation) at different crop growth stages. Specifically, for soybean, the growing season has been divided into the planting-emerging stage (April to June), blooming-podding stage (July to August), and dropping-harvesting stage (September to November) according to the Crop Calendars and Crop Progress Report published by USDA [43], [44]. Similarly, the growth season of corn can be divided into planting-emerging stage (April to May), silking stage (June to July), and mature-harvesting stage (August to October). We calculated the average of the 15 variables for different stages respectively and put the mean values into model as independent variables. Thus, a total of 45 variables (15 variables $\times$ 3 stages) were derived for each year.

c) CDL mask: All remote sensing variables were masked using the CDL data of the respective years. To use the CDL for masking remote sensing variables and, concurrently, to minimize the effects of mixed pixels, we used a pixel-by-pixel proportion discrimination method [40]. First, a 1 km $\times$ 1 km spatial grid was established within the study area. Subsequently, the total area of the soybean (or corn) CDL pixels within each grid cell was then calculated. Only grid cells with an area percentage greater than 50% were selected as soybean (or corn) dominated pixels and used to mask the remote sensing variables [45].

d) County-level aggregation: The average of each masked variable within county-level units was computed to match the annual crop yields as reported in NASS. This procedure yielded an annual set of 47 variables (45 remote sensing variables at different stages, elevation and historical average yield) for each county for yield prediction.

e) Developing the optimal models: The Cubist, RF, and XGBoost models were used to fit the 47 county-level variables with the crop yield. Subsequently, the accuracy of the models was evaluated through the "leave-one-year-out" test (as described in Section II-C1) to select the optimal model.

f) Gridded crop yield generation: For each crop, gridded-scale predictive variables, consistent with the administrative-level selected predictors, were input into the optimal estimation model [20], [21], and the gridded annual crop yield dataset, USASoy&CornYield1km, was generated, which covered the period from 2006 to 2021.

g) Results evaluation: We compared USASoy&CornYield1km with statistical yield and other gridded datasets to assess the accuracy and reliability of USASoy&CornYield1km (See Section II-C3 for details).

3) Evaluation and Assessment of Gridded Crop Yield Data: In this study, we employed three different methods to validate

the reliability of the generated 1-km gridded crop yield data. Firstly, for the USASoy&CornYield1km produced by the optimal model, the "leave-one-year-out" test was also used to assess its reliability. We aggregated USASoy&CornYield1km into county-level units and then compared them with the yearly statistical yield [21]. In addition to R^2, RMSE, and MAE, we also calculated the relative root mean square error (rRMSE) to compare the estimation accuracy of USASoyYield1km and USACornYield1 km, which are subsets of the USASoy&CornYield1km dataset. For the aggregated results, the rRMSE for each year was calculated as the ratio of the year's RMSE to the average statistical yield across all counties [46] (1). For the raster images of the USASoy&CornYield1km, the calculation of rRMSE was performed through the method described by Luo et al. [47] and Wu et al. [21], as outlined in (2). The rRMSE for each county was then assigned to the centroids, and kriging interpolation was applied to generate a spatial distribution map of rRMSE, masked using the CDL data processed in Section II-C2

$$\mathrm{rRMSE}_{\mathrm{statistics}} = \frac{\mathrm{RMS}E_j}{\overline{O_j}} \times 100\% \tag{1}$$

$$\mathrm{rRMSE}_{\mathrm{images}} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}\left(\frac{O_{i,j}-E_{i,j}}{O_{i,j}}\right)^2} \times 100\% \tag{2}$$

where i is the number of administrative units, j is the year, and n is the total number of years, $O_{i,j}$ and $E_{i,j}$ stands for the statistical yield and estimated yield in the ith administrative unit of year j, respectively, and $\overline{O_j}$ is the average statistical yield across all counties in year j.

Second, to further test the robustness and reliability of the gridded yield data, we compared the USASoy&CornYield1km dataset with soybean and corn yields derived from the SPAM,[10] which is a commonly used global dataset for gridded crop yield estimates [16].

Third, since the spatial resolution of SPAM data is relatively coarse (∼10 km) and only the 2010 dataset from SPAM is available for such a comparative analysis, we also compared USASoy&CornYield1km with the GOSIF GPP. Gross primary productivity (GPP) represents the total amount of carbon dioxide that is fixed by plants during photosynthesis, which is a fundamental process driving crop growth and yield. Therefore, GPP is closely related to crop yield, as it reflects the overall photosynthetic activity and biomass accumulation of crops over time [48], [49]. Using GPP as a validation metric allows us to assess the consistency and accuracy of our gridded yield estimates at a finer spatial resolution. The GOSIF GPP dataset,[11] with its spatial resolution of approximately 5 km and a temporal resolution of 8 days, offers detailed and timely insights into the photosynthetic performance of crops [29]. By comparing USASoy&CornYield1km with GOSIF GPP, we can evaluate how well our crop yield estimates capture the spatial and temporal variations in crop productivity. The long time-series of GOSIF GPP also allows for a comprehensive comparison with USASoy&CornYield1km throughout the entirety of our study period (2006–2021). Combining these three validation measures, we demonstrate the accuracy and robustness of our production results.

[10]Online. [Available]: https://mapspam.info/

[11]Online. [Available]: http://data.globalecology.unh.edu

III. Results

A. Performance of the Prediction Models

Fig. 3 presents the accuracy of yield predictions for soybean and corn at the county level using Cubist, XGBoost, and RF. It is evident that all models achieved satisfactory results in county-level yields prediction. Among them, the Cubist model slightly outperforms XGBoost and RF. In predicting soybean yields, the Cubist model has averaged an R^2 of 0.81, RMSE of 303.59 kg/ha, and MAE of 239.49 kg/ha over 16 years. In comparison, XGBoost averages an R^2 of 0.79, RMSE of 314.62 kg/ha, and MAE of 245.93 kg/ha, while RF has an R^2 of 0.80, RMSE of 314.06 kg/ha, and MAE of 245.70 kg/ha. This disparity is even more pronounced in corn yield predictions. For instance, Cubist achieves an average R^2 of 0.80, RMSE of 987.11 kg/ha, and MAE of 762.51 kg/ha, whereas both XGBoost and RF have an average R^2 around 0.75, with RMSE exceeding 1100 kg/ha and MAE approaching 880 kg/ha. Consequently, we adopted Cubist as the optimal estimation model in subsequent analysis and employed it in the production of gridded crop yield estimates (i.e., USASoy&CornYield1km).

B. Comparison of the Generated Gridded Crop Yield With Observations

After aggregating USASoy&CornYield1km in county-level units, we compared it with the observed yields, and the results are presented in Fig. 4 (soybean) and Fig. 5 (corn), respectively. It can be found that the ranges of R^2, RMSE, and MAE for soybean are 0.70–0.86, 282.36–592.85 kg/ha, and 220.34–519.81 kg/ha, respectively, and those for corn are 0.43–0.88, 993.08–2695.78 kg/ha, and 781.58–2192.23 kg/ha. These results indicate that the accuracy of USASoy&CornYield1km at the county level is at a high level in all years. For a fair comparison of yield prediction accuracy between soybean and corn, we calculated the rRMSE for both crops. Our findings showed that soybean has a lower rRMSE range (9.4%–21.48%, average of 13.18%) compared to corn (11.54%–38.57%, average of 17.74%), indicating that the soybean yield prediction accuracy is higher than the corn. Notably, the accuracy of the yield predictions for both crops is lowest in 2012. Coincidentally, 2012 experienced the most severe agricultural drought in the U.S. Corn Belt since 1988. Hence, the lower accuracy in that year may be attributed to the extreme climatic conditions, as previously reported by other yield prediction studies [3], [40], [42]. During the same period, corn yields reached their lowest level since 1995, while soybean yields were also affected though the impact was less significant. This situation aligns with the anticipated model accuracy. In 2012, the rRMSE of the yield prediction results for soybean and corn were as high as 21.48% and 38.75%, respectively. In contrast, the rRMSE for both crops

