ISPRS Journal of Photogrammetry and Remote Sensing 205 (2023) 50–73

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

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

A Phenology-guided Bayesian-CNN (PB-CNN) framework for soybean yield estimation and uncertainty analysis

Chishan Zhang [a], Chunyuan Diao [a,*]

[a] Department of Geography and Geographic Information Science, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA

ARTICLE INFO

Keywords:
Crop yield
Uncertainty
Phenology
Agriculture
Deep learning

ABSTRACT

Large-scale crop yield estimation is important for understanding the response of agriculture production to environmental forces and management practices, and plays a critical role in insurance designing, trade decision making, and economic planning. The empirical models (e.g., deep learning models) have been increasingly utilized for estimating crop yields with the ability to take into account a range of yield predictors and complex modeling relationships. Yet empirical estimation of crop yields still faces important challenges, particularly in accommodating spatio-temporal crop phenological development patterns as well as tackling the heterogeneity of a diversity of yield predictors. The different types of uncertainties associated with empirical yield estimations have seldom been explored. The objective of this study is to develop a Phenology-guided Bayesian-Convolutional Neural Network (PB-CNN) framework for county-level crop yield estimation and uncertainty quantification, with soybean in the US Corn Belt as a case study. The PB-CNN framework comprises three key components: Phenology Imagery construction, multi-stream Bayesian-CNN modeling, as well as feature importance (i.e., yield predictor and phenological stage) and predictive uncertainty analysis (i.e., aleatoric and epistemic uncertainty). With the innovative integration of critical crop phenological stages in modeling the crop yield response to a heterogeneous set of yield predictors (i.e., satellite-based, heat-related, water-related, and soil predictors) as well as the associated uncertainties, the developed PB-CNN framework outperforms three advanced benchmark models, achieving an average RMSE of 4.622 bu/ac, an average R^2 of 0.709, and an average bias of −2.057 bu/ac in estimating the county-level soybean yield of the US Corn Belt in testing years 2014–2018. Among the yield predictor groups, the satellite-based predictor group is the most critical in soybean yield estimation, followed by the water- and heat-related predictor groups. Throughout the growing season, the soybean blooming to dropping leaves phenological stages play a more crucial role in modeling the soybean yield. The soil predictor group as well as the early growing stages can improve the model estimation accuracy yet potentially brings more uncertainties into the yield estimation. The further uncertainty disentanglement indicates that the dominant uncertainty in yield estimation is the aleatoric uncertainty, mainly stemming from the fluctuations and variations inherent in the modeling input observations. The PB-CNN framework largely enhances our understanding of the complex soybean yield response to varying environmental conditions across crop phenological stages as well as associated uncertainties for more sustainable agricultural development.

1. Introduction

With the projected increase of the global population being about 2 billion in the next 30 years, agricultural demands are expected to drastically boost throughout the world, resulting in severe challenges to food security (Weiss et al., 2020). To meet the rising food demand, a sustainable increase in food production is required, which needs to be accompanied by more intelligent agricultural management with a comprehensive understanding of the forces constraining crop productivity (Crane-Droesch, 2018; S. Feng et al., 2021; Wei et al., 2014). Large-scale crop yield monitoring under various environmental forces allows better assessments of gaps between actual and potential yields to facilitate the optimization of farm management practices (Guan et al., 2017). Such crop yield information is also critical for the price forecasts of agricultural commodities, aiding in crop insurance designing, trade decision making, and national economic planning (Carletto et al., 2015; Sherrick et al., 2014).

Current mainstream crop yield estimation models include process-

* Corresponding author.
E-mail addresses: chishan2@illinois.edu (C. Zhang), chunyuan@illinois.edu (C. Diao).

https://doi.org/10.1016/j.isprsjprs.2023.09.025
Received 1 May 2023; Received in revised form 23 August 2023; Accepted 30 September 2023

based crop models and empirical models. Process-based crop models estimate yield through simulating soil-crop-atmospheric interactive processes (e.g., crop growth, water and energy balance, and nutrient cycling) driven by a combination of environmental and crop management factors (Basso & Liu, 2019; Bassu et al., 2014; Sheng et al., 2019; H. S. Yang et al., 2004). To simulate those complex processes, crop models usually require expert knowledge and a large amount of field observations regarding soil characteristics, weather conditions, and management practices to calibrate the models. The collection of such diverse and comprehensive field observations of adequate quality can be difficult at large scales, which may hamper the crop model calibration and induce the misspecified crop growth trajectory with yield estimation accuracy compromised (Guo et al., 2021; Huang et al., 2015; Jiang et al. (2020a)). By comparison, empirical models may overcome field data barriers and model calibration issues by building the empirical relationship between the crop yield and a variety of environmental and satellite-based predictors (Becker-Reshef et al. 2010; Jiang et al. 2020; Kaylen and Koroma 1991; Shahhosseini et al. 2020). To characterize complex yield-predictor relationships (e.g., both linear and non-linear relationships), machine learning models, particularly deep learning models, have been increasingly utilized in empirical crop yield estimation (Chlingaryan et al., 2018; Kang et al., 2020; Koirala et al., 2019). For example, the Convolutional Neural Network (CNN)-based models leverage convolutional network structures to characterize multi-level spatial texture information of yield predictors (e.g., satellite imagery and soil maps) (Barbosa et al., 2020; Kattenborn et al., 2021; Muruganantham et al., 2022; Russello, 2018; Terliksiz & Altylar, 2019). The Long Short-Term Memory (LSTM)-based models employ recurrent network structures to model the temporal dependent patterns of the sequential time series of yield predictors (e.g., time series of satellite and meteorological predictors) over the growing season (Jiang et al. (2020b); Tian et al., 2021; You et al., 2017). With uniquely designed modeling structures, deep learning-based yield estimation models have been found to outperform other empirical models (e.g., random forest) for a range of crop species (e.g., soybean, corn, wheat, and rice) in various applications (Fernandez-Beltran et al., 2021; Ma et al., 2021; Muruganantham et al., 2022; Sun et al., 2019; Tian et al., 2021).

Despite the promising estimation accuracy achieved by the deep learning-based models, empirical estimation of crop yields has its own challenges. One challenge arises from the increasing heterogeneity of yield predictors of various characteristics and resolutions. To account for the complex soil–plant-atmospheric continuum affecting crop growth and yield formation, empirical models need to accommodate various types of predictors that can reflect both crop growth status (e.g., satellite-based vegetation indices) and associated meteorological (e.g., heat and water stress) and soil (e.g., bulk density and organic carbon content) conditions. All these types of predictors are important in empirical yield estimation, yet they are heterogeneous in nature, with different characteristic associations with yields (Kang et al., 2020; Maimaitijiang et al., 2020; van Klompenburg et al., 2020; Muruganantham et al., 2022). For instance, satellite-based vegetation indices can characterize crop canopy greenness and photosynthetic capacity that directly denote crop growth status, while the meteorological heat and water stresses represent critical environmental stress conditions that negatively affect plant cell functions, leaf development, and photosynthesis. These predictors may also exhibit diverse resolutions (e.g., one-off soil maps versus daily satellite imagery) of various qualities. Despite the increasing accommodation of different types of predictors, most deep learning-based yield studies tend to characterize these predictors with a single shared deep learning structure, the capability of which in tackling the heterogeneity of yield predictors is limited.

Besides the increasingly diverse yield predictors, another challenge of empirical crop yield estimation stems from the difficulty in accommodating season-long crop phenological dynamics. At different phenological stages, the crop physiological responses (e.g., light use efficiency, photosynthesis, and evapotranspiration) to changes in meteorological conditions may vary largely, resulting in varying impacts on the change of subsequent crop yields (Cohen et al., 2021; Guo et al., 2021; Jumrani & Bhatia, 2019). For example, the same level of water stress at the reproductive stage (e.g., initiation of seed filling stage) of soybean can cause more reduction in yield than that at the vegetative stage (Jumrani & Bhatia, 2018). Many empirical yield estimation studies analyze the yield predictors through temporally aggregating them into calendar-based intervals (e.g., the monthly interval), of which different phenological stages of a crop may be mixed, particularly when the durations of phenological stages are short. The lack of explicit representation of phenology in empirical models remains an obstacle to unveiling the roles of various environmental and satellite-based predictors in estimating crop yield. In recent years, several efforts have been made to incorporate crop phenology in empirical models, including adding the timing of crop phenological events as the predictor (Butts-Wilmsmeyer et al., 2019; Guo et al., 2021; Shahhosseini et al., 2021) and normalizing the yield predictors in terms of certain phenological events (Bolton & Friedl, 2013). Despite these efforts, the phenological representation of predictors (e.g., predictors via phenological timing or limited phenological stages) in most empirical yield estimation studies is relatively simplified. Crops go through multiple phenological stages throughout the growing season, with diverse yield responses to environmental stress (Diao, 2019, 2020). Given the distinct yield response characteristics across phenological stages, the incorporation of phenological stage-specific yield predictors (e.g., satellite, meteorological, and soil predictors) along the season-long crop growth trajectory is essentially needed yet remains underexplored.

The evaluation of empirical yield estimations has been mostly focused on estimation accuracy via the comparison between estimated and reference yields. Several accuracy metrics, such as Root Mean Square Error (RMSE), coefficient of determination, and bias, have been developed for such evaluation. Yet the quantitative analysis of yield estimation uncertainties via the construction of confidence intervals of the estimated yields has been largely overlooked, despite its importance in inferring the reliability of yield estimations (Lobell et al., 2009; Ma et al., 2021; X. Wang et al., 2020). The lack of estimation uncertainties may compromise subsequent yield-relevant risk management and negatively affect the yield-based decision-making processes of agricultural, environmental, and economic significance, including agricultural market planning and government agricultural policy-making (Lencucha et al., 2020; Müller et al., 2021). The empirical yield estimation uncertainties may stem from both input data and modeling structures, and can be categorized into two types: aleatoric uncertainty and epistemic uncertainty (Depeweg et al., 2017; Hüllermeier & Waegeman, 2021; Kiureghian & Ditlevsen, 2009; Mobiny et al., 2021). Aleatoric uncertainty is a measure of the variation of data, and denotes the uncertainty caused by the noise or randomness inherent in the yield-relevant inputs, such as the errors in the crop type maps for identifying the target crop species, the fluctuations of meteorological predictors induced by extreme heat and water stress, and the potential heteroscedastic noise to the yield labels introduced by the overall yield detrending (Ma et al., 2021; X. Wang et al., 2020). By contrast, epistemic uncertainty typically refers to the uncertainty of the model (e.g., model structures and parameters), arising from incomplete knowledge of appropriate modeling systems underlying the empirical yield estimation possibly due to the lack of training data (Abdar et al., 2021; Hüllermeier & Waegeman, 2021; Kiureghian & Ditlevsen, 2009; X. Wang et al., 2020). Higher epistemic uncertainty may indicate that the training datasets are not sufficiently representative, particularly when historical yield records are limited or when environmental settings in testing datasets are significantly different from those in training datasets (Kendall & Gal, 2017; X. Wang et al., 2020). With the different characteristics of the two uncertainties, disentangling these two uncertainties could be critical for assessing the likelihood of yield estimates and evaluating the sources that cause the estimate fluctuations for subsequent yield-relevant risk management (Hüllermeier & Waegeman, 2021; Kukal & Irmak, 2018; Y.

Zhang et al., 2019). Despite the importance, there is a dearth of studies quantifying and distinguishing the uncertainties associated with the predictions of crop yields (Ma et al., 2021; X. Wang et al., 2020).

To address these aforementioned issues, the overarching goal of this study is to develop a Phenology-guided Bayesian-Convolutional Neural Network (PB-CNN) framework for the crop yield estimation and uncertainty quantification of the counties in the US Corn Belt. Using soybean in the Corn Belt as a case study, the specific objectives are to 1) innovatively construct the Phenology Imagery to incorporate the phenological stage-specific soybean yield predictors (e.g., satellite, meteorological, and soil predictors) into empirical yield estimation, 2) devise the multi-stream Bayesian-CNN architectures to integrate a diversity of yield predictors of heterogeneous nature for soybean yield and uncertainty quantification, and 3) evaluate both the aleatoric and epistemic uncertainties of PB-CNN, as well as the importance of varying phenological stages and yield predictors in soybean yield and uncertainty estimation.

2. Study area and data

2.1. Study area

The US Corn Belt is selected as the study area, which covers 12 states of the Midwestern US, including Illinois (IL), Indiana (IN), Iowa (IA), Kansas (KS), Michigan (MI), Minnesota (MN), Missouri (MO), Nebraska (NE), North Dakota (ND), Ohio (OH), South Dakota (SD), and Wisconsin (WI) (Fig. 1). It is characterized by relatively flat land with deep fertile soil. As the primary agricultural area, these states produced more than 81 % of US soybean and corn in 2020 (Naeve & Miller-Garvin, 2019). There are about 135 million acres of land planted with soybean in the study region. The areas planted with soybean range from 2 to 10 million acres across states, with the average planted soybean area of the states being 6 million acres (USDA, 2022). The average field size across states ranges from 175 to 1512 acres, with the average size of fields of the study region being 585 acres (USDA NASS, 2022). In most states (e.g., central and eastern states), soybean and corn are the dominant crops, and the average of historical soybean yields is usually higher than 40 bu/ac (shown in Fig. 1). By comparison, the western states, which are close to the arid Great Plains, have more diverse crop types (e.g., wheat and barley) with relatively lower average historical soybean yield. In this study, we select the years from 2008 to 2018 as the years of interest in consideration of the availability of various data products (e.g., soybean location maps and phenology products) for identifying soybean fields and generating diverse types of yield predictors.

2.2. Data

The records of county-level soybean yields are retrieved from the United States Department of Agriculture (USDA) National Agricultural Statistics Service (NASS) Database (https://quickstats.nass.usda.gov/). The soybean yield records of a total of 958 counties in the US Corn Belt for the years of interest (2008–2018) are collected. Yield measurements for soybean in the US are conventionally recorded in bu/ac, and 1bu/ac is equivalent to 67.25 kg/ha. Each year the Cropland Data Layer (CDL) from the USDA (https://nassgeodata.gmu.edu/CropScape/) is utilized to identify soybean field locations in the counties for retrieving soybean field-specific crop growth and environmental predictors for subsequent yield analysis (NASS, 2016). The CDL is a 30-m resolution, crop-specific land-cover data layer, which has been created annually using satellite imagery and extensive ground measurements for all the states of the continental US since 2008. The mapping accuracies of soybean in the CDLs exceed 90 % in the Corn Belt region for the study period (Boryan et al., 2011; NASS, 2016).