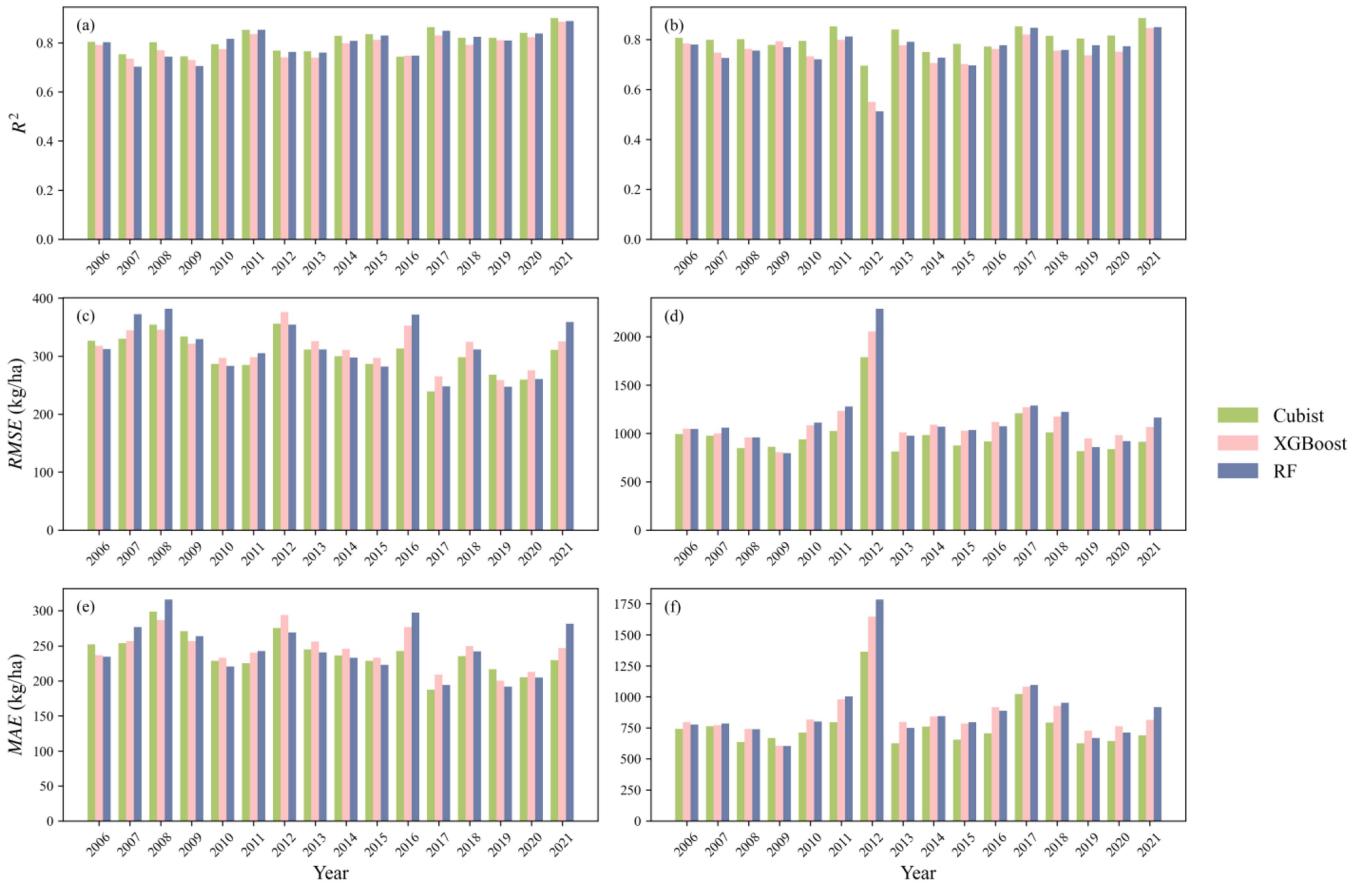

Fig. 3. Model performance (R^2, RMSE, and MAE) of county-level yields estimation for soybean (a), (c), (e) and corn (b), (d), (f) through interannual cross-validation.

was less than 20% in other years, indicating good consistency between the predicted yields and the observed yields [50].

C. Comparison of the Generated Gridded Crop Yield With SPAM

To further validate the accuracy of the generated gridded crop yield estimates, we compared USASoy&CornYield1km with SPAM yield spatial dataset. Fig. 6(a)–(c) show the distribution of soybean yields in USASoyYield1km (soybean subset of USASoy&CornYield1km) and SPAM across the U.S. Corn Belt, compared with county-level yield statistics. The yield distribution in South Dakota is highlighted, which demonstrates greater spatial heterogeneity along its border with neighboring states, as shown in Fig. 6(a1)–(c1). The spatial distribution comparison of USACornYield1 km (corn subset of USASoy&CornYield1km) with SPAM and county-level yield statistics is displayed in Fig. 7. It is apparent that the spatial distribution of USASoy&CornYield1km is very similar to the SPAM but with enhanced capability to display more spatial details of crop yields, and shows a high level of consistency with the observed yield data.

Moreover, we aggregated USASoyYield1km within grid cells corresponding to the size of SPAM pixels and conducted a pixel-by-pixel comparison, as shown in Fig. 8(a). And the comparison between both datasets and statistical yields is presented in Fig. 8(b). A similar comparison between USACornYield1 km and SPAM is shown in Fig. 9. These comparisons indicate a high correlation between the yields in USASoy&CornYield1km and those in SPAM, with correlation coefficients of 0.84 for soybean and 0.78 for corn. Additionally, both SPAM and USASoy&CornYield1km show high correlation when compared with observed yields. These results suggest that USASoy&CornYield1km is highly reliable at the pixel scale.

D. Comparison of the Generated Gridded Crop Yield With GOSIF GPP

As detailed in Section II-C3, the gross primary productivity is a vital remote sensing variable indicating vegetation growth status and is closely related to crop yield. In this study, we compared the GOSIF GPP with our USASoy&CornYield1km dataset to evaluate the ability of the generated gridded crop yield dataset to capture the spatial and temporal variations in crop productivity. We aggregated the GOSIF GPP data during different growth stages of soybeans and corn, and then calculated average values within each county-level unit. This allowed us to explore the correlation between the aggregated GOSIF GPP results at various growth stages and the statistical yields, as shown in Fig. 10(a). Apparently, the highest correlation between aggregated GPP and annual yield is observed during the mid-growing season (July to August for soybeans

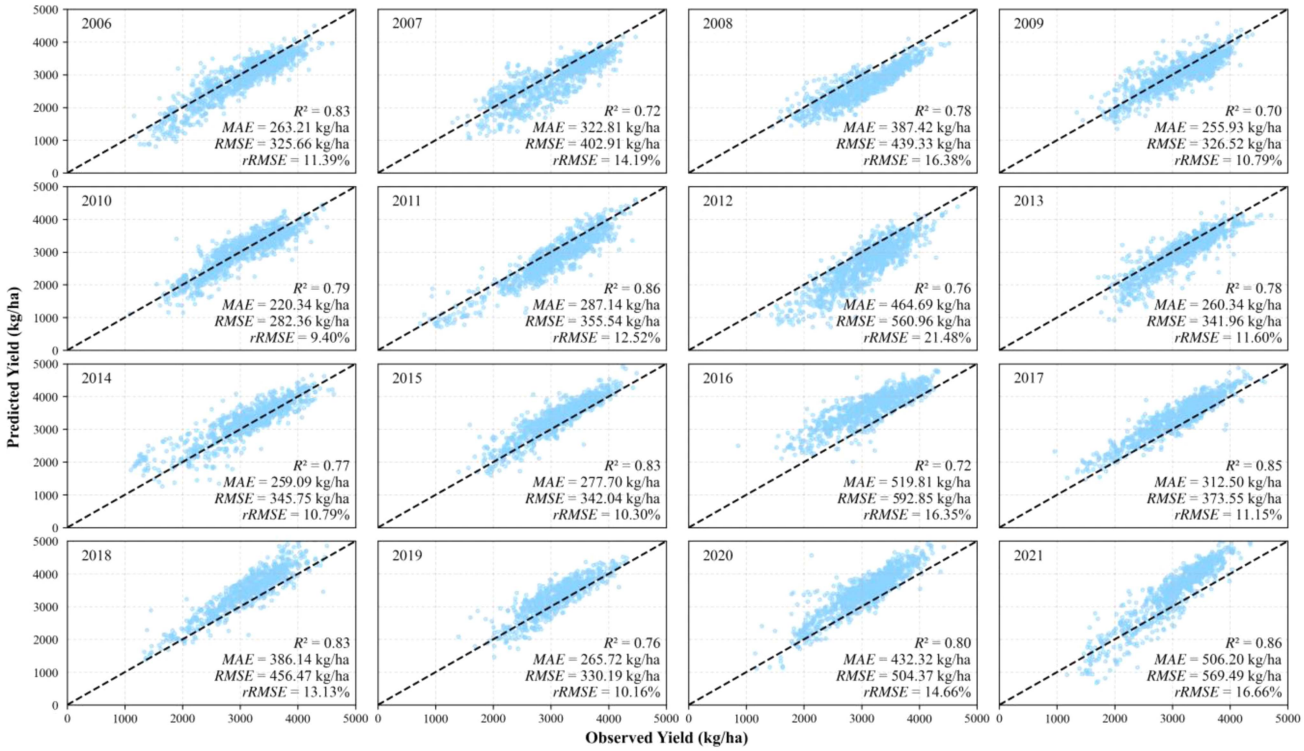

Fig. 4. Interannual cross-validation results of USASoyYield1km with county-level statistical yields. Note that USASoyYield1km is a soybean subset of the USASoy&CornYield1km dataset.

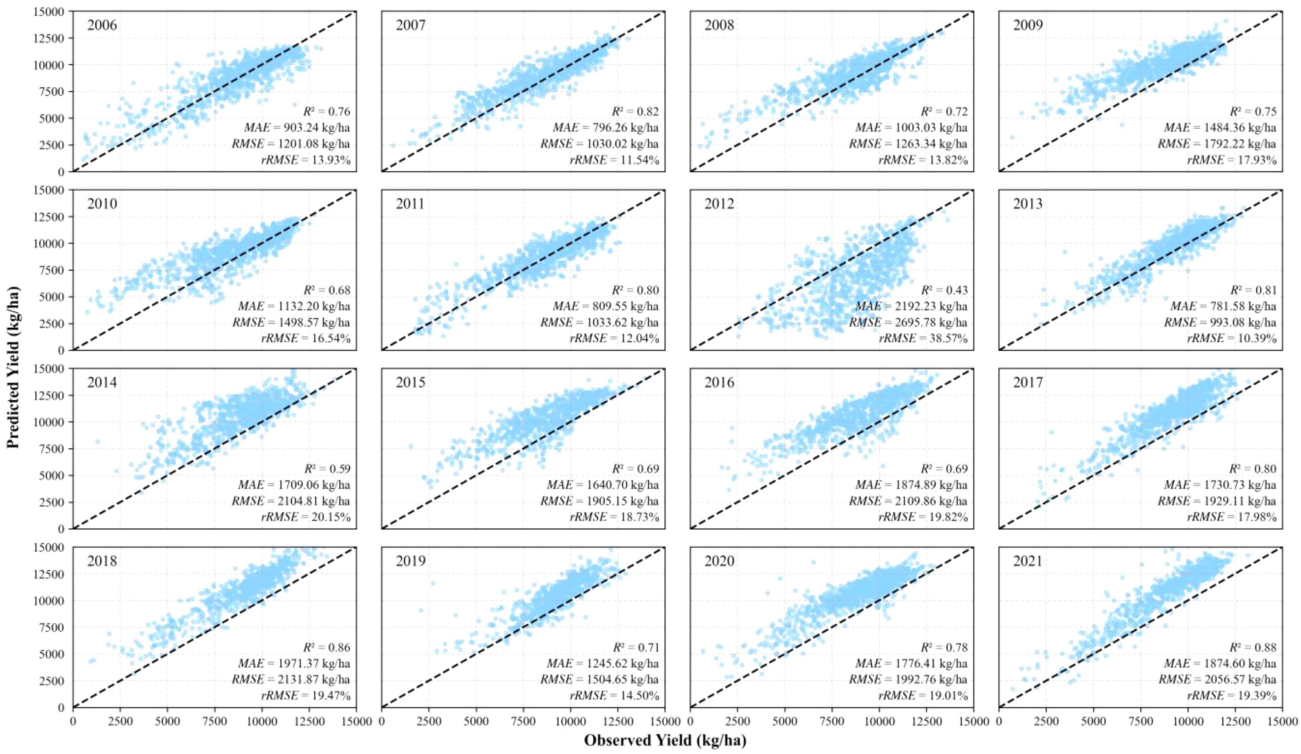

Fig. 5. Interannual cross-validation results of USACornYield1km with county-level statistical yields. Note that USACornYield1 km is a corn subset of the USASoy&CornYield1km dataset.