In this study, we collect satellite products, water and heat-related meteorological data, as well as soil data to generate the yield predictors that could take into account both crop growth status and associated environmental conditions (e.g., meteorological and soil conditions) (Table 1). For the predictor reflecting soybean growth status, we mainly consider the 2-band Enhanced Vegetation Index (EVI2), which is a widely used remotely sensed vegetation index for quantifying crop growth vigor and greenness. As an enhanced vegetation index, EVI2 reduces the influence of canopy background and atmospheric interference and is less susceptible to saturation at high biomass regions (Son et al., 2014). The EVI2 predictor in this study is generated from the Moderate Resolution Imaging Spectroradiometer (MODIS) MCD43A4 Nadir Bidirectional Reflectance Distribution Function (BRDF) Adjusted

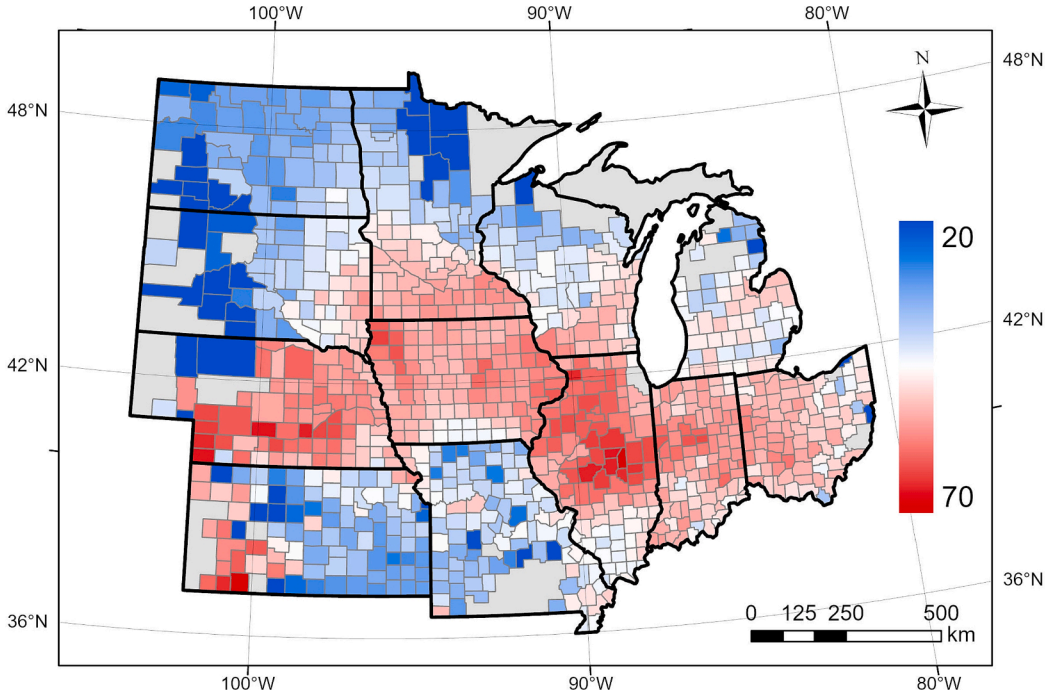

Fig. 1. Map of the study area and the average county-level soybean yield (bu/ac) over 2008–2018 in this area. Areas in light gray indicate no data available.

Table 1
Description of all input data.

Category	Variable	Spatial Resolution	Temporal Resolution	Source
Crop	Yield Record (bu/ac)	County-level	Annual	USDA Quick Statistic Database
	Cropland Map	30 *m*	Annual	NASS CDL Program
Phenology	Satellite Phenology	500 *m*	Annual	MODIS MCD12Q2
	Ground Crop Phenology	State-level	Annual	NASS CPR
Vegetation Index	EVI2	500 *m*	Daily	MODIS MCD43A4
Heat-related	LST_Day (Kelvin)	1 km	Daily	MODIS MOD11A1
	LST_Night (Kelvin)			
	Maximum air temperature(°C)	1 km	Daily	Daymet V4
	Minimum air temperature(°C)			
Water-related	Precipitation (mm)	1 km	Daily	Daymet V4
	VPDmax (hPa)	5 km	Daily	PRISM
	VPDmin (hPa)			
	Evapotranspiration ($kg/m^2/s$)	0.25°	3-Hour	GLDAS
	Potential ET (W/m^2)			
	Soil Moisture (kg/m^2)			
Soil	Clay Content (%)	250 *m*	Static	OpenLandMap
	Sand Content (%)			
	pH in H_2O			
	Bulk Density (kg/m^3)			
	Organic Carbon Content (g/kg)			

Reflectance product, which provides the nadir land surface reflectance data of 500 m spatial resolution and daily temporal resolution with view angle effects removed using the BRDF algorithm (Schaaf & Wang, 2015). The equation for calculating EVI2 is as follows:

$$EVI2 = 2.5 \times \frac{NIR - Red}{NIR + 2.4 \times Red + 1} \quad (1)$$

where Red and NIR represent the reflectance in red and near-infrared bands, respectively. We further pre-process the EVI2 time series to eliminate the influence of abnormal observations (see details in Appendix A-1). The pre-processed EVI2 time series is then composited by the average EVI2 value on 3-day intervals, which can preserve the phenological dynamic patterns, reduce the subtle fluctuations, as well as decreasing the data volume.

As for the predictors reflecting critical environmental conditions, we take into account the water and heat-related meteorological predictors, as well as the soil predictors (Table 1). The meteorological predictors consist of four heat-related predictors and six water-related predictors. The heat-related predictors include daily maximum 2-meter air temperature (Tmax), daily minimum 2-meter air temperature (Tmin), daily daytime land surface temperature (LST_Day), and daily nighttime land surface temperature (LST_Night). Tmax and Tmin are acquired from the daily 1 km Daymet dataset, and LST_Day and LST_Night are extracted from the daily 1 km MODIS MYD11A1 product. The water-related predictors include daily total precipitation (Precipitation), daily minimum vapor pressure deficit (VPDmin), daily maximum vapor pressure deficit (VPDmax), evapotranspiration (ET), potential ET (PET), and soil moisture at four depths (10, 40, 100, and 200 cm) (Rodell et al., 2004). Precipitation, VPDmin, and VPDmax are acquired from the daily 5 km Parameter-elevation Regressions on Independent Slopes Model (PRISM) dataset. ET, potential ET, and soil moisture of four depths are obtained from the 3-hour 0.25 degree Global Land Data Assimilation System (GLDAS) dataset. These heat- and water-related predictors have been employed to describe the temperature and water conditions for crop growth and analyze the influence of heat and water stress on crop production (Anderson et al., 2016; Guan et al., 2017; Ray et al., 2015). For example, VPD affects the amount of water required for crop development and carbon fixation, and a high VPD caused by water stress may lead to a subsequent reduction in crop yields (Guan et al., 2017; Kang et al., 2020).

Besides satellite and meteorological predictors, several soil-based predictors characteristic of soil conditions are also taken into account to evaluate the soil's capacity to supply water, air, and nutrients to crops (Table 1) (Pourmohammadali et al., 2019; Shirani et al., 2015). These soil predictors include clay content mass fraction, sand content mass fraction, pH in H_2O, bulk density, and organic carbon content of six depths (0, 10, 30, 60, 100, and 200 cm). These soil-based predictors are collected from OpenLandMap, which is a one-off soil product with a spatial resolution of 250 m (Hengl, 2018; Jaafar & Mourad, 2021). Among the predictors, bulk density, sand content, and clay content affect water holding capacity and air capacity for the water and oxygen supply. Soil pH and organic carbon content have a substantial effect on soil nutrient availability, and the latter could also affect the extent of soil water retention (Pourmohammadali et al., 2019; Shirani et al., 2015).

As the soybean yield response to the satellite- and environmental-based predictors may vary across its phenological stages, we further leverage the MODIS MCD12Q2 Land Cover Dynamics product to divide the growing season into different phenological stages and to analyze the sub-seasonal changes of all the yield predictors via these soybean stages. The MCD12Q2 product provides annual 500 m pixel-level estimations of key phenology phase transition dates (Ganguly et al., 2010). These transition dates are generally associated with the critical change of soybean greenness and vigor, and have been found to be closely linked with the transition of several soybean vegetative and reproductive stages. The ground-based Crop Progress Reports (CPRs) released by USDA are collected as the reference data to evaluate these satellite-based phenological transition dates (Diao et al., 2021; NASS, 2018). The CPRs document the proportion of soybean reaching several phenological stages on a weekly basis throughout the growing season at the state level.

3. Methodology

In this study, we propose a Phenology-guided Bayesian-Convolutional Neural Network (PB-CNN) framework that can integrate a diversity of crop growth and environmental predictors along soybean phenological development trajectory to estimate the yields as well as associated uncertainties. The PB-CNN framework is mainly composed of three components: Phenology Imagery construction, multi-stream Bayesian-CNN modeling, as well as the feature (i.e., yield predictor and phenological stage) and uncertainty (i.e., aleatoric and epistemic uncertainty) analysis (Fig. 2). Several methods in these components are designed particularly for the study objectives, including the construction of Phenology Imagery for incorporating season-long crop phenological dynamics, the multi-stream modeling architecture to integrate a diversity of yield predictors, and the Bayesian-CNN for crop yield and uncertainty quantification analysis. This framework is designed not only to estimate crop yields, but also to evaluate the estimation uncertainties as well as contributions of varying phenological stages and yield predictors to the empirical model predictions. The data processing based on Google Earth Engine and the Python implementation of our model are publicly available at https://github.com/rssiuiuc/PBCNN.

3.1. Phenology Imagery construction

Throughout the soybean growth trajectory, the environmental and satellite-based predictors may exhibit diverse roles across phenological

stages in estimating the yields, which makes it important to construct phenological stage-specific yield predictors. In this study, the soybean growing season is divided into different phenological stages according to the satellite-based vegetation phenology product (i.e., the MODIS MCD12Q2 product). The MCD12Q2 provides pixel-level estimations of phenology stage transition dates using the threshold-based method for

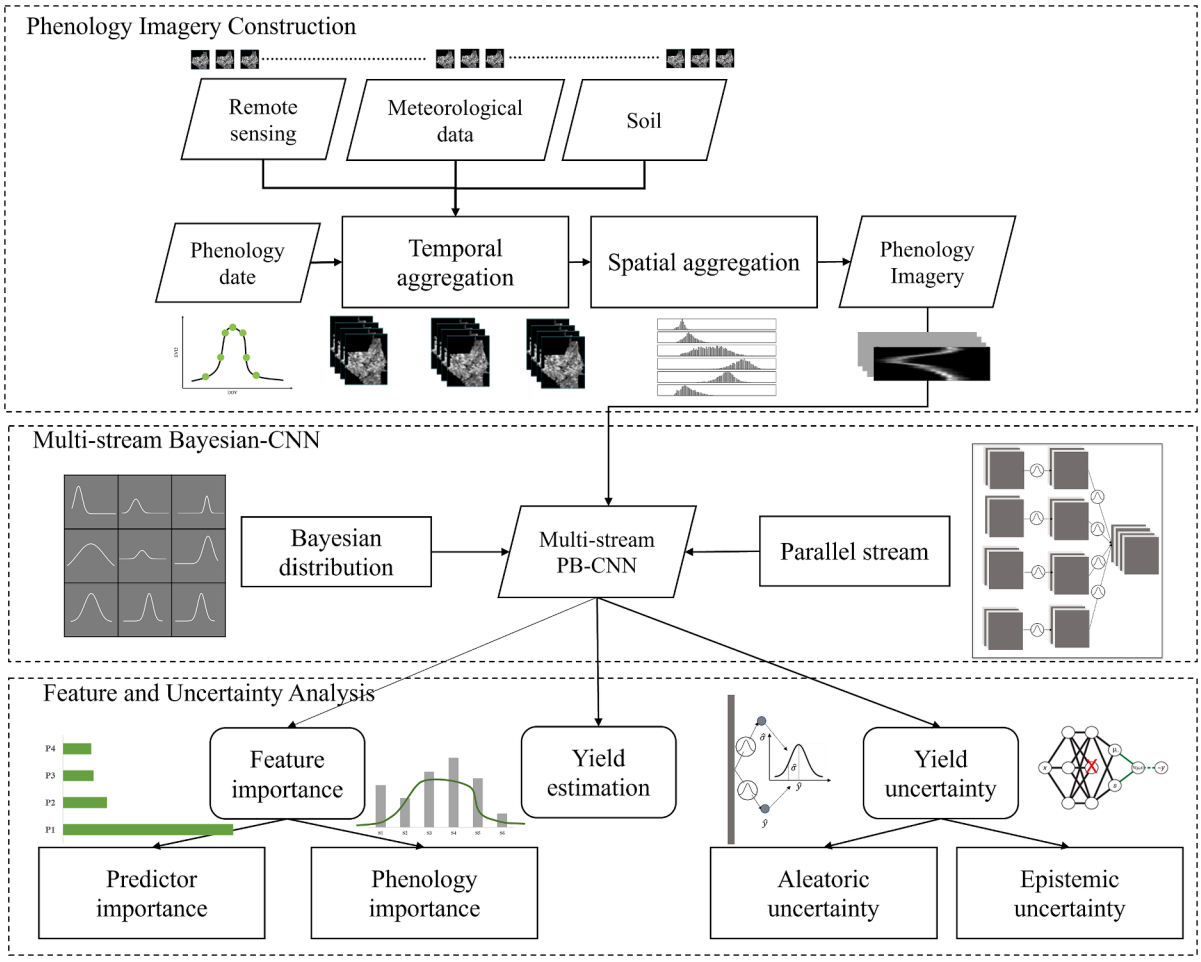

Fig. 2. Flowchart of the PB-CNN framework.