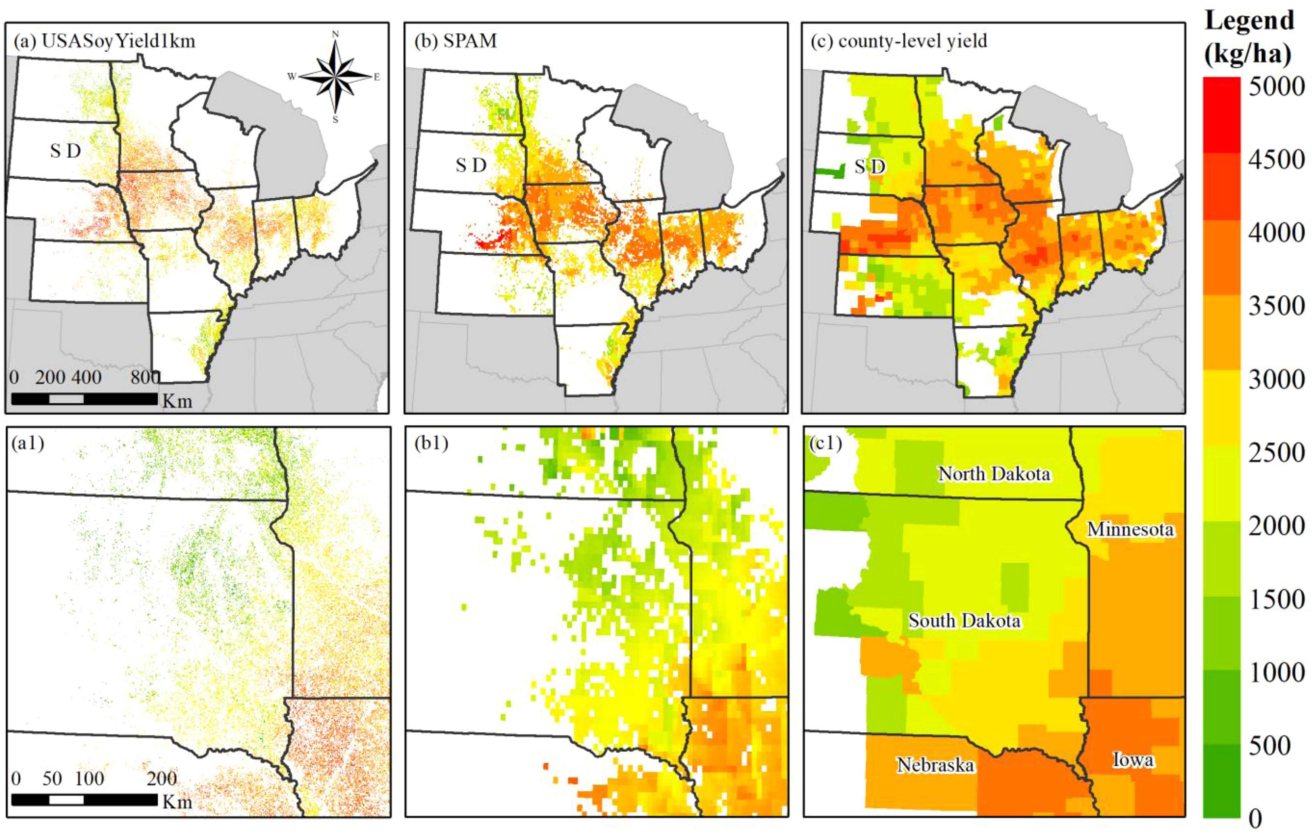

Fig. 6. Yield distribution of (a) USASoyYield1km, (b) SPAM, and (c) observed yields in 2010. Panels (a1) to (c1) show zoomed-in views of the yield distribution along the South Dakota (SD) border, where (a1) is USASoyYield1km, (b1) is SPAM, and (c1) is county-level statistical yields. Note that, due to the limited temporal coverage of SPAM, only soybean yields from the 2010 SPAM dataset are available for comparison with USASoyYield1km.

and June to July for corn). Consequently, we compared the mid-season aggregated GPP with the USASoy&CornYield1km at pixel scale, using a method similar to that used with the SPAM dataset. The comparison results, presented in Fig. 10(b), indicate a strong correlation between the yields of the two crops in the USASoy&CornYield1km dataset and the GOSIF GPP. The correlation coefficients range from 0.66 to 0.87 for soybean, with an average of 0.80, and from 0.59 to 0.83 for corn, with an average of 0.70. The year with the lowest correlation was notably the extreme drought year of 2012. The observed correlation between GOSIF GPP and USASoy&CornYield1km suggests that the reliability of USASoy&CornYield1km at the pixel scale is not limited to the year 2010 but extends to all predicted years, indicating a consistent level of reliability across the study periods.

E. Spatial Patterns and Uncertainty Evaluation

Fig. 11 shows the spatial pattern of the multiyear average values and annual distributions of USASoy&CornYield1km between 2006 and 2021, indicating a remarkably similar distribution pattern for both corn and soybean yields. For soybeans, Nebraska recorded the highest average yield (3667.00 kg/ha), followed by Illinois (3558.79 kg/ha) and Iowa (3517.20 kg/ha). In the case of corn, Illinois achieved the highest average yield of 10 579.59 kg/ha, followed by Iowa (10 547.53 kg/ha) and Arkansas (10 410.27 kg/ha). In contrast, the states of North Dakota, Kansas, and South Dakota showed the lowest average yields for both soybeans and corn, with yields of 2102.58, 2526.08, 2729.40 kg/ha for soybeans, and 6173.09, 6907.28, 7152.77 kg/ha for corn, respectively. Overall, these data are highly consistent with statistics provided by the NASS, with an R^2 of 0.97 for soybeans and 0.95 for corn, and RMSE of 218.79 kg/ha for soybeans and 345.27 kg/ha for corn. Therefore, the USASoy&CornYield1km dataset exhibits high accuracy not only at the county level but also at the state level. The central region of the Corn Belt, particularly Nebraska, Illinois, and Iowa, presents as highly productive areas for both soybeans and corn. This may be attributed to favorable climatic conditions and advanced agricultural techniques [51], [52].

Moreover, we evaluated the yield estimation uncertainty, with its spatial distribution shown in Fig. 12. In more than three-quarters of the regions, the rRMSE for soybeans is less than 15% and for corn less than 25%. The regional average rRMSE is 14.04% for soybeans and 20.49% for corn, indicating a generally low level of uncertainty in the USASoy&CornYield1km dataset. The uncertainty for both soybeans and corn is relatively low in central states such as Iowa and Nebraska. High uncertainty is distributed in North Dakota and South Dakota for both crops. Notably, the uncertainty for corn and soybean in Missouri and southern Illinois is comparatively high, and significantly higher for corn than for soybeans in the same region. According to

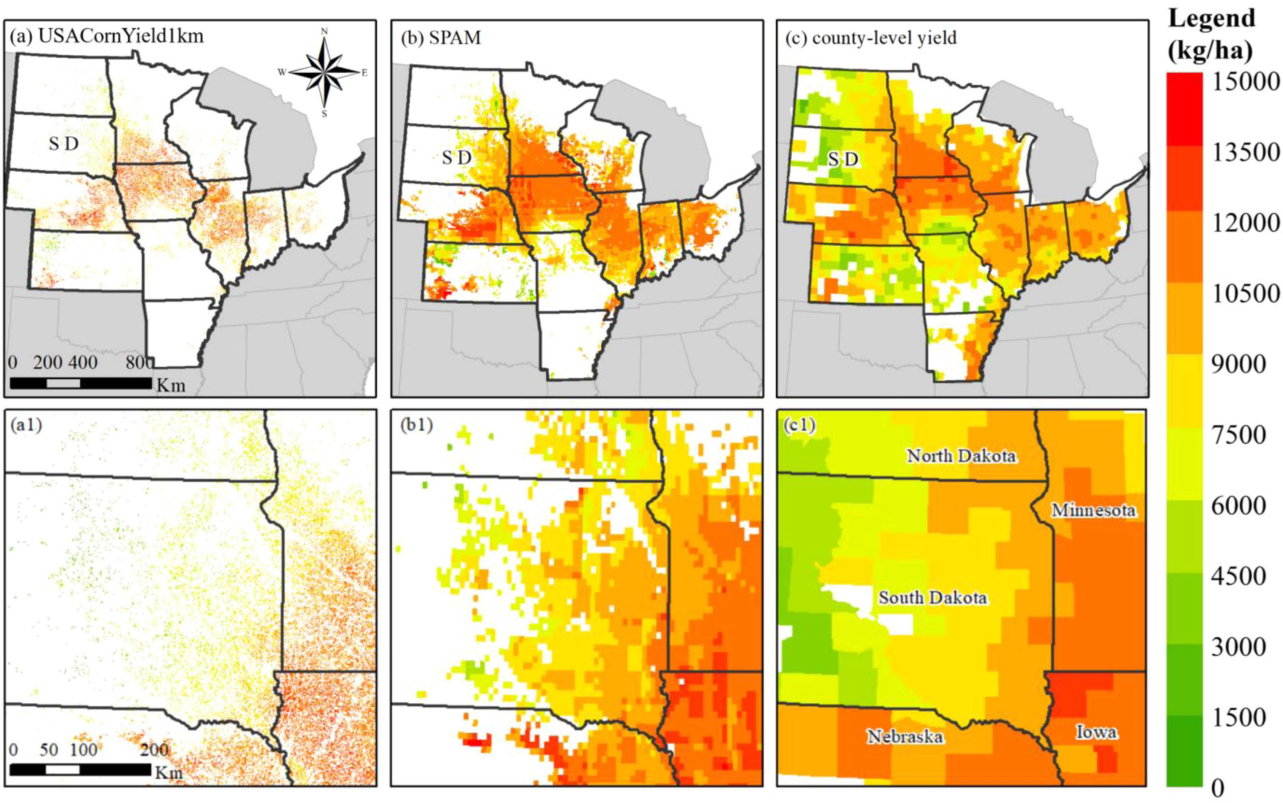

Fig. 7. Yield distribution of (a) USACornYield1 km, (b) SPAM and (c) observed yields in 2010. Panels (a1) to (c1) show zoomed-in views of the yield distribution along the South Dakota (SD) border, where (a1) is USACornYield1 km, (b1) is SPAM, and (c1) is county-level statistical yields. Note that, due to the limited temporal coverage of SPAM, only corn yields from the 2010 SPAM dataset are available for comparison with USACornYield1 km.

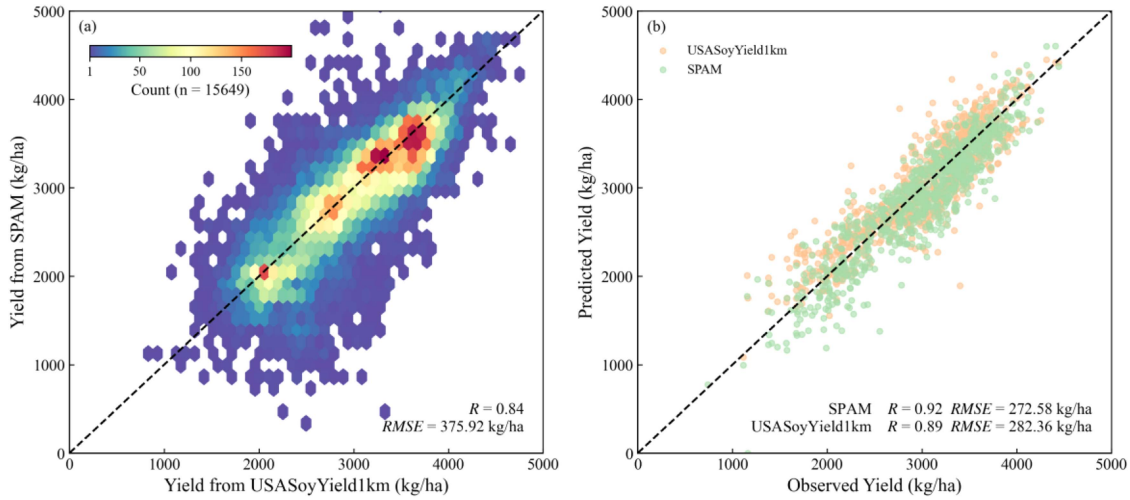

Fig. 8. Quantitative comparison between USASoyYield1km and SPAM. Panel (a) shows the correlation between USASoyYield1km and SPAM, and panel (b) shows the correlation between the two and observed yields.

the crop yield distribution data published by the USDA, Illinois, despite a major producer of soybeans and corn, has its production mainly concentrated in the northern counties, with significantly smaller planting areas for both crops in the southern part of the state. In Missouri, the planting areas for both crops are relatively small as well. This suggests that regions with smaller crop planting areas tend to exhibit relatively higher uncertainty [4], [21]. This pattern is also evident in the central regions of North Dakota and South Dakota, as well as in Kansas, where the crop planting is more scattered.

F. Importance Analysis of the Predictors

In this study, the 47 variables used for crop yield prediction show significant variations in contributions, as illustrated by the

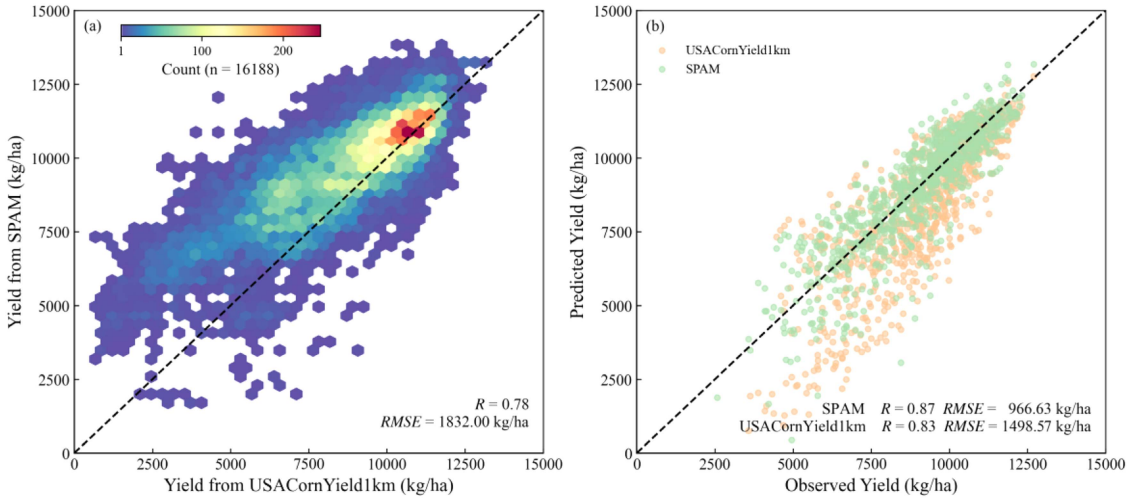

Fig. 9. Quantitative comparison between USACornYield1 km and SPAM. Panel (a) shows the correlation between USACornYield1 km and SPAM, and panel (b) shows the correlation between the two and observed yields.

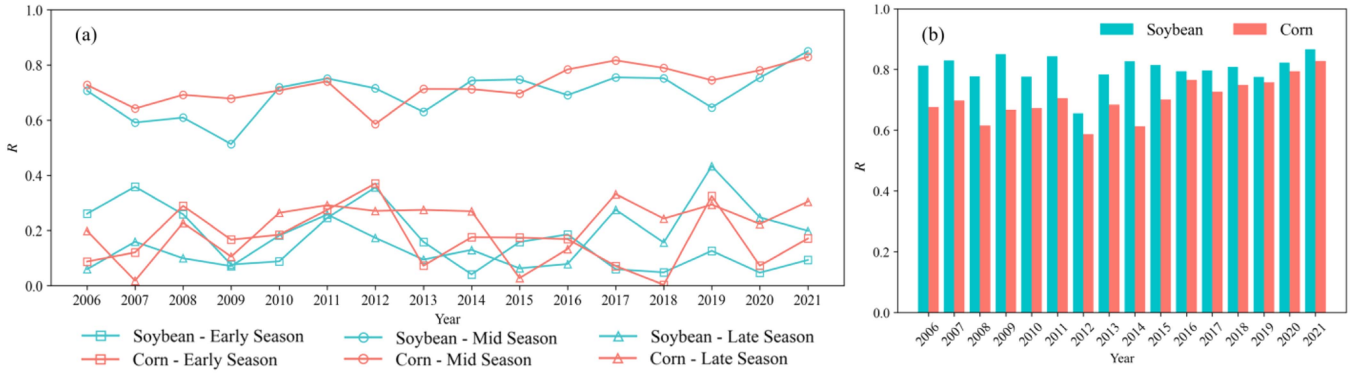

Fig. 10. Annual correlation of GOSIF GPP with USASoy&CornYield1km and statistical yield. Panel (a) shows the correlation between GOSIF GPP aggregated at different growth stages and statistical yield, and panel (b) shows the correlation between GOSIF GPP during the mid-growing season and USASoy&CornYield1km.

variable importance ranking from the optimal Cubist model (see Fig. 13). For both soybean and corn, historical yield is the most important variable. For soybeans, EVI during the bloom-pod stage (July to August) has a high contribution, followed by ST1, PPT during the same period and DEM. For corn, EVI is equally important in both the silk and mature-harvest stages, and PDSI also makes a significant contribution. Notably, the contribution of DEM for soybean yield prediction is significantly higher than for corn. The variables that contribute least to yield prediction for both crops are the LSTs, regardless of whether it is daytime or nighttime.

The pie charts in Fig. 13 show the relative importance of vegetation indices, climate data, and soil data at different stages of crop growth. It is evident that for both soybeans and corn, climate data contribute the least to yield prediction at every stage. For soybeans, soil data is most influential during the plant-emerge stage, while vegetation indices are predominant in the bloom-pod stage. In the drop-harvest stage, the importance of both these factors is approximately equal. In the case of corn, soil data remains the most significant contributor during the plant-emerge stage. However, for both the silk and mature-harvest stages, the importance of vegetation indices surpasses that of soil data. This pattern suggests that soil attributes have a more substantial impact on crop yield during the early stages of growth. As the crops enter their peak growth phase, vegetation indices become more indicative of the final yield. While climate does affect crop yields, it is less effective as a standalone predictor [3].