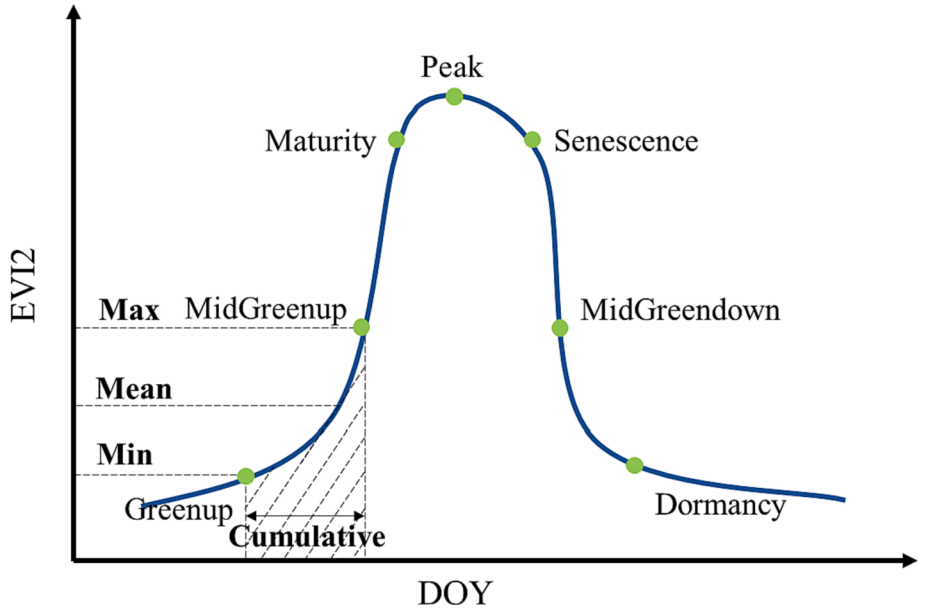

Fig. 3. Four types of temporal aggregations of the EVI2 predictor in each phenological stage, with the stage from greenup to greenup midpoint (midgreenup) as an example.

characterizing the timing of plant greenness and vigor change throughout its phenological cycle. These phenology transition dates include the onset of greenness (geenup), greenup midpoint (midgreenup), maturity, peak greenness (peak), senescence, greendown midpoint (midgreendown), and dormancy transition dates, and these transition dates can be connected with the transitioning of several soybean vegetative and reproductive stages (Diao, 2020; Diao & Li, 2022). For instance, the greenup and midgreenup dates can be associated with the timing of soybean emergence and subsequent change in leaf numbers during its vegetative stages. The rest of the transition dates (e.g., maturity, peak greenness, and midgreendown) can be linked with the timing of soybean blooming, setting pods, and dropping leaves during its reproductive stages. The connections between satellite phenology measures and field soybean growth stages are discussed in detail in Appendix A-5.

With the aforementioned seven phenology transition dates, the soybean growing season is divided into six phenological stages, namely the stage from greenup to greenup midpoint (denoted as S1), greenup midpoint to maturity (S2), maturity to peak greenness (S3), peak greenness to senescence (S4), senescence to greendown midpoint (S5), and greendown midpoint to dormancy (S6), at the pixel level. Based on the division of the phenological stages, the Phenology Imagery of each yield predictor is constructed via a combination of temporal (Step I) and spatial (Step II) aggregation.

Step I Temporal aggregation: To construct pixel-level phenological stage-specific yield predictors, all the environmental and satellite-based predictors for each pixel are temporally aggregated into the six soybean phenological stages based on the corresponding start and end dates. For each environmental predictor, the type of temporal aggregation (e.g., maximum, minimum, mean, or cumulative aggregation) is determined in terms of the predictor characteristics and its yield implications. Specifically, for the predictor of precipitation, cumulative aggregation is implemented for each phenological stage to reflect the accumulation effect of precipitation on soybean growth and productivity in each stage (Meng et al., 2017). For all the other environmental predictors, the mean aggregation is applied for each phenological stage to represent the average conditions of the predictors within each stage. As EVI2 is closely related to soybean productivity, four types of aggregations (i.e., maximum, minimum, mean, and cumulative aggregations) are employed for the EVI2 predictor within each phenological stage to comprehensively characterize soybean canopy greenness and growth status (shown in Fig. 3). Specifically, maximum and minimum aggregations of EVI2 can denote the change extent of canopy greenness within each phenological stage. The mean and cumulative aggregations of EVI2 can be indicative of soybean average greenness and aboveground biomass, respectively, and they together can further indicate the greenness variation and duration of each phenological stage (Benedetti & Rossini, 1993; Rasmussen, 1992). With the temporal aggregations, all the yield predictors of different temporal resolutions are aggregated accordingly into six soybean phenological stages per pixel and year (Fig. 3). The temporally aggregated predictor images of six phenological stages can thus be formed for each county and year for subsequent

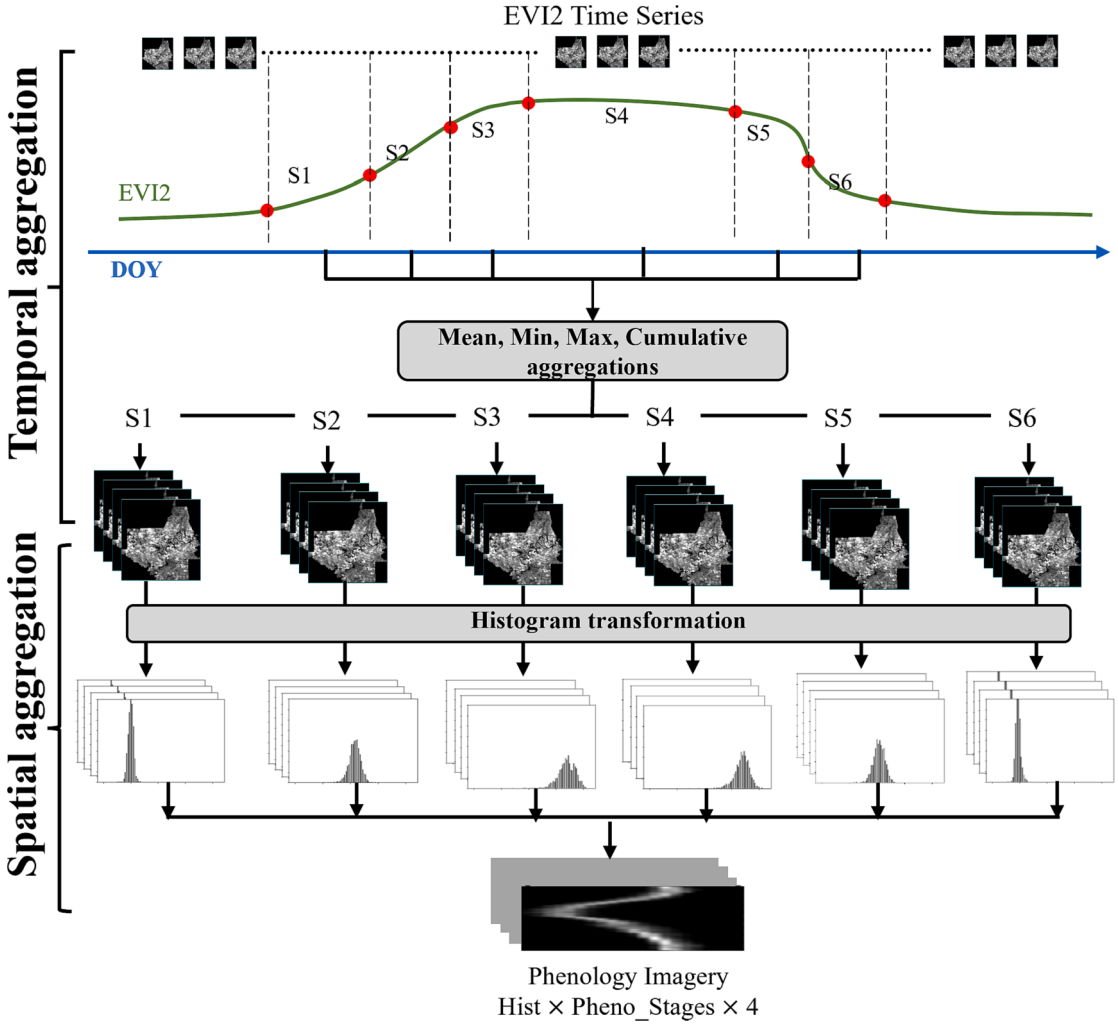

Fig. 4. The temporal and spatial aggregations for constructing the Phenology Imagery of the EVI2 predictor group.

county-level soybean yield estimations.

Step II Spatial aggregation: Due to the irregular boundary of each county, the temporally aggregated predictor images should be further spatially aggregated for each county. Most of the previous studies use the simple average spatial aggregation to gather county-level information of yield predictors. Yet this average county-level spatial aggregation fails to accommodate the variations in the values of the yield predictors and may not reflect spatial heterogeneity or changes in crop growth status or environmental conditions within a county. In this study, a histogram-based transformation is therefore employed as the spatial aggregator to characterize the distributions of yield predictor values per county and year, with both average and spatial variation information of the predictors preserved. As a spatial dimension reduction method, histogram-based transformation can also reduce the complexity of deep learning models, improve the efficiency of model training, and reduce the risk of overfitting. Assuming the whole range of pixel values in a temporally aggregated predictor image could be divided into N specific sub-ranges (i.e., the bin numbers in histograms), the predictor image at the *t-th* phenological stage will be transformed as a vector $H(t)$ with $N \times 1$ dimensions (N is set to 11 after a set of experiments). By concatenating the $H(t)$ of all six phenological stages ($t \in [1, T]$) along the soybean growth trajectory (i.e., along columns) for each yield predictor, we will construct the Phenology Imagery $P \in \mathbb{R}^{N \times T \times 1}$, where N is the bin numbers of the histogram ($N = 11$) and T is the number of phenological stages ($T = 6$). For each county and year combination, the Phenology Imagery of a predictor characterizes the corresponding distribution changes of predictor values across soybean phenological stages, and thus this image-based yield predictor is phenology stage-specific.

In this study, we consider four groups of temporally and spatially aggregated yield predictors, namely satellite-based, heat-related, water-related, and soil predictor groups (Table 1). As the predictors within a group may possess shared characteristics in understanding the crop-environment interactive system, we further concatenate the Phenology Images of all individual predictors of a group along the third dimension to form a multi-band Phenology Imagery for each predictor group. Thus we will construct four Phenology Imageries (with dimensions of $N*T*G$) per county and year, including the Phenology Imagery of the EVI2 predictor group ($N*T*G_{EVI}$; shown in Fig. 4), the heat-related predictor group ($N*T*G_{Heat}$), the water-related predictor group ($N*T*G_{Water}$), and the soil predictor group ($N*T*G_{Soil}$). The first dimension N of Phenology Imagery represents the number of histogram bins ($N = 11$). The second dimension T is the number of phenological stages ($T = 6$). The third dimension G denotes the number of predictors in each predictor group. For the EVI2 group, the number of predictors is four (i.e., $G_{EVI} = 4$), which represents four different temporal aggerated EVI2 predictors. The heat-related group has four predictors (i.e., $G_{Heat} = 4$), including two daily air temperature predictors (maximum and minimum), as well as two daily land surface temperature predictors (daytime and nighttime). The water-related group includes nine predictors (i.e., $G_{Water} = 9$), namely precipitation, VPDmin, VPDmax, ET, PET, and four soil moisture predictors of different depths. As for the soil group, given that all five soil predictors are collected at six different depths, the total number of predictors is 30 (i.e., $G_{Soil} = 30$). Since soil predictors are static across phenological stages, they are spatially aggregated and repeated at each phenological stage to ensure consistency with other predictor groups. The constructed Phenology Imageries of four predictor groups per county and year will be the input of the subsequent deep learning model for county-level yield estimations.

3.2. Multi-stream Phenology-guided Bayesian-CNN model

To tackle the heterogeneity of multi-source yield predictors and model the estimation uncertainty, a two-part PB-CNN model architecture is designed in this study (Fig. 5). Part I includes a multi-stream architecture to extract and fuse the features from the constructed Phenology Imageries with each stream's sub-network optimized for the imagery of each predictor group. Part II is designed to model the crop yield as well as the predictive uncertainty based on the features extracted and fused from part I.

1) Part I for multi-stream feature fusion: The first part of our model encompasses a multi-stream architecture, which is employed to extract

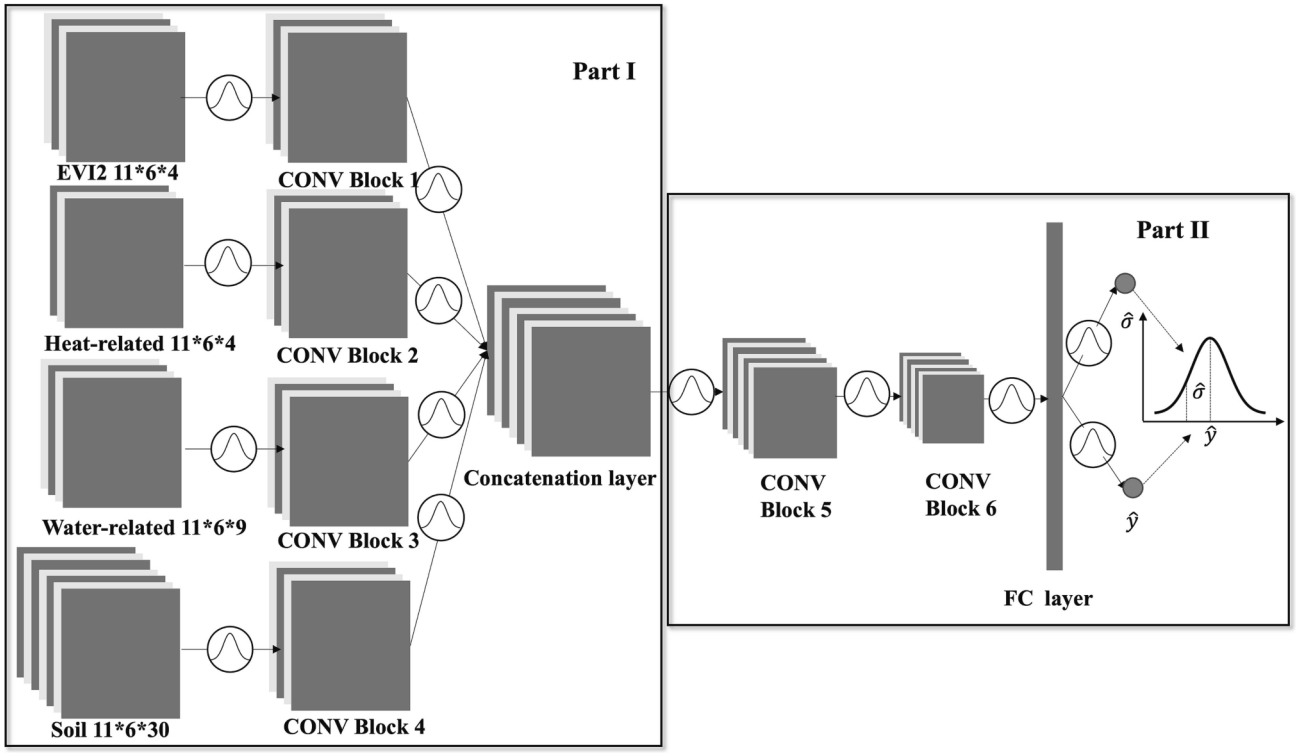

Fig. 5. The architecture of the multi-stream PB-CNN.