IV. Discussion

A. Advancements of the Gridded Crop Yield Estimates

High-resolution crop yield distribution information can enhance agricultural management by facilitating optimized resource use and promoting environmental sustainability [19], [53]. In this study, we proposed a statistical downscaling framework and generated a high-resolution dataset for soybean and corn yields from 2006 to 2021, known as USASoy&CornYield1km, covering the primary planting regions of the United States. This was accomplished through the integration of multisource geospatial data and advanced machine learning techniques. Compared to other existing and publicly

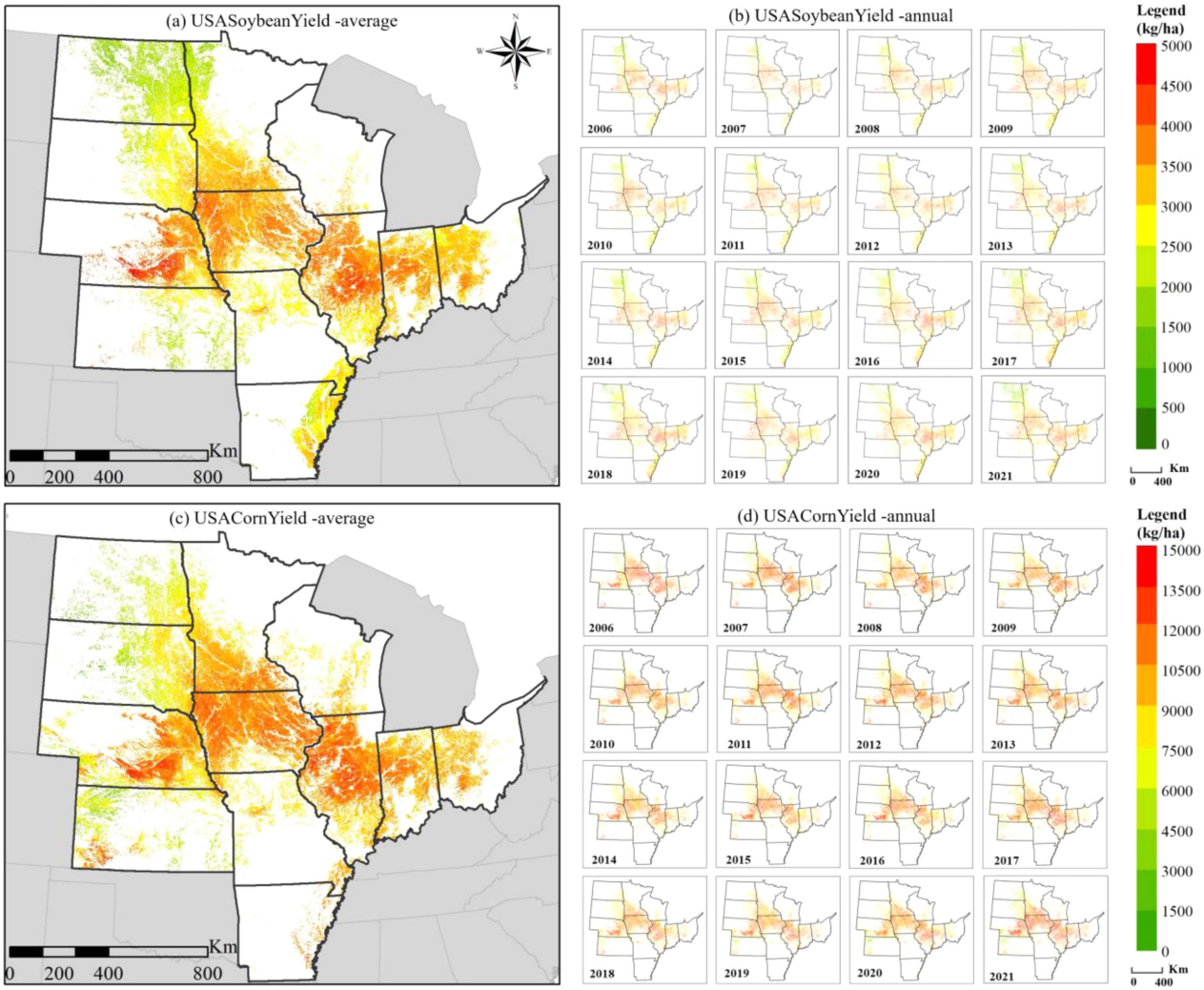

Fig. 11. Spatial distribution of multiyear average yields and annual distribution of USASoy&CornYield1km for the period 2006 to 2021. Left panels (a) and (c) represent the spatial distribution of average yields for soybeans and corn, respectively. Right panels (b) and (d) represent the annual yield distribution for soybeans and corn, respectively.

available crop yield datasets, such as SPAM and GDHY, the USASoy&CornYield1km dataset has several advantages.

First, the gridded crop yield estimates generated in this study has a high spatio-temporal resolution with a long time span. USASoy&CornYield1km provides a spatial resolution of 1 km and a time span of 16 years (2006–2021), which is superior to existing public datasets. The annual temporal resolution coupled with the long time span allows for the interannual dynamic analysis of crop yields. By accurately downscaling crop yield statistics from region to pixel scale, a high spatial resolution of 1 km has been achieved, which enables USASoy&CornYield1km to provide more accurate spatial information on crop yields and reflect internal yield variability within administrative units (e.g., counties). This approach overcomes the limitations of traditional statistical data and low spatial resolution datasets in conducting spatial analysis within counties, thereby facilitating precise monitoring of crop yields over large areas [15], which can guide agricultural production and improve productivity.

Moreover, the generated spatial yield dataset has a stable accuracy. The Cubist model constructed in our study, compared to other models, demonstrated optimal performance with high accuracy in the "leave-one-year-out" test, ensuring the precision of yield estimation [40], [41]. The study also incorporated four categories of variables for crop yield inversion. By aggregating these variables at different growth stages of the crops, comprehensive growth information throughout the season was provided, improving the accuracy of the predictions. The high correlation between USASoy&CornYield1km and county-level yield statistics indicates high accuracy at the county scale; comparisons with SPAM data demonstrate the reliability of USASoy&CornYield1km at the pixel scale. Such stable accuracy indicates that USASoy&CornYield1km can provide a reliable basis for agricultural management and production [54]. In addition, in this study, we discriminated soybean and corn planting areas pixel by pixel, accounting for spatial and temporal dynamics of the crop planting conditions. This method incorporates richer crop planting information compared to constant

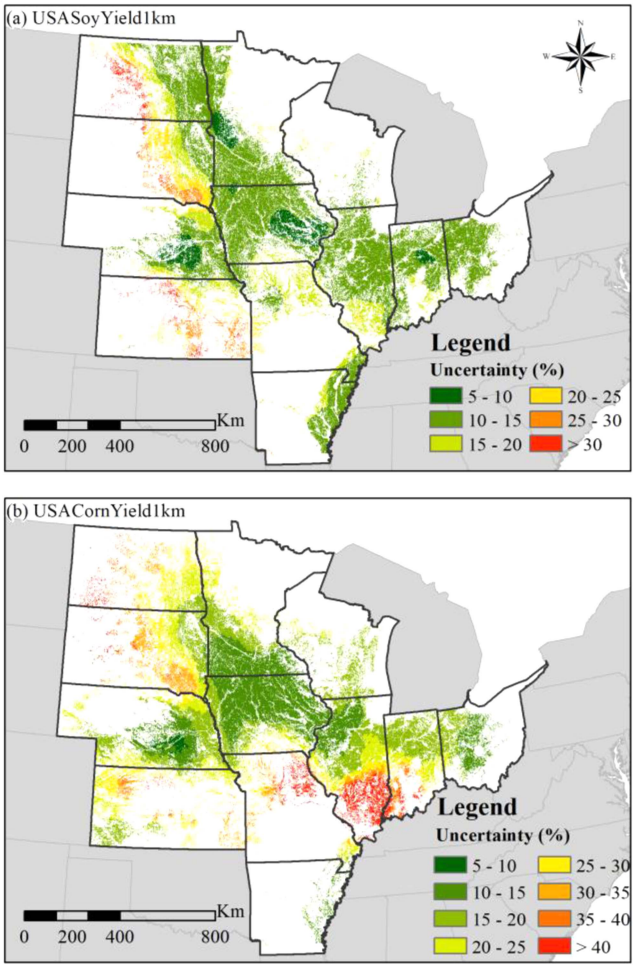

Fig. 12. Spatial distribution of uncertainty (rRMSE, %) in USASoy& CornYield1km. (a) Represents the uncertainty of soybean yield and (b) represents the uncertainty of corn yield.

cultivation area maps [55], better reflecting the spatiotemporal changes in crop planting areas and thereby greatly enhancing the accuracy of our yield estimates.

B. Generalizability and Transferability of the Method

The spatial transferability of the proposed method is a critical aspect that determines its applicability across different geographical regions. Unlike the U.S. Corn Belt, other regions may have different agricultural and climatic conditions and often lack high-precision crop distribution data similar to the CDL. This makes the spatial transfer of the method challenging. However, inspired by the work of Hao et al. [56], who successfully predicted crop spatial distribution in other regions using CDL as training samples, we found that the same crop grown in different regions of the world exhibits similar temporal growth patterns. This similarity allows for the prediction of crop planting spatial distribution in any region using a similar approach proposed by Hao et al. [56]. Besides crop spatial distribution data, all the training data used in this study can be downloaded from GEE, covering any region globally. These data generally do not suffer from missing values and adequately reflect the varying agricultural and climatic conditions of different regions. Therefore, the crop yield prediction method proposed in this study holds potential for application within the existing transfer learning framework.