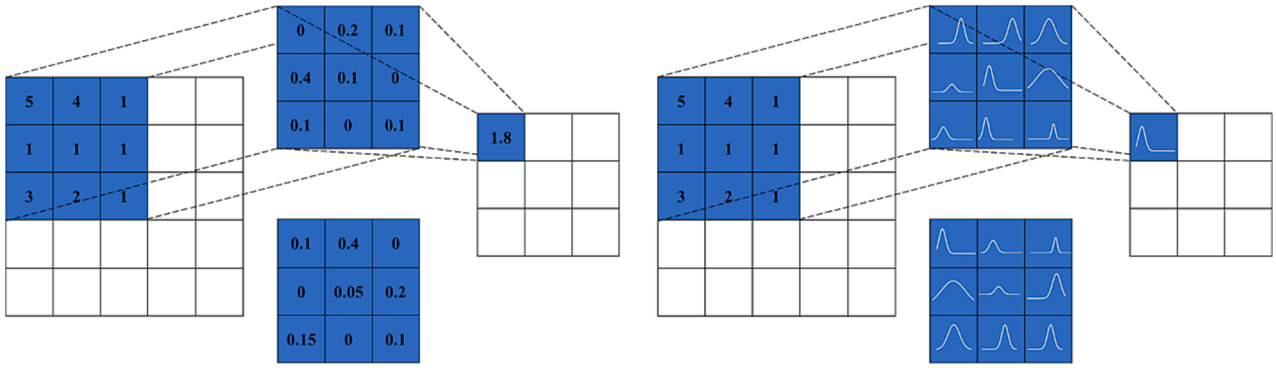

Fig. 6. Comparison between conventional and Bayesian CNNs. Left: conventional CNN kernels with point estimates. Right: Bayesian CNN kernels with probability distributions.3.

and learn the intricate features from the Phenology Imageries of yield predictor groups. Considering the heterogeneity of yield predictor groups, this part of PB-CNN is composed of a four-parallel-stream architecture, with the sub-network in a stream independently designed for a yield predictor group (i.e., satellite-based, heat-related, water-related, or soil predictor group) (Fig. 5). The Phenology Imagery of a predictor group (generated in Section 3.1) will be the input for the sub-network of the corresponding stream of PB-CNN.

Each stream's sub-network contains one independent convolutional block, which consists of one convolutional layer, one batch normalization layer, and one dropout layer. The latter two layers are employed to reduce overfitting and speed up the learning process. All convolutional layers within part I adopt the 1×1 convolutional kernels suggested by Lin et al. (2013) and Yu et al. (2017), which can be viewed as cross-channel pooling processes to learn salient and distinct feature representations across channels of a predictor group. For example, the cross-channel pooling for the satellite-based predictor group can integrate the four types of aggregated EVI2 predictors to extract discriminative features for comprehensively representing the crop growth status. After the independent convolutional operation of each stream, the extracted features of the four parallel streams are then fused together by a concatenation layer.

2) Part II for yield and uncertainty modeling: Part II of PB-CNN is composed of several convolutional blocks as well as one specially designed fully connected layer with two outputs. Each convolutional block consists of one convolutional layer with a kernel size of 3×3 and a stride of 1, one batch normalization layer, and one dropout layer. Through the convolutional blocks, the fused features from part I are further integrated with distinct features important for yield and uncertainty estimations being learned. The fully connected layer includes two outputs $[y, \sigma]$, namely the mean y and the standard deviation σ of yield estimate (Fig. 5). The yield estimate typically follows the Gaussian distribution to account for the observation noise or randomness stemming from different yield predictor inputs, with the mean y denoting the average yield estimate and the standard deviation σ characterizing the uncertainty/noise in the input observations (Kendall & Gal, 2017; Mobiny et al., 2021; X. Wang et al., 2020). To model the predictive uncertainty of model parameters, the Bayesian theory is employed in the PB-CNN. Unlike the CNN models, all the weights W in the PB-CNN model are represented by probability distributions (shown in Fig. 6), which can be used to characterize the variance of the model weights. According to the Bayesian theory, given the input x and the training data $D=\{X, Y\}=\{x_i, y_i\}_{i=1}^{N}$, the outputs $[y, \sigma]$ of PB-CNN will be each represented as predictive distributions (i.e., $p(y|x, X, Y)$ and $p(\sigma|x, X, Y)$) by integrating the model likelihoods (i.e., $p(y|x, W)$ and $p(\sigma|x, W)$) over the posterior distribution of the model weights W (i.e., $p(\mathrm{W}|X, Y)$) (Eqs. (2) and (3)):

$$p(y|x, X, Y) = \int p(y|x, \mathrm{W}) p(\mathrm{W}|X, Y) d\mathrm{W} \tag{2}$$

$$p(\sigma|x, X, Y) = \int p(\sigma|x, \mathrm{W}) p(\mathrm{W}|X, Y) d\mathrm{W} \tag{3}$$

With the predictive distributions of mean y and standard deviation σ of yield estimates, the total predictive uncertainty can then be calculated. This total uncertainty comprises both the aleatoric uncertainty and the epistemic uncertainty. The aleatoric uncertainty measures the uncertainty of data (e.g., the output standard deviation σ in the study), usually stemming from the noise or randomness in the input observations. By contrast, the epistemic uncertainty corresponds to the uncertainty in the model structures and parameters (e.g., the variation of W in the study), caused by a lack of knowledge about the best model underlying the yield estimations given the collected data. With the PB-CNN modeling design and Bayesian inference, the aleatoric uncertainty can be calculated as the mean of $p(\sigma|x, X, Y)$ (i.e., $\mathrm{E}[\sigma]$), and the epistemic uncertainty can be calculated as the standard deviation of $p(y|x, X, Y)$ (i.e., $\sigma(y)$) (Abdar et al., 2021; Depeweg et al., 2017; Gal & Ghahramani, 2016; Hüllermeier & Waegeman, 2021; Kendall & Gal, 2017). The total uncertainty can thus be calculated using the aleatoric and epistemic uncertainty (Eq. (4))

$$\sigma_{total} = \sqrt{\left(\sigma_{p(\mathrm{W}|X,Y)}(y)\right)^2 + \left(\mathrm{E}_{p(\mathrm{W}|X,Y)}[\sigma]\right)^2} \tag{4}$$

where σ_{total} is the total predictive uncertainty of the estimated yield; y and σ are the estimated mean and standard deviation outputs of the fully connected layer of PB-CNN, respectively. Given W $p(\mathrm{W}|X, Y)$, $\sigma_{p(\mathrm{W}|X,Y)}(y)$ is the epistemic uncertainty, measured as the standard deviation of the estimated mean value y; $\mathrm{E}_{p(\mathrm{W}|X,Y)}[\sigma]$ is the aleatoric uncertainty, calculated as the mean value of the estimated standard deviation σ.

However, due to a large number of parameters of Bayesian-CNN models, the posterior distribution $p(\mathrm{W}|X, Y)$ cannot be evaluated analytically, and the estimations of $p(y|x, X, Y)$ and $p(\sigma|x, X, Y)$ are typically approximated using the Monte Carlo (MC) sampling approach. MC sampling can approximately estimate $p(y|x, X, Y)$ and $p(\sigma|x, X, Y)$ by generating T sets of $[y, \sigma]$ pairs $\{[y_1, \sigma_1], \cdots [y_T, \sigma_T]\}$ based on T sets of network weights $\{\mathrm{W}_1, \cdots, \mathrm{W}_T\}$ whose empirical distribution (i.e., $q_\theta^*(\mathrm{W})$, parameterized by θ) approaches $p(\mathrm{W}|X, Y)$. In practice, the sets of network weights $\{\mathrm{W}_1, \cdots, \mathrm{W}_T\}$ can be generated by dropout inference, which is conducted via the random dropout of neurons during both the training and the testing stages, resulting in different weight configurations (Gal & Ghahramani, 2015, 2016; Shridhar et al., 2019). Thus, the final predictive distributions can be approximated by the integration of the sets of MC simulations, and Eq. (4) can be rewritten as Eq. (5) (Depeweg et al., 2017; Kendall & Gal, 2017):

$$\sigma_{total} = \sqrt{\left(\sigma_{q^*_\theta(W)}(y)\right)^2 + \left(\mathrm{E}_{q^*_\theta(W)}[\sigma]\right)^2} \tag{5}$$

$$\sigma_{q^*_\theta(W)}(y) \approx \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(\widehat{y}_t - \frac{1}{T}\sum_{t=1}^{T}\widehat{y}_t\right)^2} \tag{6}$$

$$\mathrm{E}_{q^*_\theta(W)}[\sigma] \approx \frac{1}{T}\sum_{t=1}^{T}\widehat{\sigma}_t \tag{7}$$

where T represents the number of MC samplings; $\widehat{y}_t$ and $\widehat{\sigma}_t$ are the estimated mean and standard deviation outputs of the fully connected layer of PB-CNN at the t-th simulation when $W = W_t$. As PB-CNN includes two outputs, the weights of PB-CNN are optimized by using a simplified negative log-likelihood loss function (Eq. (8)) (Kendall & Gal, 2017):

$$L(W) = \sum_{i=1}^{N}\left(\frac{1}{2\widehat{\sigma}_i^2}\|\widehat{y}_i - y_i\|^2 + \frac{1}{2}log\widehat{\sigma}_i^2\right) \tag{8}$$

where $\widehat{y}_i$ and y_i denote the estimated mean yield and observed yield for sample i, respectively. N is the number of samples in the training data. $\widehat{\sigma}_i$ denotes the estimated standard deviation for sample i.

3.3. Feature and uncertainty analysis

3.3.1. Feature importance

In this study, we investigate the importance of each yield predictor group as well as each phenological stage in estimating the soybean yield using the permutation importance feature selection method (Breiman, 2001; Shahhosseini et al., 2020; X. Wang et al., 2020). As a non-linear model inspection method, the permutation feature selection method evaluates the relative feature importance through measuring the change in model performance caused by the random shuffling of a feature (see details in Appendix A-2). Given the similar characteristics of the predictors within a group associated with the yield, we assess the relative importance of each predictor group in soybean yield estimation, instead of each individual yield predictor. Besides, we also investigate the contributions of each predictor group in each phenological stage with the permutation feature importance method to comprehensively unravel the roles of diverse types of predictors throughout the soybean growing cycle in yield estimations.

3.3.2. Uncertainty analysis

As introduced in Section 3.2, the total predictive uncertainty comprises both aleatoric uncertainty and epistemic uncertainty, and these two types of uncertainties of PB-CNN are modeled in this study. We first evaluate the spatial distributions of county-level yield estimation uncertainties of PB-CNN, including total predictive uncertainty, aleatoric uncertainty, and epistemic uncertainty, and compare the spatial patterns of these three uncertainty measures. We also summarize these uncertainty measures at the state level to identify the regions with relatively high aleatoric or epistemic uncertainty, which is further analyzed under the associated environmental conditions characterized by several representative yield predictors. Moreover, as aleatoric uncertainty comes from the input observations, we explore the contribution of the inputs of each yield predictor group as well as each phenological stage to the aleatoric uncertainty using the permutation importance method. Similar to the feature importance analysis in Section 3.3.1, we analyze the importance of each yield predictor group, each phenological stage, as well as their combination (per yield predictor group and phenological stage), by measuring the aleatoric uncertainty change caused by the corresponding random permutation.

3.4. Model training and evaluation

To evaluate the devised model, we compare the performance of PB-CNN with that of three advanced benchmark models for conducting county-level soybean yield estimations. The hyperparameter settings of PB-CNN and model training processes are introduced in Appendix A-3. The benchmark models include Support Vector Regression (SVR), Random Forest (RF), and LSTM (see the detailed model settings in Appendix A-4). To ensure a relatively comprehensive comparison with consideration of the size of training samples, we assess the performance of all these models for five testing years (2014–2018). For each testing year, all the county-level data of corresponding historical years are utilized for training the models. Given that the testing year 2018 takes into account more historical years (2008–2017) for training the models, we subsequently leverage this testing year to carry out a more comprehensive evaluation of PB-CNN in terms of both model accuracy and uncertainty, as well as conducting feature importance and uncertainty analysis. In total, there are 8,202 training data samples from 2008 to 2017 and 728 testing data samples in 2018 for the study site.

The RMSE, coefficient of determination (R^2), and bias are selected as the metrics to evaluate the model performance. The three metrics are calculated as follows:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\widehat{y}_i - y_i)^2} \tag{9}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(\widehat{y}_i - y_i)^2}{\sum_{i=1}^{N}(\bar{y} - y_i)^2} \tag{10}$$

$$Bias = \sum_{i=1}^{N}\frac{(y_i - \widehat{y}_i)}{N} \tag{11}$$

where N denotes the number of samples; y_i is the observed yield and $\widehat{y}_i$ is the estimated yield of the sample i; $\bar{y}$ denotes the mean value of all observed yields. Besides yield estimations, the developed framework also focuses on evaluating estimation uncertainties of crop yields. As described in Section 3.2, the results of 1000 MC simulations are used to calculate the predictive uncertainties (Eqs. 5–7) (Gal & Ghahramani, 2016). We also calculate the P-factor to assess the goodness of the total predictive uncertainty of the model (Liu et al., 2017; Sheng et al., 2019; J. Yang et al., 2008). The P-factor is defined as the percentage of observed yield enveloped within the 95 % confidence interval bounded by the predictive uncertainty and is calculated using Eq. (12):

$$\mathrm{P-factor} = \frac{NQ_{in}}{n} \times 100 \tag{12}$$

where NQ_{in} is the number of observations enveloped by the 95 % confidence interval, and n is the number of observations.

4. Framework evaluation

With the construction of Phenology Imagery in the PB-CNN framework, we first evaluate the ability of the satellite-based vegetation phenology product in characterizing the soybean growing stages using the ground-based CPRs. Specifically, we compare the medians of satellite-derived phenology stage transition dates with the CPR-documented median dates of soybean growing stages (i.e., emerged, blooming, setting pods, dropping leaves, and harvest stages) from 2008 to 2018 at the state-level for our study site.