Another noteworthy issue is that the CDL used in the study is typically released in January or February of the following year. For example, the CDL for 2023 was published in February 2024. This means that the models built in this study can only improve the spatial resolution of crop yield from the previous year, and cannot directly predict crop yield spatial distribution maps with high spatial resolution for the following year. This limitation hinders the application of this method for in-season crop yield prediction. However, thanks to the work of Zhang et al. [13], who used historical CDL as training samples and employed multilayer artificial neural networks to predict the new year's crop planting distribution, predicting the spatial distribution before the growing season has become possible. This study combined historical yield statistics with remote sensing information to construct a Cubist model that downscales the crop yield statistics for the new year. By integrating this method with the approach proposed by Zhang et al. [13], it may become possible to obtain high spatial resolution crop yield distribution maps during the mid-growing season. This would be of significant importance for agricultural production decision-making.

C. Uncertainties and Limitations

In this study, we conducted an uncertainty analysis of the generated crop yield estimates across the spatial domain, as shown in Fig. 12. The results demonstrate that, for most regions, the uncertainty indicated by the rRMSE is below 15% for soybeans and less than 25% for corn. This suggests an overall low level of uncertainty in the gridded crop yield dataset, USASoy&CornYield1km. While USASoy&CornYield1km offers considerable advantages, some limitations still exist. For instance, areas with smaller scale and scattered crop cultivation, such as in South Dakota, North Dakota, and Kansas, exhibit relatively higher uncertainty in crop yield prediction. This may be attributable to the occurrence of mixed pixels during the crop yield prediction processes due to the different spatial resolutions of multisource data, introducing uncertainty. Some of the meteorological and soil data used in our study, although commonly used in previous studies, have a coarser spatial resolution. Mixed pixels are inevitably generated during the prediction process and adversely affect the accuracy of the gridded crop yield estimates. This mixed-pixel effect has also been reported by previous yield estimation studies [20], [21]. The use of CDL data for classifying 1 km crop planting areas may also create mixed pixels. However, we have minimized this uncertainty as much as possible by employing the per-pixel proportion discrimination method [45].

Furthermore, we observed that counties with higher prediction uncertainty are predominantly located in the western and southern Corn Belt, where crops experience more severe heat and drought during the summer [6], [57]. This suggests that the abnormal climate conditions can also impact the crop

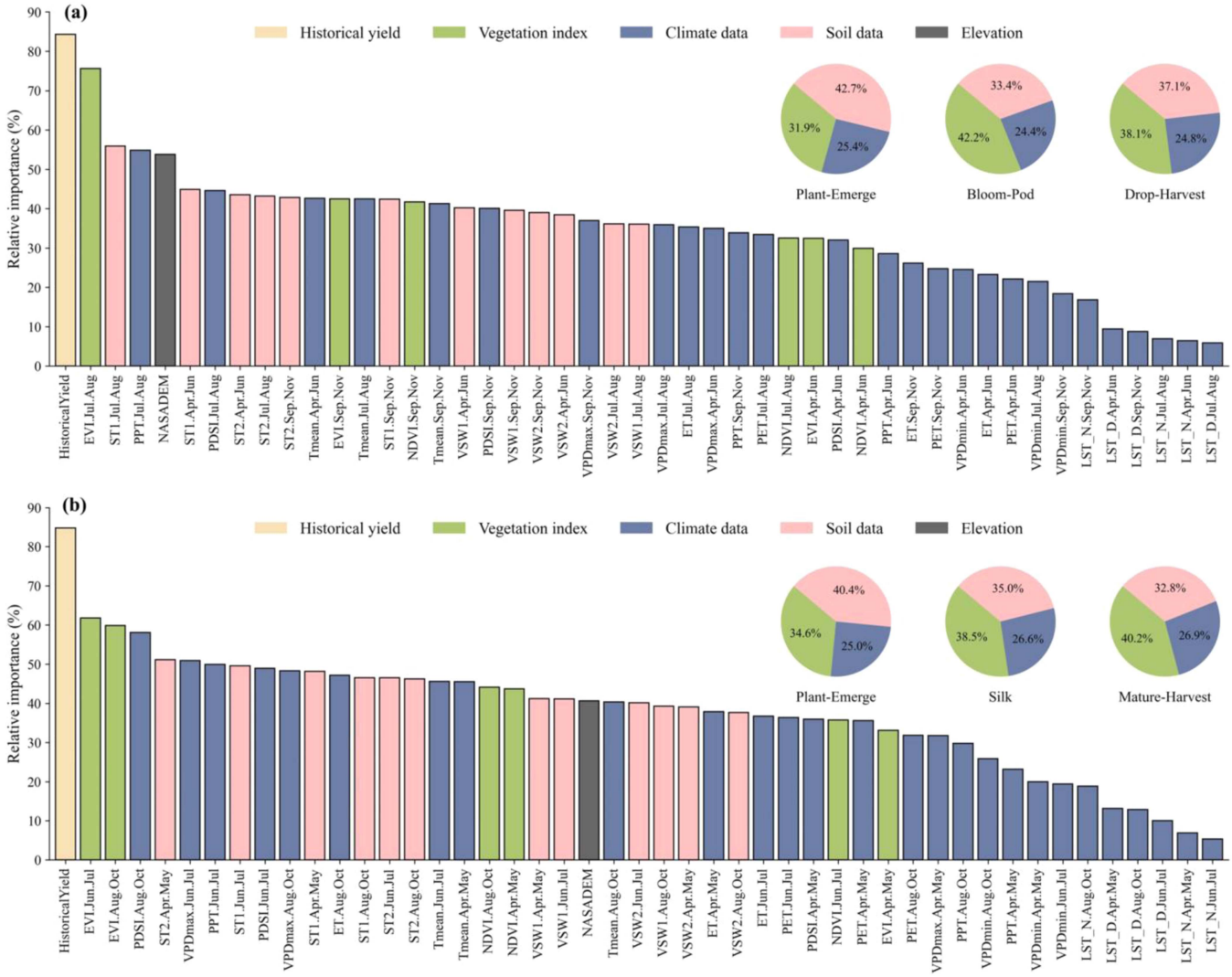

Fig. 13. Relative importance of predictors used in inversion of soybean (a) and corn (b) yields. The pie charts in the figure show the proportioned importance of the vegetation index, climate data and soil data at different growth stages.

yield prediction accuracy [20]. In the various accuracy validations carried out in our experiments, the year 2012, which was characterized by extreme drought, exhibited the lowest accuracy (see Figs. 4 and 5). This indicates that the reliability of USASoy&CornYield1km was negatively impacted by the occurrence of extreme weather events. Although we have tried to incorporate climate factors such as the drought index PDSI to represent the impact of abnormal weather conditions on crop yields, it appears that in the case of significant meteorological disasters, the introduction of a few predictive factors does not effectively improve the accuracy. This phenomenon has also been observed in previous studies and remains an issue worthy of further exploration.

Additionally, this study only predicted the spatial distribution of yield for corn and soybeans. Predicting the yield distribution for these two crops is relatively straightforward due to their regular corn-soybean rotation pattern in the U.S. Corn Belt [58]. For other crops, especially those which are intercropped and have complex distribution patterns within the same region [59], although the same methods can be directly applied, the model's performance remains to be further explored. Despite these limitations, our study still achieved satisfactory performance in crop yield estimation (see Figs. 6–9), demonstrating fine spatial resolution and covering a long time span.

D. Implications of the 1-km Gridded Crop Yield Dataset

The 1-km resolution crop yield dataset generated in this study holds significant implications for both the scientific community and local agricultural management.

For the scientific community, this dataset offers a valuable resource for conducting detailed studies on crop yield variability, climate change impacts on agriculture, and the development of new agricultural models [60]. The high spatial and temporal resolution allows researchers to perform fine-scale analyses and understand the dynamics of crop yields at a more granular level, facilitating more accurate forecasting and policy-making [61]. For instance, researchers can utilize this dataset to examine the

effects of climatic factors, such as temperature and precipitation variability, on crop yields over time, and to develop more precise models for predicting future yields under different climate scenarios. Furthermore, this dataset can aid in improving our understanding of the interactions between crop growth and environmental factors, supporting the development of more resilient agricultural systems in the face of climate change.

For local agricultural management, this dataset provides a powerful tool to optimize resource allocation, enhance decision-making, and improve overall productivity [62], [63]. By offering detailed insights into yield variability within smaller administrative units, such as counties or even individual fields, farmers and agricultural managers can identify high-yield and low-yield areas. This enables targeted interventions, such as precision fertilization, irrigation management, and pest control, ultimately leading to more efficient use of resources and increased crop yields [64]. For example, understanding yield variability at a fine spatial scale allows for the implementation of site-specific management practices that can address the unique needs of different areas within a field, thereby improving productivity and sustainability. Additionally, the long-term data can help assess the impacts of past management practices and environmental changes, guiding future strategies to ensure sustainable agricultural development [65]. Historical yield data, in conjunction with environmental and management information, can help in developing adaptive management practices that mitigate the impacts of climate variability, such as droughts or heatwaves, ensuring food security and agricultural sustainability.

V. Conclusion

In this study, we developed a statistical downscaling framework to generate spatially explicit crop yield estimates by leveraging multisource environmental covariates and ensemble machine learning methods. Using the optimal Cubist model, we produced a soybean and corn yield distribution dataset (USASoy&CornYield1km) for the U.S. Corn Belt from 2006 to 2021 at a 1 km resolution. The crop yield estimation model exhibited satisfactory performance, with R^2 values ranging from 0.70 to 0.89 (average of 0.80) for corn and 0.74 to 0.90 (average of 0.81) for soybeans. Compared to other publicly available crop yield datasets, USASoy&CornYield1km provides higher spatio-temporal resolution and shows high spatial consistency with county-level yield statistics. To assess the pixel-scale reliability of USASoy&CornYield1km, we compared it with the SPAM dataset, finding correlations of up to 0.78 for corn and 0.84 for soybeans. Further comparison with GOSIF GPP data across the entire study period demonstrates its capability to characterize regional crop productivity from 2006 to 2021. Moreover, spatial uncertainty analysis shows that the estimation uncertainty is 14.04% for soybeans and 20.49% for corn, indicating a generally low level of uncertainty in the USASoy&CornYield1km dataset.