We also evaluate the effectiveness of multi-stream architecture as well as the influence of different predictor groups by designing three combinations of input predictors (i.e., PB-CNN_EVI2, PB-CNN_EVI2&MET, and PB-CNN_ALL): I) PB-CNN_EVI2: only the satellite-based predictor group (i.e., four EVI2 predictors) is employed in the PB-CNN model to estimate the soybean yield. The streams of other predictor

groups are removed from PB-CNN accordingly. II) PB-CNN_EVI2&MET: In this model setting, satellite-based and meteorological water- and heat-related predictor groups are utilized as the modeling input. The stream of the soil predictor group is removed from PB-CNN. III) PB-CNN_ALL: All four yield predictor groups are employed as the input under this modeling setting (Fig. 5). For PB-CNN_EVI2 and PB-CNN_EVI2&MET, only sub-networks of corresponding predictor groups are kept in the model architectures. For each of these three combinations of input predictors, we further compare the performance of the multi-stream architecture and that of the corresponding single-stream architecture. The single-stream architecture is formed with all sub-networks of yield predictor groups merged together as one stream, and maintains the same number of convolutional filters as the multi-stream counterpart. It's worth noting that multi-stream and single-stream PB-CNN_EVI2 share the same architecture because only one predictor group is utilized as the input.

The contribution of phenophase-guided structure is investigated by comparing the performance of PB-CNN model with that of two model variants, which are based on the same predictors yet with calendar-based intervals (i.e., 8-day and monthly intervals). The temporal aggregation steps of the phenophase, 8-day, and monthly composites are 6, 30, and 8, respectively. Except for temporal aggregation steps, the two model variants share the same modeling architectures and settings as our PB-CNN model. For fair comparisons, all the hyper-parameters of the three models are optimized separately. To further assess the contribution of each crop phenological stage, we also evaluate the change of yield estimation performance of PB-CNN with the phenological stage progressing from S1 to S6. This evaluation can further shed light on the phenology-guided within-season soybean yield estimation.

In addition, the impact of training sample size on the PB-CNN performance is investigated by the designed varying historical yield scenarios, in which the amount of training samples is reduced degressively to evaluate the change of PB-CNN modeling accuracies and uncertainties. In this experiment, we degressively reduce the existing 10-year to 2-year historical data to train our PB-CNN model, simulating the cases of lacking historical yield records in practice.

5. Results

5.1. Model performance of PB-CNN

The state-level median dates of satellite-based phenological metrics as well as CPR-based soybean phenological measures throughout the study area are both shown in Fig. 7. Overall, the satellite- and corresponding CPR-based median phenological transition dates exhibit comparable variation patterns across years and states. The CPR-documented soybean growing stages can be approximated by satellite-based phenological measures (shown in Table 2, and see detailed discussions in Appendix A-5). Specifically, the satellite-based greenup/mid-greenup metrics can be connected to soybean's emerged stage. The satellite-based maturity metric has implications for the blooming stage of soybean. The timing of peak greenness metric aligns with the soybean setting pods stage. The mid-greendown metric can be connected to the dropping leaves stage of soybean. The soybean harvest stage can be approximated by the satellite-based dormancy metric. Except for the soybean emerged stage, the median transition dates of all the other phenology stages can be aligned by the corresponding satellite-derived metrics, with the mean difference less than 5 days and standard deviation less than 10 days. It indicates that the satellite-based phenology product metrics used in this study could adequately rebuild the soybean phenological development trajectory. In the constructed Phenology Imagery, the S1-S2 phenological stages correspond to soybean vegetative stages, and the S3-S6 stages connect with the soybean reproductive stages.

After we evaluate the ability of the satellite phenology product in characterizing the soybean growing stages, the model performances of PB-CNN and three advanced benchmark models (i.e., SVR, RF, and LSTM) are compared in Table 3. The results show that the models differ in predicting the soybean yields from 2014 to 2018, with the average RMSE ranging from 4.622 to 5.554 bu/ac, the average R^2 from 0.579 to 0.709, and the average bias from −2.999 to −2.057 bu/ac. In general, the developed PB-CNN model outperforms all the other models, with the lowest average RMSE (4.622 bu/ac), the highest average R^2 (0.709), and the lowest average bias (-2.057 bu/ac). It's noted that the performance of all the models declines in 2016, potentially due to abnormal climate conditions (e.g., warm winter/early spring, drought, and flooding triggered by El Niño) especially in the western and central regions of the Corn Belt (Kogan & Guo, 2016; Ortez et al., 2022). With the testing year 2018 taking into account more historical years (2008–2017) for training the models, we further compare the scatter plots of observed and predicted yields of all the models for this testing year (Fig. 8). The PB-CNN model shows better agreement between observed and predicted yields in the scatter plots. Compared to other models, fewer large-error predictions out of the two dashed lines (absolute estimation errors larger than 5.0 bu/ac) are in the PB-CNN scatter plot. The LSTM model achieves the lowest bias (-0.535 bu/ac) among all the models, but the overestimation of LSTM in some low-yield regions results in a higher RMSE and a lower R^2 for the model. We also find that all the models tend to underestimate the soybean yield in high-yield counties, which is in agreement with the previous studies (Maimaitijiang et al., 2020; Y. Wang et al., 2020). This underestimation may partly be caused by the saturation issue of optical remote sensing (Maimaitijiang et al., 2020; Y. Wang et al., 2020) and imbalanced training dataset (Fig. 9), in which there are relatively fewer high-yield training samples, leading to challenges in modeling their variations. For instance, the training dataset does not have yield records higher than 72 bu/ac (the dashed line in Fig. 9), making those high yield records in the testing dataset more difficult to estimate.

To further evaluate the performance of PB-CNN, we also map the spatial distributions of the modeling yield predictions and associated errors across counties of the study site in 2018 (Fig. 10). In general, the spatial distributions of predicted yields exhibit comparable patterns as those of observed yields, with RMSE (i.e., absolute error) for most of the counties lower than 4.0 bu/ac. The low-yield counties are mainly distributed in the northern and southern parts (e.g., North Dakota, South Dakota, Kansas, and Missouri) of the study site, while the soybean yields in the central part (e.g., Iowa, Illinois, and Indiana) are relatively high. The counties with larger RMSEs of yield predictions are mainly located in Kansas, Illinois, and North Dakota. Especially, higher errors are observed in the very high-yield counties in Illinois, partly due to the imbalanced training samples, as shown in Fig. 9. As PB-CNN achieves improved yield prediction accuracy upon comparison with the three benchmark models, we further evaluate the PB-CNN in terms of the impact of its modeling architecture, phenology design, and yield predictors using the testing year 2018 in Sections 5.2–5.4. We will then analyze the predictive uncertainty of PB-CNN for this testing year in Section 5.5.

5.2. Assessment of model architecture

We assess the model architecture of PB-CNN by comparing the yield prediction accuracies of multi-stream and single-stream architectures under various combinations of input predictors in 2018 (Table 4). When comparing the different input combinations, we find that the incorporation of meteorological and soil predictor groups tends to improve the performance of both multi-stream and single-stream models. Compared to the model with only satellite-based predictor group, the model with all predictor groups decreases the RMSE by 1.435 bu/ac, increases the R^2 by 0.157, and reduces the bias by 2.196 bu/ac for the multi-stream architecture. As for the single-stream architecture, the model with all predictor groups also tends to reduce the prediction error (RMSE decreased by 0.419 bu/ac and bias decreased by 0.767 bu/ac), yet with a

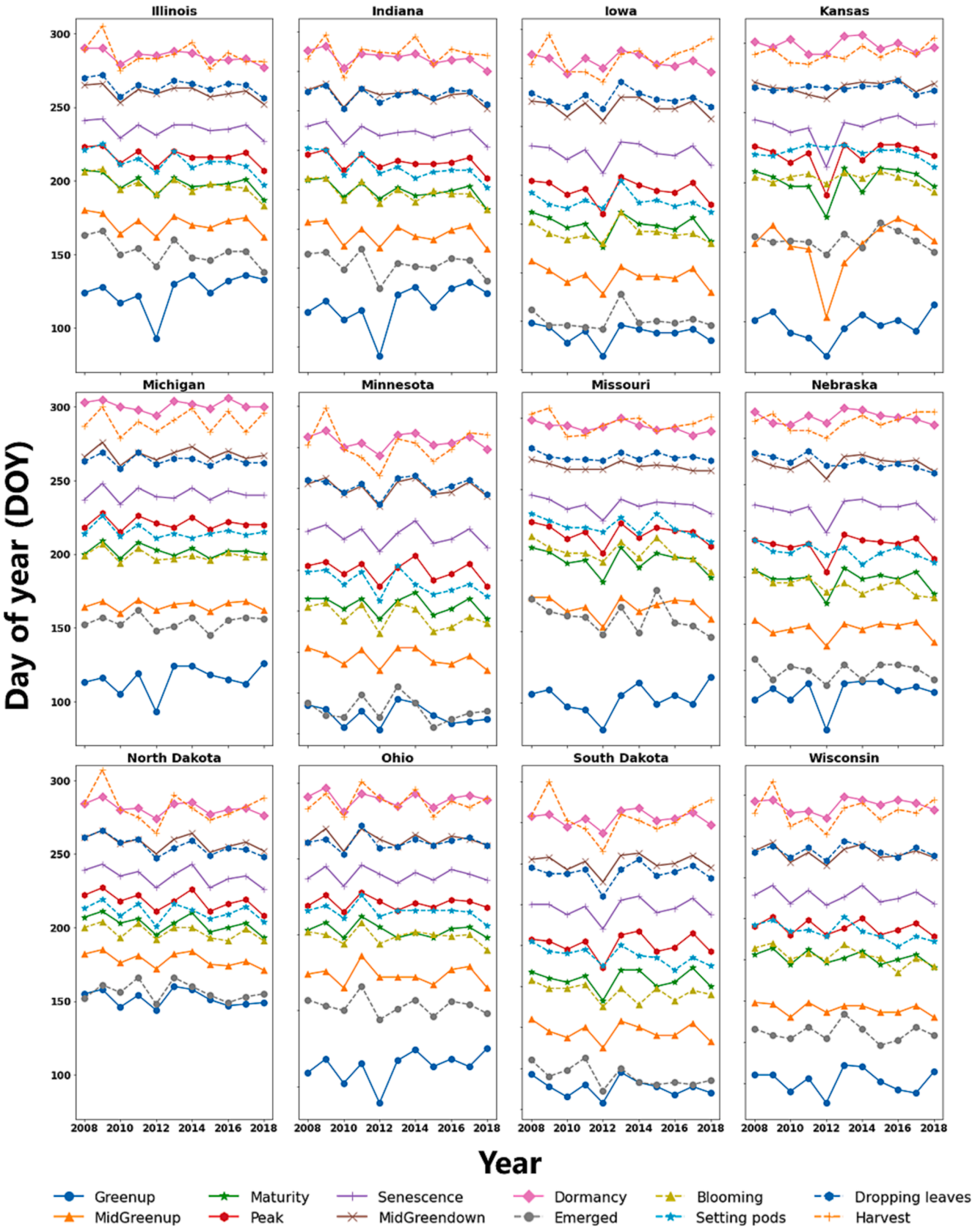

Fig. 7. Inter-annual median transition date comparisons between the satellite- and CPR-based soybean phenological measures across the states in the Corn Belt.

Table 2
The mean and standard deviation of the median dates of satellite- and CPR-based soybean phenological measures in the study area from 2008 to 2018.

CPR-based transition date	Mean	Standard deviation	Satellite-based transition date	Mean	Standard deviation
Emerged	154	7	Mid-Greenup/Greenup	171/129	10/20
Blooming	198	6	Maturity	201	6
Setting pods	215	7	Peak	219	6
Dropping leaves	262	6	Mid-Greendown	262	5
Harvest	286	9	Dormancy	288	8

Table 3
The accuracy of PB-CNN, LSTM, SVR, and RF in estimating the soybean yields for the testing years 2014–2018.

Accuracy Metric	Year	Model			
		LSTM	SVR	RF	PB-CNN
RMSE (bu/ac)	2014	5.027	4.379	4.499	**3.981**
	2015	7.012	5.982	5.263	**5.140**
	2016	5.455	6.127	5.929	**5.076**
	2017	5.444	4.906	4.861	**4.562**
	2018	4.831	4.721	4.777	**4.35**
	Average	5.554	5.223	5.066	**4.622**
R^2	2014	0.707	0.779	0.767	**0.817**
	2015	0.396	0.552	0.653	**0.669**
	2016	0.451	0.345	0.387	**0.551**
	2017	0.693	0.727	0.732	**0.763**
	2018	0.646	0.713	0.667	**0.743**
	Average	0.579	0.623	0.641	**0.709**
Bias (bu/ac)	2014	−2.438	−1.842	−1.837	**−1.072**
	2015	−5.399	−4.293	−3.310	**−3.242**
	2016	−3.784	−5.007	−4.780	**−3.767**
	2017	−2.292	−2.483	−2.198	**−1.275**
	2018	**−0.54**	−1.372	−1.641	−0.928
	Average	−2.891	−2.999	−2.753	**−2.057**

decreased R^2 value. The incorporation of the soil predictor group into the models with satellite and meteorological predictor groups can further improve the model performance for the multi-stream architecture, yet not much for the single-stream architecture. When more than two yield predictor groups are considered, the models with multi-stream architectures consistently outperform those of single-stream ones. Especially when all the predictor groups are considered as the input, the multi-stream model significantly improves the yield prediction accuracy with RMSE decreased by 1.02 bu/ac, R^2 increased by 0.19, and bias decreased by 1.43 bu/ac. Unlike the shared network structures for all the predictor groups in the single-stream models, the sub-networks are employed and optimized for different predictor groups in the multi-stream models. The PB-CNN with the multi-stream architecture thus takes into account the heterogeneous characteristics of different predictor groups, achieving an RMSE of 4.350 bu/ac, an R^2 of 0.743, and a bias of −0.928 bu/ac when all the predictor groups are considered.

5.3. Evaluation of Phenology design

The performances of the PB-CNN models under three temporal aggregations (i.e., phenophase, 8-day, and monthly aggregations) are shown in Table 5. Among the temporal aggregations, the PB-CNN with the phenophase design achieves the highest prediction accuracy, with the RMSE being 4.350 bu/ac, R^2 being 0.743, and bias being −0.928 bu/ac, while the model with the 8-day calendar interval suffers performance degradation (RMSE is 5.535 bu/ac, R^2 is 0.522, and bias is −2.540 bu/ac). Compared to the 8-day temporal aggregation, the model with the monthly calendar interval achieves slightly higher accuracy, with RMSE decreased by 0.569 bu/ac, R^2 increased by 0.067, and bias decreased by 0.66 bu/ac. Despite the more subtle temporal information provided by the 8-day interval setting, the temporal aggregations of smaller calendar intervals may not yield better performance, partly due to the increasing inconsistency in crop phenology representation over space and time with smaller intervals. With crop phenological dynamics explicitly considered in the phenophase aggregation design, the proposed PB-CNN model accounts for phenological stage-specific crop growth and environmental conditions and outperforms the other two calendar-interval based models in yield predictions. Though the PB-CNN model of phenophase design maintains similar temporal aggregation steps as the model of monthly design, it achieves higher yield estimation accuracy with RMSE reduced from 4.966 to 4.350 bu/ac, R^2 increased from 0.589 to 0.743, and bias decreased from −1.880 to −0.928 bu/ac. The comparisons indicate that the PB-CNN of phenophase design could facilitate the modeling of the distinct yield response characteristics across phenological stages to improve the estimation accuracy.

As the soybean yield response to crop growth and environmental conditions varies across phenological stages, we further evaluate yield estimation performance across soybean phenological stages throughout the season in 2018 (Fig. 11). Overall, the PB-CNN model shows improved performance (i.e., RMSE, R^2, and bias) with more phenological stages incorporated, resulting in the highest accuracy achieved with the consideration of all phenological stages. Yet the modeling performance improvements vary across the stages, with a relatively large accuracy increase in S3 and S4, when soybean experiences the critical transition from vegetative to reproductive stages (Table 2). Specifically, the incorporation of the S3 phenological stage decreases the RMSE by 0.73 bu/ac, increases the R^2 by 0.15, and reduces the bias by 1.01 bu/ac. On the other hand, the change of modeling accuracy with the incorporation of S5 and S6 becomes smaller, as those stages are related to the soybean dropping leaves and harvest stages (Table 2).

5.4. Feature importance

We further utilize the permutation importance method to analyze the importance of each yield predictor group per phenological stage in estimating the soybean yield (Fig. 12). Overall, the satellite-based EVI2 predictor group has high contributions to the soybean yield estimation. The EVI2 predictors during S2 to S5 are among the most important predictors due to the embedded critical crop growth information that is directly indicative of soybean yield. For the meteorological predictors, the water- and heat-related predictor groups show relatively comparable patterns in predictor importance across phenological stages. Both water- and heat-related predictors within phenological stages 3 to 5 are more important for yield prediction, possibly due to the more significant impacts of environmental stress on soybean growth and yield formation during those stages (Jumrani & Bhatia, 2018). The soil predictor group is less important, with its values being unchanged across phenological stages and years. Among the phenological stages, the S3-S5 stages are the most crucial ones, corresponding to soybean blooming, setting pods, and turning yellow stages (Table 2). The S1 and S6 stages, representing the soybean emergence and harvest stages, are less critical in predicting the soybean yield.

5.5. Predictive uncertainty analysis

Besides the predicted yields, we also analyze the predictive

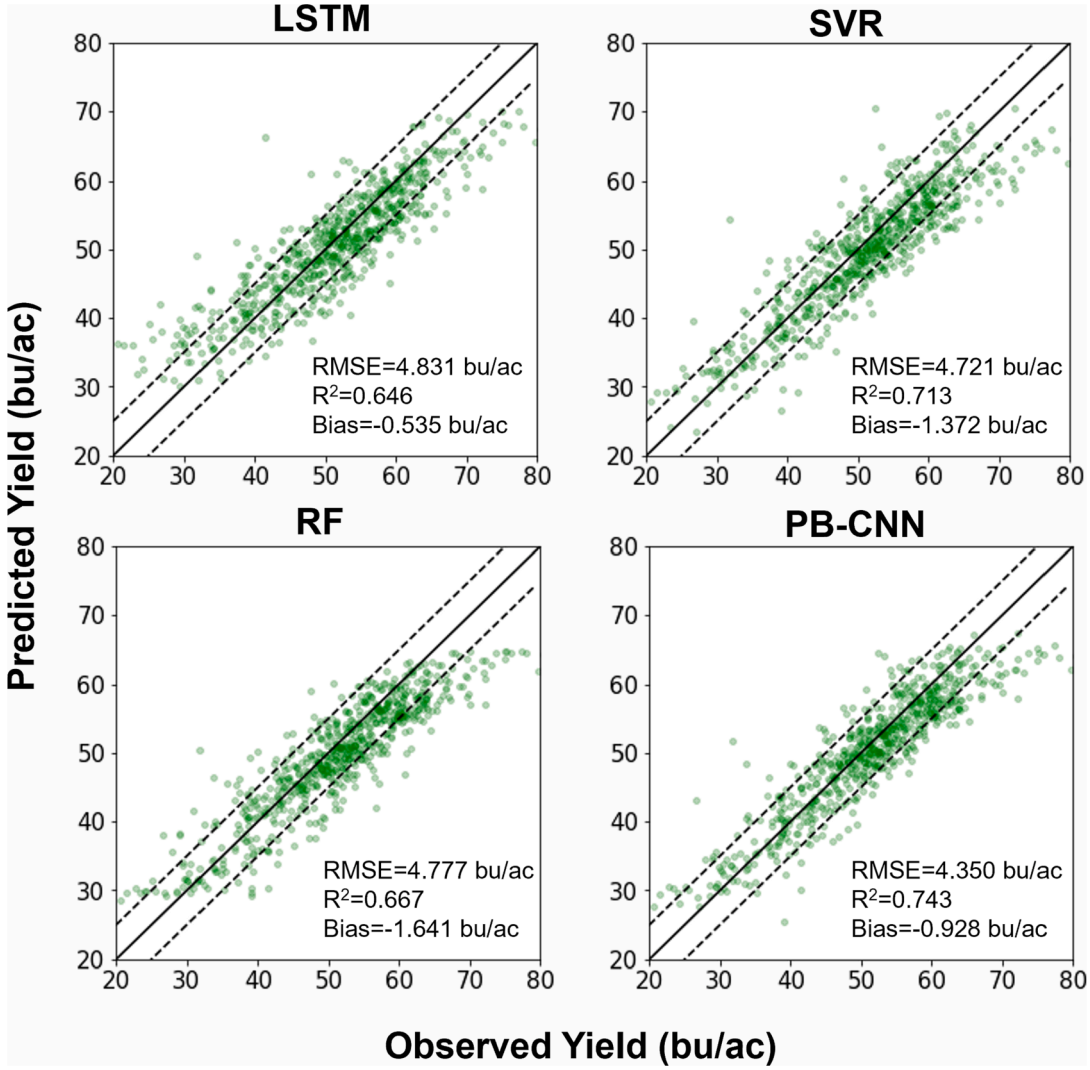

Fig. 8. The accuracy of PB-CNN, along with three benchmark models (LSTM, SVR, and RF), in estimating the soybean yields in 2018. Two oblique dashed lines indicate the absolute estimation errors being 5.0 bu/ac.

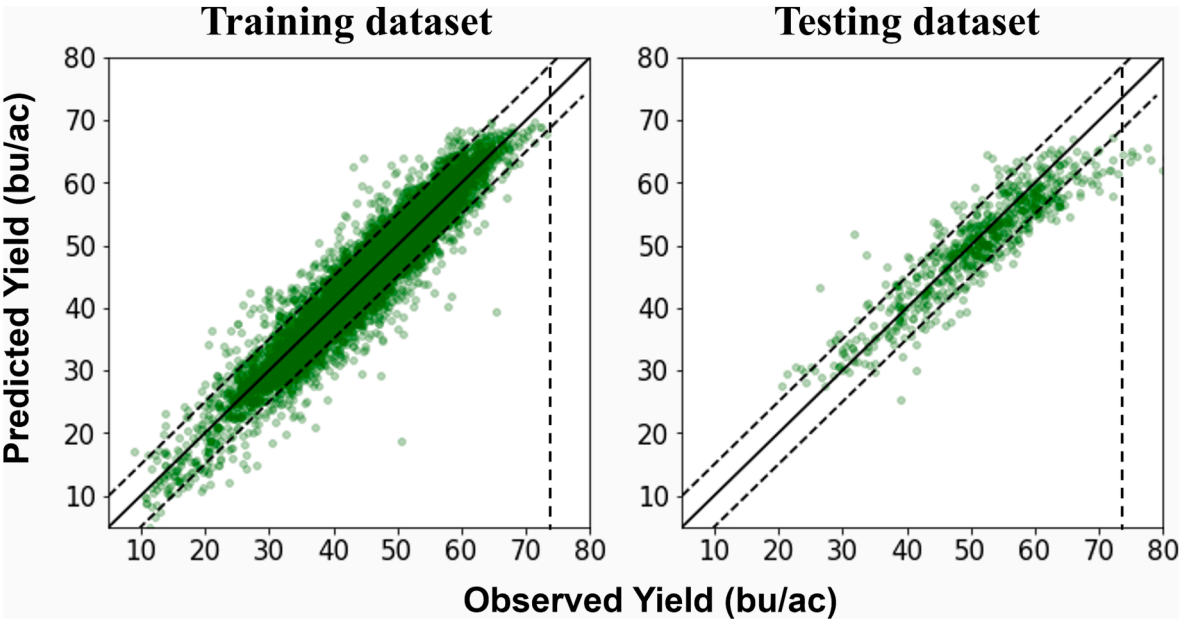

Fig. 9. Scatter plots of observed versus predicted yields for the training (2008–2017) and testing (2018) datasets using PB-CNN. Two oblique dashed lines indicate the absolute estimation errors being 5.0 bu/ac. The vertical dashed line represents the observed yield being 72 bu/ac.

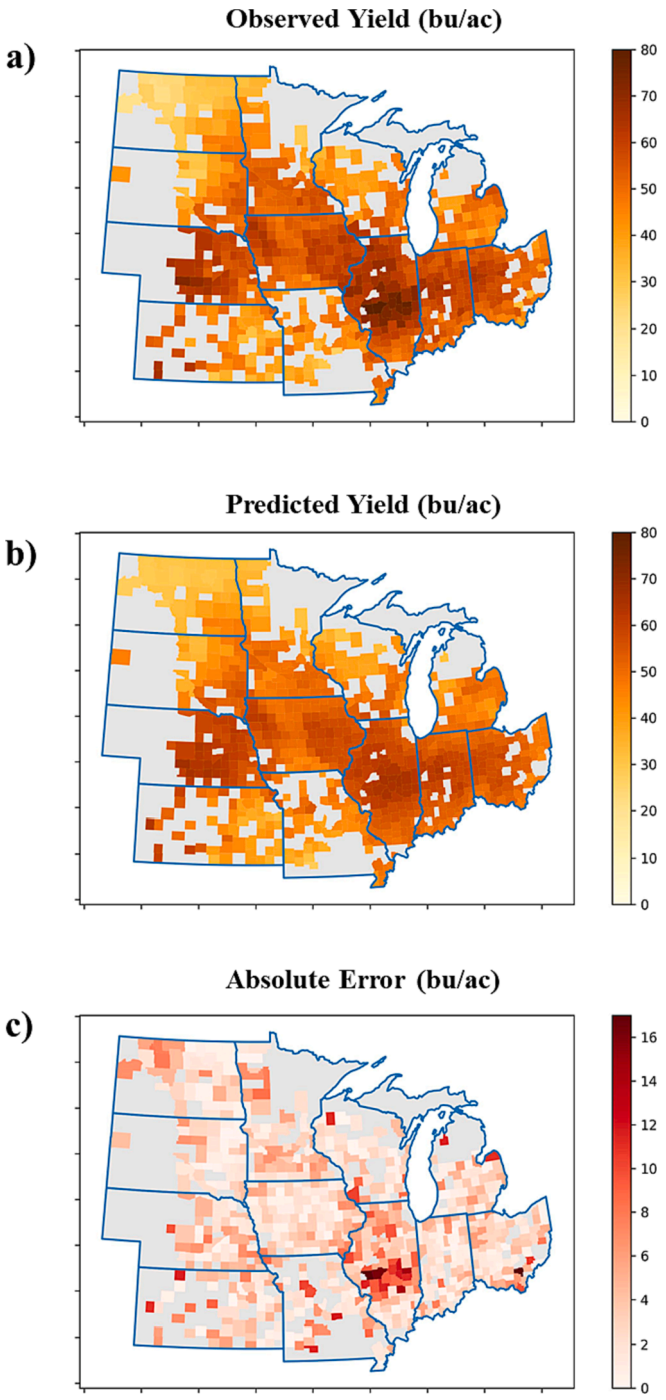

Fig. 10. Spatial distributions of (a) observed yield, (b) predicted yield, and (c) RMSE (i.e., absolute error) using PB-CNN across counties of the study site for the year 2018. Areas in light gray indicate no data available.

uncertainty derived from the PB-CNN framework. In 2018, the P-factor of PB-CNN is 96.7 %, indicating that more than 96 % of the county-level observed yields are enveloped within the 95 % confidence interval given by our model. This close to 100 % P-factor value empirically demonstrates the good performance of PB-CNN in estimating the yield distribution and predictive uncertainty. Additionally, the total predictive uncertainty map over the study area is shown in Fig. 13. The spatial distribution of the predictive uncertainty exhibits a relatively homogenous pattern compared to that of RMSE (Fig. 10). The predicted yields in most counties have relatively small uncertainty (less than 5.0 bu/ac). Relatively larger yield uncertainty regions are mainly in the western (e.g., South Dakota and Kansas) and southern (e.g., Missouri) parts of the study area, where the climate and soil characteristics are significantly different from other regions in the Corn Belt (Challinor et al., 2014). Some low-yield regions, such as the northern part of Wisconsin, also have relatively larger total predictive uncertainty, which may be caused by the lack of similar samples in the training dataset.

We further partition the total predictive uncertainty into aleatoric uncertainty and epistemic uncertainty (Fig. 14). Overall, the ranges of the two uncertainties show significant differences. For most counties, the epistemic uncertainty is typically smaller than 3.0 bu/ac. Yet the aleatoric uncertainty is larger than 3.5 bu/ac, indicating that the aleatoric uncertainty is the dominant uncertainty of PB-CNN. Meanwhile, similar spatial patterns between the aleatoric and epistemic uncertainties are observed in most of our study areas. For the counties in

Table 4
Comparisons of PB-CNN modeling performance with multi-stream and single-stream architectures under various combinations of input predictors for soybean yield estimation in 2018.