Overall, our USASoy&CornYield1km dataset successfully downscaled yield data from county-level administrative units to kilometer-scale high-resolution data, overcoming the limitations of traditional yield statistical data that lack the capability for spatially detailed analysis. It reflects the spatial variability of crop yields within counties, which is beneficial for agricultural production and further academic research.

Appendix

The USASoy&CornYield1km dataset generated in this study will be shared upon reasonable request.

Acknowledgment

The authors would like to thank the data curators for providing the MODIS products, meteorological data, SPAM data, and county-level crop yield statistics by the United States Department of Agriculture.

References

[1] Z. Ji, Y. Pan, X. Zhu, D. Zhang, and J. Dai, "Prediction of corn yield in the USA corn belt using satellite data and machine learning: From an evapotranspiration perspective," *Agriculture*, vol. 12, no. 8, Aug. 2022, Art. no. 1263, doi: 10.3390/agriculture12081263.

[2] X.-P. Song et al., "Massive soybean expansion in South America since 2000 and implications for conservation," *Nature Sustainability*, vol. 4, no. 9, pp. 784–792, Sep. 2021, doi: 10.1038/s41893-021-00729-z.

[3] Y. Li et al., "A county-level soybean yield prediction framework coupled with XGBoost and multidimensional feature engineering," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 118, Apr. 2023, Art. no. 103269, doi: 10.1016/j.jag.2023.103269.

[4] Y. Ma, Z. Zhang, Y. Kang, and M. Özdoğan, "Corn yield prediction and uncertainty analysis based on remotely sensed variables using a Bayesian neural network approach," *Remote Sens. Environ.*, vol. 259, Jun. 2021, Art. no. 112408, doi: 10.1016/j.rse.2021.112408.

[5] Food and Agriculture Organization of the United Nations (FAO), "FAOSTAT," 2018. Accessed: Dec. 7, 2023. [Online]. Available: https://www.fao.org/faostat/zh/#data/QCL

[6] D. B. Lobell et al., "Greater sensitivity to drought accompanies maize yield increase in the U.S. midwest," *Science*, vol. 344, no. 6183, pp. 516–519, May 2014, doi: 10.1126/science.1251423.

[7] E. Vogel et al., "The effects of climate extremes on global agricultural yields," *Environ. Res. Lett.*, vol. 14, no. 5, May 2019, Art. no. 054010, doi: 10.1088/1748-9326/ab154b.

[8] R. P. Rötter, M. Appiah, E. Fichtler, K. C. Kersebaum, M. Trnka, and M. P. Hoffmann, "Linking modelling and experimentation to better capture crop impacts of agroclimatic extremes—A review," *Field Crops Res.*, vol. 221, pp. 142–156, May 2018, doi: 10.1016/j.fcr.2018.02.023.

[9] S. Fritz et al., "A comparison of global agricultural monitoring systems and current gaps," *Agricultural Syst.*, vol. 168, pp. 258–272, 2019, doi: 10.1016/j.agsy.2018.05.010.

[10] Y. Li, K. Guan, G. D. Schnitkey, E. DeLucia, and B. Peng, "Excessive rainfall leads to maize yield loss of a comparable magnitude to extreme drought in the United States," *Glob. Change Biol.*, vol. 25, no. 7, pp. 2325–2337, 2019, doi: 10.1111/gcb.14628.

[11] H. Jiang et al., "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. Change Biol.*, vol. 26, no. 3, pp. 1754–1766, 2020, doi: 10.1111/gcb.14885.

[12] D. M. Johnson et al., "USA crop yield estimation with MODIS NDVI: Are remotely sensed models better than simple trend analyses?," *Remote Sens.*, vol. 13, no. 21, Oct. 2021, Art. no. 4227, doi: 10.3390/rs13214227.

[13] C. Zhang, L. Di, L. Lin, and L. Guo, "Machine-learned prediction of annual crop planting in the US corn belt based on historical crop planting maps," *Comput. Electron. Agriculture*, vol. 166, Nov. 2019, Art. no. 104989, doi: 10.1016/j.compag.2019.104989.

[14] C. Müller et al., "The global gridded crop model intercomparison phase 1 simulation dataset," *Sci. Data*, vol. 6, no. 1, May 2019, Art. no. 50, doi: 10.1038/s41597-019-0023-8.

[15] T. Iizumi and T. Sakai, "The global dataset of historical yields for major crops 1981–2016," *Sci. Data*, vol. 7, no. 1, Mar. 2020, Art. no. 97, doi: 10.1038/s41597-020-0433-7.

[16] Q. Yu et al., "A cultivated planet in 2010 – Part 2: The global gridded agricultural-production maps," *Earth Syst. Sci. Data*, vol. 12, no. 4, pp. 3545–3572, Dec. 2020, doi: 10.5194/essd-12-3545-2020.

[17] *Global Agro-Ecological Zone V4 – Model Documentation*. FAO, Rome, Italy, 2021, doi: 10.4060/cb4744en.

[18] D. B. Lobell, D. Thau, C. Seifert, E. Engle, and B. Little, "A scalable satellite-based crop yield mapper," *Remote Sens. Environ.*, vol. 164, pp. 324–333, Jul. 2015, doi: 10.1016/j.rse.2015.04.021.

[19] M. Cheng et al., "High-resolution crop yield and water productivity dataset generated using random forest and remote sensing," *Sci. Data*, vol. 9, no. 1, Oct. 2022, Art. no. 641, doi: 10.1038/s41597-022-01761-0.

[20] Y. Zhao et al., "ChinaWheatYield30m: A 30 m annual winter wheat yield dataset from 2016 to 2021 in China," *Earth Syst. Sci. Data*, vol. 15, no. 9, pp. 4047–4063, Sep. 2023, doi: 10.5194/essd-15-4047-2023.

[21] H. Wu et al., "AsiaRiceYield4km: Seasonal rice yield in Asia from 1995 to 2015," *Earth Syst. Sci. Data*, vol. 15, no. 2, pp. 791–808, Feb. 2023, doi: 10.5194/essd-15-791-2023.

[22] T. van Klompenburg, A. Kassahun, and C. Catal, "Crop yield prediction using machine learning: A systematic literature review," *Comput. Electron. Agriculture*, vol. 177, Oct. 2020, Art. no. 105709, doi: 10.1016/j.compag.2020.105709.

[23] Y. Cai et al., "Integrating satellite and climate data to predict wheat yield in Australia using machine learning approaches," *Agricultural Forest Meteorol.*, vol. 274, pp. 144–159, Aug. 2019, doi: 10.1016/j.agrformet.2019.03.010.

[24] T. Sakamoto, "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. Photogrammetry Remote Sens.*, vol. 160, pp. 208–228, Feb. 2020, doi: 10.1016/j.isprsjprs.2019.12.012.

[25] K. Gavahi, P. Abbaszadeh, and H. Moradkhani, "DeepYield: A combined convolutional neural network with long short-term memory for crop yield forecasting," *Expert Syst. Appl.*, vol. 184, Dec. 2021, Art. no. 115511, doi: 10.1016/j.eswa.2021.115511.

[26] J. Cao et al., "Integrating multi-source data for rice yield prediction across China using machine learning and deep learning approaches," *Agricultural Forest Meteorol.*, vol. 297, Feb. 2021, Art. no. 108275, doi: 10.1016/j.agrformet.2020.108275.

[27] S. Jeong, J. Ko, and J.-M. Yeom, "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.*, vol. 802, Jan. 2022, Art. no. 149726, doi: 10.1016/j.scitotenv.2021.149726.

[28] United States Department of Agriculture (USDA), USDA/NASS QuickStats Ad-hoc Query Tool, 2022. Accessed: Dec. 7, 2023. [Online]. Available: https://quickstats.nass.usda.gov/

[29] X. Li and J. Xiao, "Mapping photosynthesis solely from solar-induced chlorophyll fluorescence: A global, fine-resolution dataset of gross primary production derived from OCO-2," *Remote Sens.*, vol. 11, no. 21, Oct. 2019, Art. no. 2563, doi: 10.3390/rs11212563.

[30] C. Boryan, Z. Yang, R. Mueller, and M. Craig, "Monitoring US agriculture: The US department of agriculture, national agricultural statistics service, cropland data layer program," *Geocarto Int.*, vol. 26, no. 5, pp. 341–358, 2011, doi: 10.1080/10106049.2011.562309.

[31] W. Han, Z. Yang, L. Di, and R. Mueller, "CropScape: A web service based application for exploring and disseminating US conterminous geospatial cropland data products for decision support," *Comput. Electron. Agriculture*, vol. 84, pp. 111–123, Jun. 2012, doi: 10.1016/j.compag.2012.03.005.

[32] B. Maestrini and B. Basso, "Drivers of within-field spatial and temporal variability of crop yield across the US midwest," *Sci. Rep.*, vol. 8, no. 1, Oct. 2018, Art. no. 14833, doi: 10.1038/s41598-018-32779-3.

[33] Y. Shao, J. B. Campbell, G. N. Taff, and B. Zheng, "An analysis of cropland mask choice and ancillary data for annual corn yield forecasting using MODIS data," *Int. J. Appl. Earth Observation Geoinf.*, vol. 38, pp. 78–87, Jun. 2015, doi: 10.1016/j.jag.2014.12.017.

[34] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.

[35] M. Belgiu and L. Drăguţ, "Random forest in remote sensing: A review of applications and future directions," *ISPRS J. Photogrammetry Remote Sens.*, vol. 114, pp. 24–31, Apr. 2016, doi: 10.1016/j.isprsjprs.2016.01.011.

[36] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, Aug. 2016, pp. 785–794, doi: 10.1145/2939672.2939785.

[37] A. B. Parsa, A. Movahedi, H. Taghipour, S. Derrible, and A. (Kouros) Mohammadian, "Toward safer highways, application of XGBoost and SHAP for real-time accident detection and feature analysis," *Accident Anal. Prev.*, vol. 136, Mar. 2020, Art. no. 105405, doi: 10.1016/j.aap.2019.105405.