Predictors	Single-Stream			Multi-Stream		
	RMSE (bu/ac)	R^2	Bias (bu/ac)	RMSE (bu/ac)	R^2	Bias (bu/ac)
EVI2	5.785	0.586	−3.124	5.785	0.586	−3.124
EVI2&MET	5.419	0.479	−1.728	5.056	0.548	−1.216
ALL	5.366	0.554	−2.357	4.350	0.743	−0.928

Table 5
Comparisons of PB-CNN modeling performance under three temporal aggregations for soybean yield estimation in 2018.

Temporal aggregation	Phenology	8-day	Monthly
RMSE (bu/ac)	4.350	5.535	4.966
R^2	0.743	0.522	0.589
Bias (bu/ac)	−0.928	−2.540	−1.880

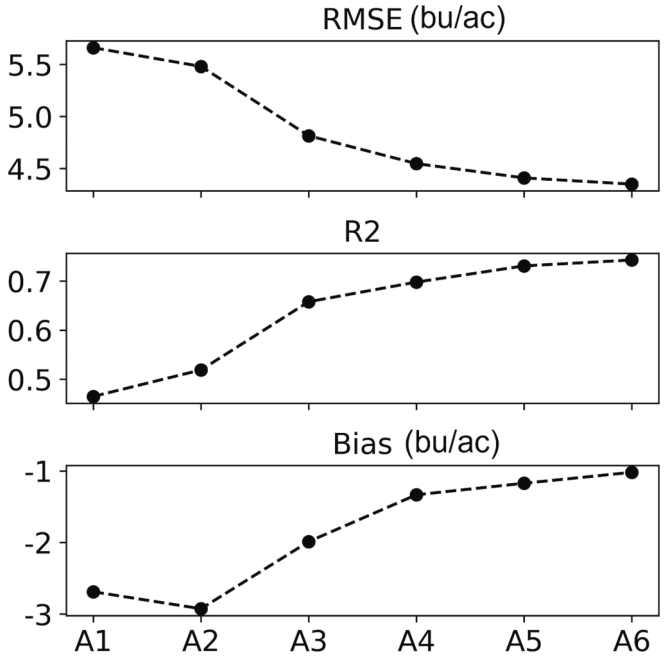

Fig. 11. The performance of PB-CNN in soybean yield estimation across phenological stages throughout the season in 2018. A1 denotes only the first phenological stage being considered, while A6 denotes all the phenological stages being incorporated.

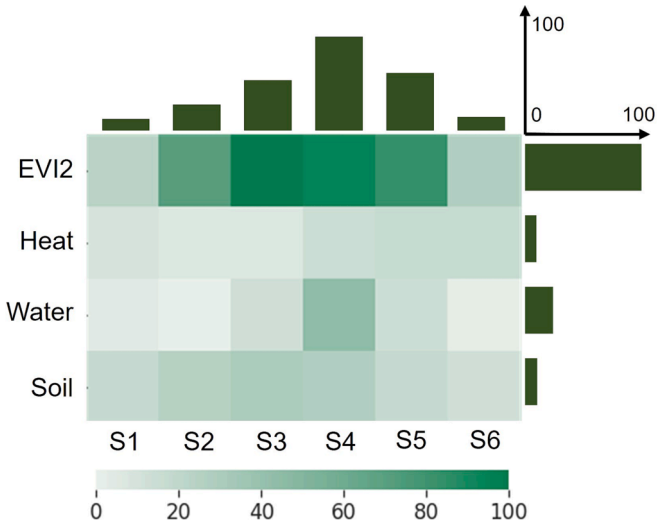

Fig. 12. The permutation importance of yield predictor groups across phenological stages on estimating the soybean yield. The heatmap shows the importance of each predictor group within each phenological stage. The importance of each predictor group and each phenological stage is shown above and to the right of the heatmap, respectively. S1-S6 represent the phenological stages 1–6.

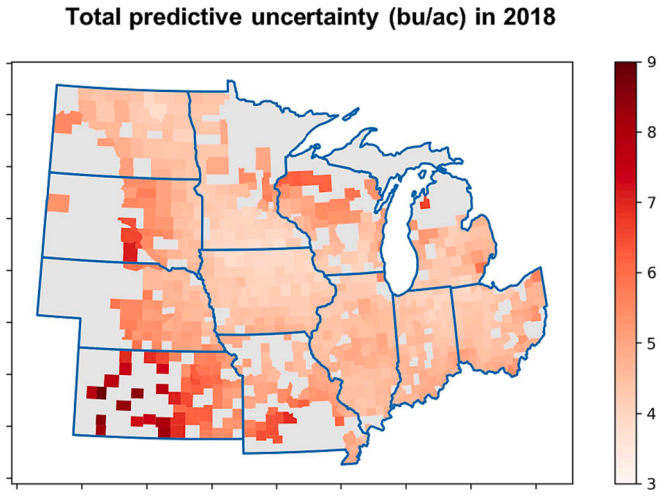

Fig. 13. The spatial distribution of the total predictive uncertainty across counties of the study site in 2018. Areas in light gray indicate no data available.

the southwestern part (e.g., Nebraska, Missouri, South Dakota, and Kansas), both uncertainties are larger than those in other regions. This could potentially be caused by the different environmental conditions in those states, which are mostly located in the Great Plains with more severe heat and water stress during the summer (Lobell et al., 2014; Ma et al., 2021; Wright & Wimberly, 2013). As for the central part of the study area, the distribution of the aleatoric uncertainty is relatively homogeneous, with the average being about 3.5 bu/ac. The epistemic uncertainty also shows similar spatial patterns in most areas but exhibits more local heterogeneities. For instance, some high epistemic uncertainty estimates of PB-CNN are in the central part of Illinois. Those samples in the testing dataset have way higher soybean yields and are likely different from those in the training dataset (Fig. 9), which could cause uncertainty for model parameters.

To further analyze the spatial distributions of uncertainties, we summarize the uncertainties and some representative predictors at the state level in Fig. 15. The state-level results show that the yield estimates in some states (e.g., Nebraska, Missouri, South Dakota, Kansas, and Wisconsin) exhibit relatively higher aleatoric and epistemic uncertainties, consistent with previous spatial pattern analyses. Among these states, the yield estimates in Kansas and Missouri also exhibit relatively large RMSE and relative RMSE (i.e., RMSE% = RMSE/observed yield) due to different climate and soil conditions in the region. The P-factors are higher than 95 % in all the states except Illinois, which is partly due to the imbalanced training samples. According to the summarized mean and standard deviation of representative yield predictors (i.e., maximum temperature, PET, and soil organic carbon content), the states with smaller uncertainties are mostly distributed in temperate regions with relatively benign environmental conditions, while the states with higher uncertainties are mainly located in areas more subject to environmental stress (Lobell et al., 2014), such as Nebraska, South Dakota, and Kansas. Heat stress (i.e., high maximum temperature) and water stress (i.e., high PET) are more likely to occur in those states with high aleatoric uncertainties (Goparaju & Ahmad, 2019; Lobell et al., 2014). The soil organic carbon content of those states also tends to be relatively low with high variations.

As aleatoric uncertainty is the main source of uncertainty and it stems from input observations, we further explore the contribution of the inputs of each yield predictor group as well as each phenological stage to the aleatoric uncertainty change using the permutation importance method (Fig. 16). Regarding the phenological stages, the aleatoric uncertainty of PB-CNN is mainly from the first two stages, corresponding to the emerged and vegetative growth stages of soybean. Because the EVI2 and other environmental predictors in the early growth stages are less associated with the final soybean yield, the predictors in those stages may bring more uncertainty into the yield estimation. As for the yield predictor groups, the inputs of the soil predictor group are the main source of the aleatoric uncertainty. Since the predictors in the soil group

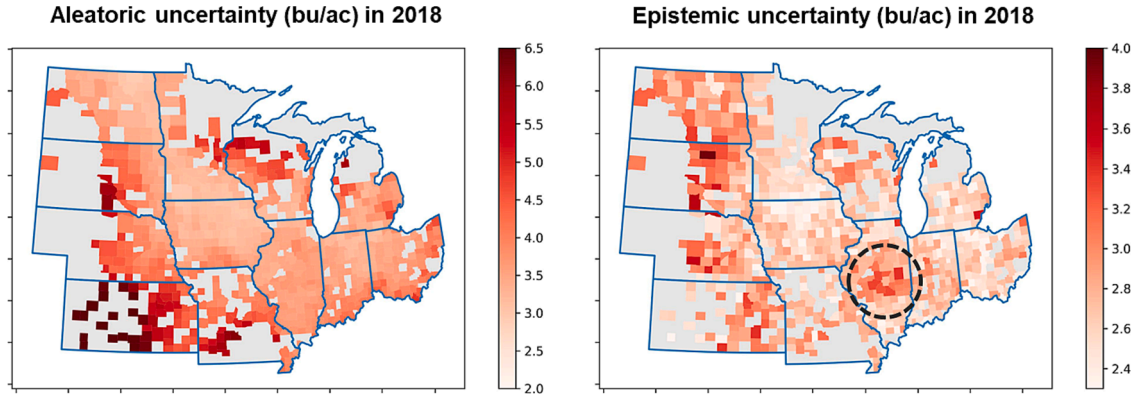

Fig. 14. The spatial distributions of aleatoric uncertainty and epistemic uncertainty across counties of the study site in 2018. The circled area represents the counties with high epistemic uncertainty in the central part of Illinois. Areas in light gray indicate no data available.

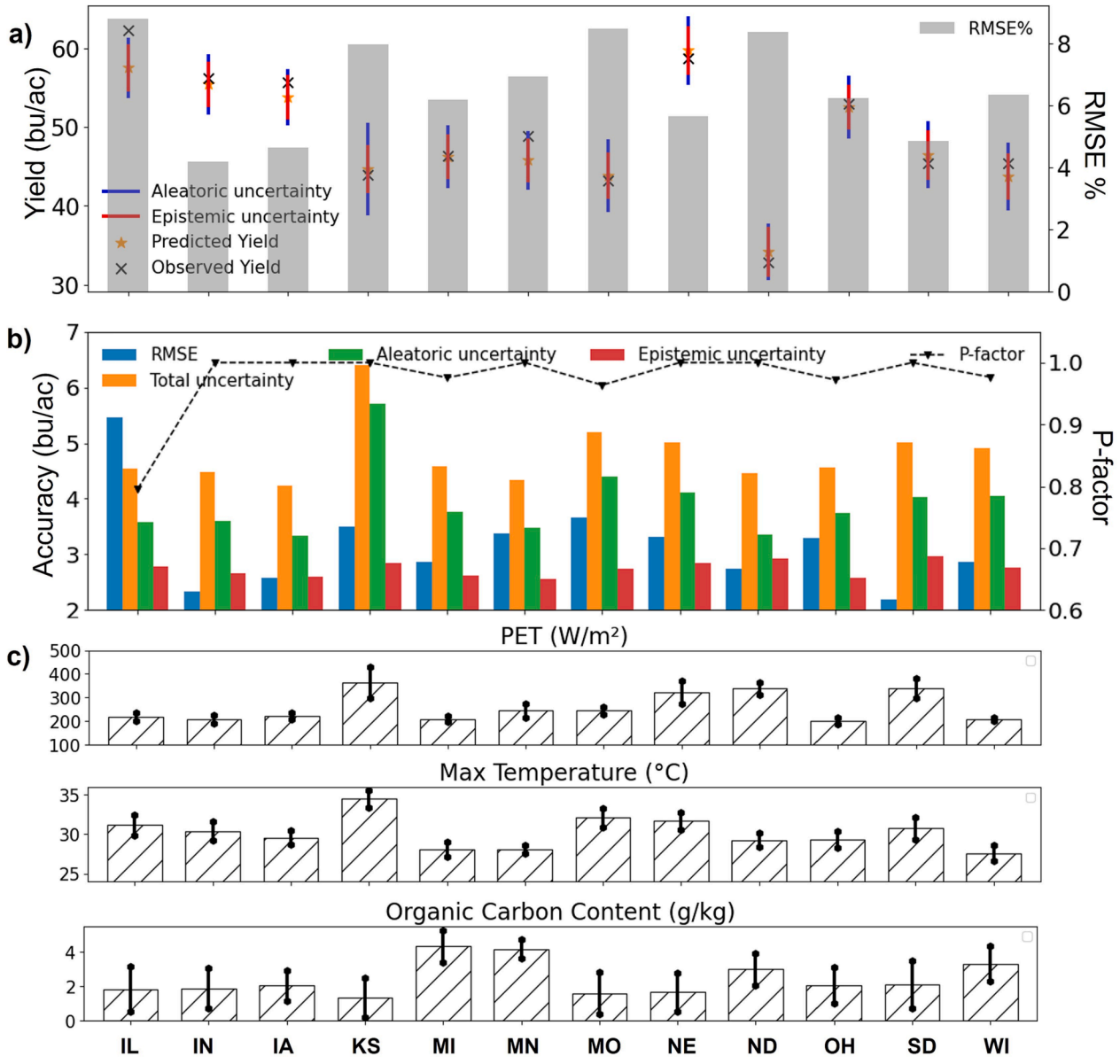

Fig. 15. Distributions of predictive yield, uncertainties and environmental predictors across the states of the study site in 2018. a) The mean observed yield, predicted yield, associated uncertainties, and RMSE% for each state; b) The mean RMSE, uncertainties, and P-factor for each state; c) The mean values of PET, maximum temperature, and soil organic carbon content for each state. The error bars represent the standard deviation.