[38] J. R. Quinlan, "Improved use of continuous attributes in C4.5," *J Artif. Intell. Res.*, vol. 4, pp. 77–90, Mar. 1996, doi: 10.1613/jair.279.

[39] R. Houborg and M. F. McCabe, "A hybrid training approach for leaf area index estimation via Cubist and random forests machine-learning," *ISPRS J. Photogrammetry Remote Sens.*, vol. 135, pp. 173–188, Jan. 2018, doi: 10.1016/j.isprsjprs.2017.10.004.

[40] D. M. Johnson, "An assessment of pre- and within-season remotely sensed variables for forecasting corn and soybean yields in the United States," *Remote Sens. Environ.*, vol. 141, pp. 116–128, Feb. 2014, doi: 10.1016/j.rse.2013.10.027.

[41] Y. Li et al., "Toward building a transparent statistical model for improving crop yield prediction: Modeling rainfed corn in the U.S," *Field Crops Res*, vol. 234, pp. 55–65, Mar. 2019, doi: 10.1016/j.fcr.2019.02.005.

[42] Y. Kang, M. Ozdogan, X. Zhu, Z. Ye, C. Hain, and M. Anderson, "Comparative assessment of environmental variables and machine learning algorithms for maize yield prediction in the US midwest," *Environ. Res. Lett.*, vol. 15, no. 6, Jun. 2020, Art. no. 064005, doi: 10.1088/1748-9326/ab7df9.

[43] United States Department of Agriculture (USDA), United States - Crop Calendar, 2022. Accessed: Dec. 16, 2023. [Online]. Available: https://ipad.fas.usda.gov/rssiws/al/crop_calendar/us.aspx

[44] United States Department of Agriculture (USDA), United States - National Agricultural Statistics Service - Crop Progress and Condition, 2022. Accessed: Dec. 16, 2023. [Online]. Available: https://www.nass.usda.gov/Charts_and_Maps/Crop_Progress_&_Condition/index.php

[45] S. Skakun et al., "Early season large-area winter crop mapping using MODIS NDVI data, growing degree days information and a gaussian mixture model," *Remote Sens. Environ.*, vol. 195, pp. 244–258, Jun. 2017, doi: 10.1016/j.rse.2017.04.026.

[46] M.-F. Li, X.-P. Tang, W. Wu, and H.-B. Liu, "General models for estimating daily global solar radiation for different solar radiation zones in mainland China," *Energy Convers. Manage.*, vol. 70, pp. 139–148, Jun. 2013, doi: 10.1016/j.enconman.2013.03.004.

[47] Y. Luo et al., "Identifying the spatiotemporal changes of annual harvesting areas for three staple crops in China by integrating multi-data sources," *Environ. Res. Lett.*, vol. 15, no. 7, Jul. 2020, Art. no. 074003, doi: 10.1088/1748-9326/ab80f0.

[48] K. Guan et al., "Improving the monitoring of crop productivity using spaceborne solar-induced fluorescence," *Glob. Change Biol.*, vol. 22, no. 2, pp. 716–726, Feb. 2016, doi: 10.1111/gcb.13136.

[49] L. He et al., "From the ground to space: Using solar-induced chlorophyll fluorescence to estimate crop productivity," *Geophysical Res. Lett.*, vol. 47, no. 7, 2020, Art. no. e2020GL087474, doi: 10.1029/2020GL087474.

[50] A. B. Heinemann, P. A. J. van Oort, D. S. Fernandes, and A. de H. N. Maia, "Sensitivity of APSIM/ORYZA model due to estimation errors in solar radiation," *Bragantia*, vol. 71, pp. 572–582, 2012, doi: 10.1590/S0006-87052012000400016.

[51] D. B. Egli, "Comparison of corn and soybean yields in the United States: Historical trends and future prospects," *Agronomy J.*, vol. 100, no. S3, pp. S-79–S-88, 2008, doi: 10.2134/agronj2006.0286c.

[52] P. Grassini, J. E. Specht, M. Tollenaar, I. Ciampitti, and K. G. Cassman, "Chapter 2 - high-yield maize–soybean cropping systems in the US corn belt," in *Crop Physiology (Second Edition)*, V. O. Sadras and D. F. Calderini, Eds. San Diego, CA, USA: Academic, 2015, pp. 17–41, doi: 10.1016/B978-0-12-417104-6.00002-9.

[53] M. L. Hunt, G. A. Blackburn, L. Carrasco, J. W. Redhead, and C. S. Rowland, "High resolution wheat yield mapping using sentinel-2," *Remote Sens. Environ.*, vol. 233, Nov. 2019, Art. no. 111410, doi: 10.1016/j.rse.2019.111410.

[54] B. Sisheber, M. Marshall, D. Mengistu, and A. Nelson, "Assimilation of earth observation data for crop yield estimation in smallholder agricultural systems," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 557–572, Nov. 2023, doi: 10.1109/JSTARS.2023.3329237.

[55] T. Iizumi et al., "Historical changes in global yields: Major cereal and legume crops from 1982 to 2006," *Glob. Ecol. Biogeogr.*, vol. 23, no. 3, pp. 346–357, Mar. 2014, doi: 10.1111/geb.12120.

[56] P. Hao, L. Di, C. Zhang, and L. Guo, "Transfer learning for crop classification with cropland data layer data (CDL) as training samples," *Sci. Total Environ.*, vol. 733, 2020, Art. no. 138869, doi: 10.1016/j.scitotenv.2020.138869.
[57] S. C. Zipper, J. Qiu, and C. J. Kucharik, "Drought effects on US maize and soybean production: Spatiotemporal patterns and historical changes," *Environ. Res. Lett.*, vol. 11, no. 9, Sep. 2016, Art. no. 094021, doi: 10.1088/1748-9326/11/9/094021.
[58] S. V. Archontoulis et al., "Predicting crop yields and soil-plant nitrogen dynamics in the US corn belt," *Crop Sci.*, vol. 60, no. 2, pp. 721–738, Mar. 2020, doi: 10.1002/csc2.20039.
[59] C. Zhang et al., "Towards automation of in-season crop type mapping using spatiotemporal crop information and remote sensing data," *Agricultural Syst.*, vol. 201, Aug. 2022, Art. no. 103462, doi: 10.1016/j.agsy.2022.103462.
[60] C. Zhang, A. Marzougui, and S. Sankaran, "High-resolution satellite imagery applications in crop phenotyping: An overview," *Comput. Electron. Agriculture*, vol. 175, Aug. 2020, Art. no. 105584, doi: 10.1016/j.compag.2020.105584.
[61] H. T. Pham, J. Awange, M. Kuhn, B. V. Nguyen, and L. K. Bui, "Enhancing crop yield prediction utilizing machine learning on satellite-based vegetation health indices," *Sensors*, vol. 22, no. 3, Jan. 2022, Art. no. 719, doi: 10.3390/s22030719.
[62] S. F. Ahmad and A. H. Dar, "Precision farming for resource use efficiency," *Resour. Use Efficiency Agriculture*, pp. 109–135, 2020. [Online]. Available: https://link.springer.com/chapter/10.1007/978-981-15-6953-1_4#citeas
[63] S. S. Kamble, A. Gunasekaran, and S. A. Gawankar, "Achieving sustainable performance in a data-driven agriculture supply chain: A review for research and applications," *Int. J. Prod. Econ.*, vol. 219, pp. 179–194, Jan. 2020, doi: 10.1016/j.ijpe.2019.05.022.
[64] B. Richard, A. Qi, and B. D. L. Fitt, "Control of crop diseases through integrated crop management to deliver climate-smart farming systems for low-and high-input crop production," *Plant Pathol.*, vol. 71, no. 1, pp. 187–206, Jan. 2022, doi: 10.1111/ppa.13493.
[65] A. Kayad et al., "Ten years of corn yield dynamics at field scale under digital agriculture solutions: A case study from North Italy," *Comput. Electron. Agriculture*, vol. 185, Jun. 2021, Art. no. 106126, doi: 10.1016/j.compag.2021.106126.

Jie Pei received the Ph.D. degree in cartography and geographical information system from the University of Chinese Academy of Sciences, Beijing, China, in 2020.

He is currently an Assistant Professor with the School of Geospatial Engineering and Science, Sun Yat-sen University, Zhuhai, China. His research interests include agricultural remote sensing, crop yield prediction, and food security.

Yaopeng Zou is currently working toward the master's degree in surveying and mapping engineering with Sun Yat-sen University, Zhuhai, China.

His research interests include agricultural remote sensing.

Yibo Liu is currently working toward the doctoral degree in cartography and geographical information system in the University of Chinese Academy of Sciences, Beijing, China.

His research interests include agricultural remote sensing.

Yinan He received the Ph.D. degree from the University of North Carolina at Charlotte, Charlotte, NC, USA, in 2019.

He is currently a Postdoc with the Lawrence Berkeley National Laboratory, Berkeley, CA, USA. His research focuses on quantitative remote sensing.

Shaofeng Tan is currently working toward the master's degree in remote sensing science and engineering in Sun Yat-sen University, Zhuhai, China.

His research interests include agricultural remote sensing.

Tianxing Wang (Member, IEEE) received the Ph.D. degree in GIS/remote sensing from Beijing Normal University, Beijing, China, in 2011.

He is currently a Professor with the School of Geospatial Engineering and Science, Sun Yat-sen University, Zhuhai, China. He has authored or coauthored more than 100 peer-reviewed papers. His research focuses on quantitative remote sensing.

Jianxi Huang (Senior Member, IEEE) received the Ph.D. degree in agricultural remote sensing from the Institute of Remote Sensing Applications, Chinese Academy of Sciences, Beijing, China, in 2006.

He is a Professor with the College of Land Science and Technology, China Agricultural University, Beijing. He has authored more than 100 articles in scientific journals. His research focuses on crop yield prediction using remote sensing and data assimilation.