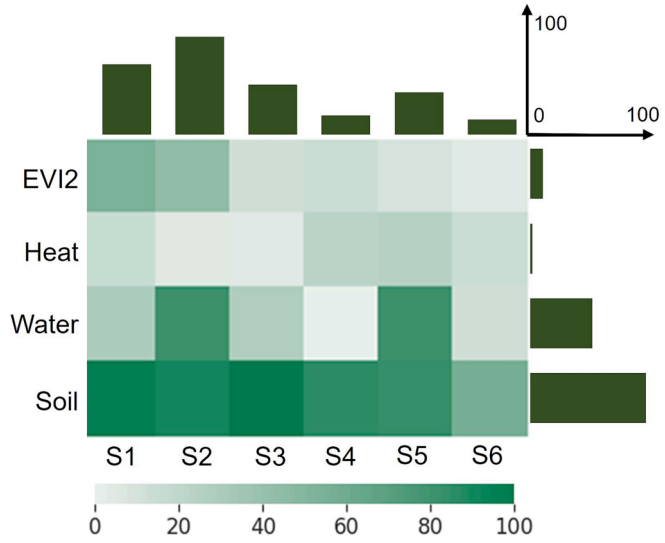

Fig. 16. The permutation importance of yield predictor groups across phenological stages to the change of aleatoric uncertainty. Heatmap shows the importance of each predictor group within each phenological stage. The importance of each predictor group and each phenological stage is shown above and to the right of the heatmap, respectively.

remain unchanged across phenological stages as well as from year to year for each county, the one-off soil data with coarse temporal resolution may bring more uncertainty into the yield estimation. The inputs of the water-related predictor group also exert a relatively large influence on the aleatoric uncertainty change, partly due to the high variations and fluctuations of precipitation across counties of the study site, especially in S2 and S5 phenological stages.

5.6. Evaluation of PB-CNN under varying historical yield scenarios

To investigate the model robustness under various training sample sizes, we further degressively reduce the training samples of historical yield records from 10-year to 2-year and assess the model performance in the testing year 2018. The model performance and estimation uncertainties for varying historical yield scenarios are shown in Table 6 and Fig. 17. Generally, aleatoric uncertainty is still the main source of uncertainty, and larger model errors and uncertainties are observed as the training sample size decreases. The RMSE gradually increases from the lowest 4.350 bu/ac with 10-year training data to the highest 5.731 bu/ac with 2-year training data (Table 6). Despite the increase of RMSE, the P-factor is still consistently higher than 92 % for all the training sample scenarios (i.e., training sample size from 8202 to 1591), indicating that more than 92 % of the observed yields are always within the 95 % confidence interval estimated by PB-CNN. This robust performance benefits from the model's accurate prediction of the estimation uncertainties. As the training sample size decreases, the uncertainties

Table 6
The PB-CNN model accuracy (bu/ac) and uncertainties (bu/ac) for different training samples.

Training Years	Number of training samples	RMSE	Total uncertainty	Aleatoric uncertainty	Epistemic uncertainty	P-factor
2008–2017	8202	4.350	4.749	3.869	2.715	0.95
2009–2017	7383	4.587	4.318	3.646	2.28	0.94
2010–2017	6552	4.764	4.291	3.615	2.264	0.92
2011–2017	5665	4.981	5.104	4.165	2.918	0.96
2012–2017	4806	4.539	4.867	3.865	2.937	0.96
2013–2017	3952	4.671	5.474	4.366	3.28	0.97
2014–2017	3155	4.869	5.504	4.451	3.222	0.97
2015–2017	2341	5.747	5.971	4.869	3.427	0.96
2016–2017	1591	5.731	6.119	5.036	3.443	0.96

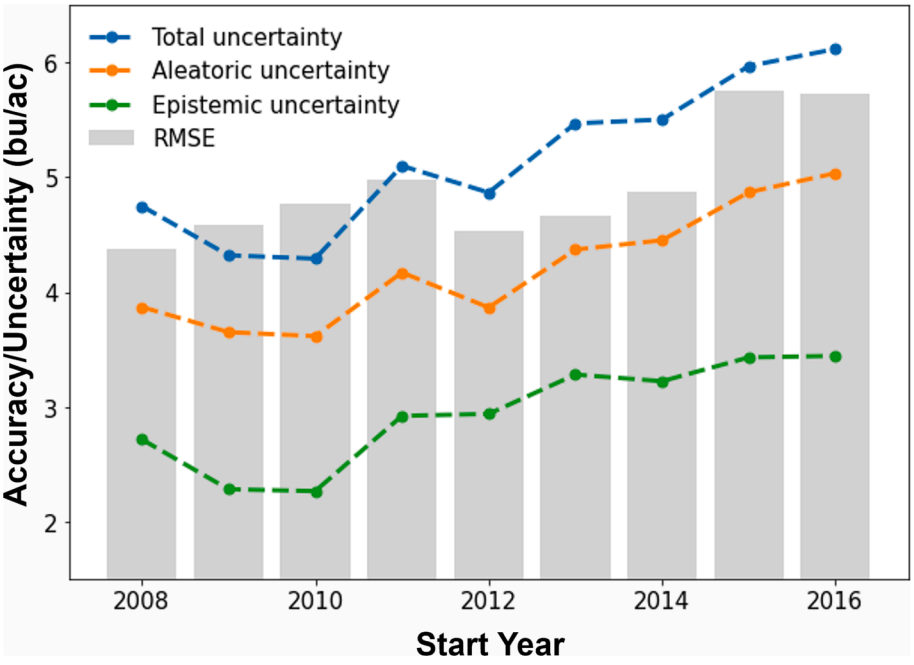

Fig. 17. The change of PB-CNN accuracy and uncertainties with the training sample start year changing from 2008 to 2016.

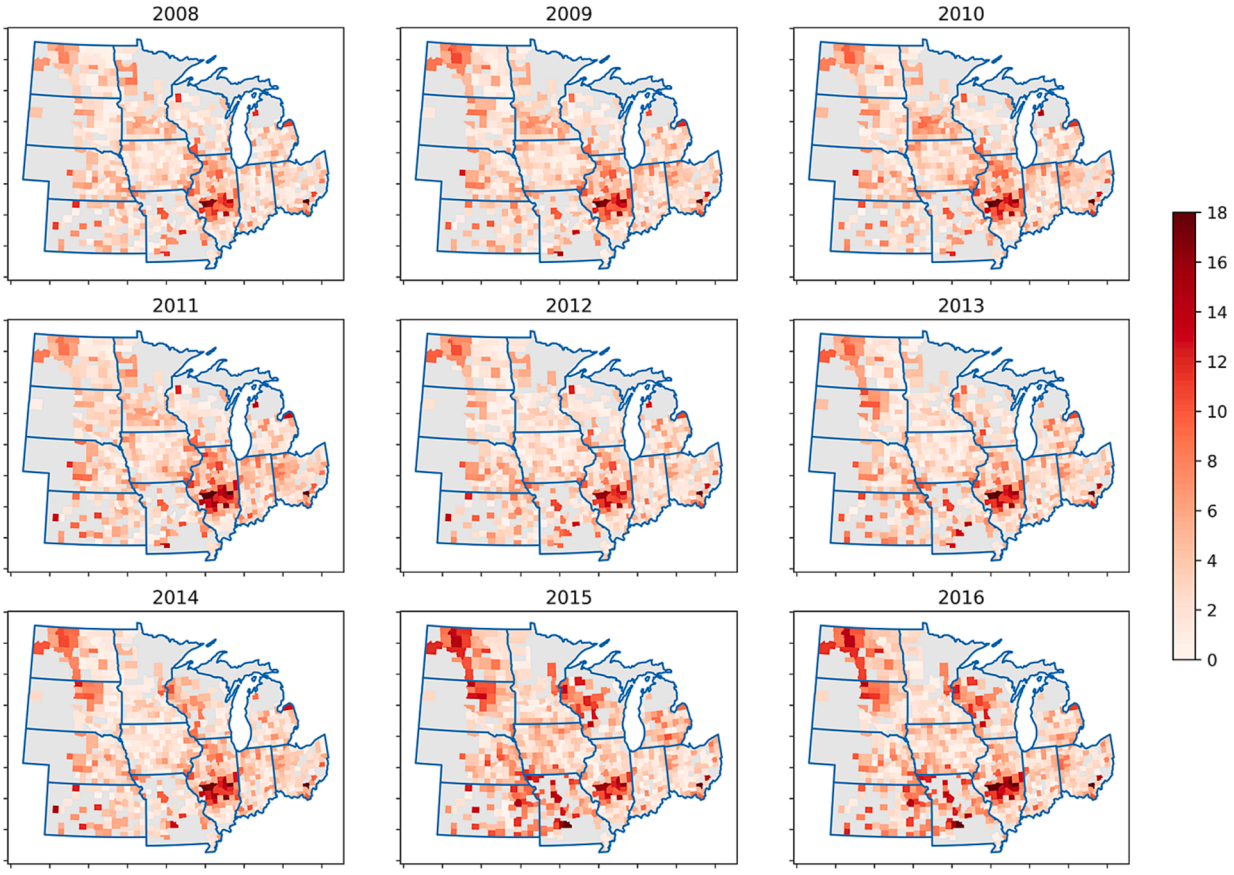

Fig. 18. The evolution of the spatial distribution of RMSE (bu/ac) for the testing year 2018 with the training sample start year changing from 2008 to 2016. Areas in light gray indicate no data available.

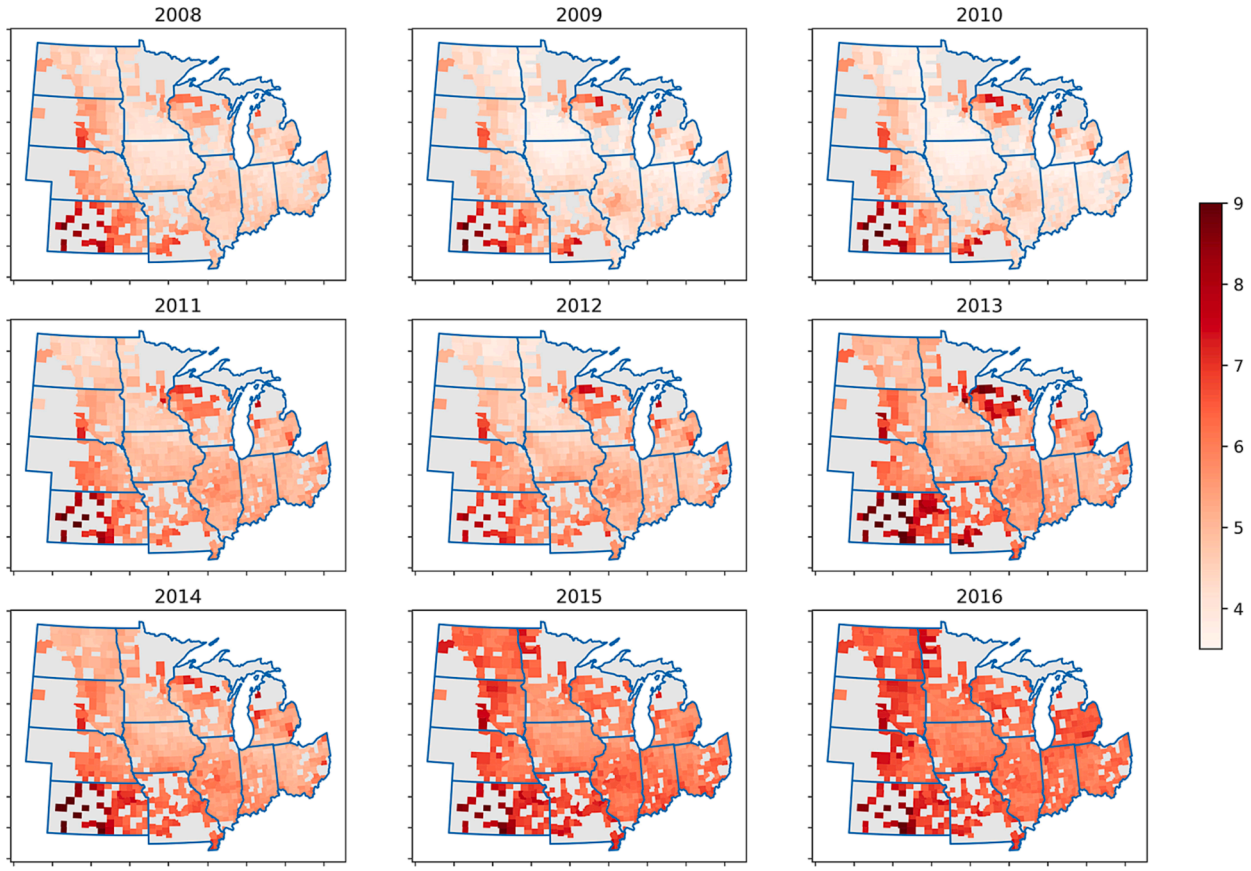

Fig. 19. The evolution of the spatial distribution of total predictive uncertainty (bu/ac) for the testing year 2018 with the training sample start year changing from 2008 to 2016. Areas in light gray indicate no data available.

estimated by PB-CNN become larger for potential risk warning: the total predictive uncertainty increases from 4.749 to 6.119 bu/ac; the aleatoric uncertainty increases from 3.869 to 5.036 bu/ac; the epistemic uncertainty increases from 2.715 to 3.443 bu/ac. The fluctuations in RMSE and total predictive uncertainty (e.g., increased RMSE and uncertainty with the start year 2011) may be due to changing variability in meteorological and management conditions when the start year changes. Also the evolutions of spatial distributions of RMSE and the total uncertainty of yield estimation with the change of training sample sizes are shown in Figs. 18 and 19, respectively. Similar to Fig. 17, as the size of historical yield records decreases, larger model errors and uncertainties are observed across the study area. Yet the spatial patterns of model errors and uncertainties remain relatively consistent with the change in the training sample size. The counties with relatively large yield estimation errors are usually located in Kansas, Illinois, and North Dakota. The counties with relatively large total uncertainty are mainly concentrated in the northern part of Michigan as well as the southwestern part (e.g., Nebraska, Missouri, South Dakota, and Kansas) of the study site.

6. Discussion

In this study, we develop an innovative PB-CNN framework for county-level soybean yield estimation and uncertainty quantification. The PB-CNN framework encompasses three key components, namely Phenology Imagery construction, multi-stream Bayesian-CNN modeling, as well as feature and uncertainty analysis. By integrating the phenological, deep learning, and Bayesian modeling designs, the PB-CNN framework can largely advance the modeling of crop yield responses to crop growth and environmental conditions across critical soybean phenological growth stages. With the multi-stream network modeling architecture, it also facilitates the accommodation of heterogenous yield predictors of various characteristics and resolutions. The PB-CNN framework can further advance the analysis of the importance of varying phenological stages and yield predictors in soybean yield estimation, as well as the modeling of associated uncertainties.

Under the devised framework, we innovatively construct the Phenology Imagery for each yield predictor via spatial and temporal aggregations to characterize the predictor county-level variations across soybean growth stages. This novel construction of Phenology Imagery takes into account the spatial and temporal heterogeneity of crop growth progress at the pixel level, as well as enabling the modeling of diverse responses of crop yield to weather and environmental stress of varying phenological stages. The phenophase-based design thus considerably improves the model performance in estimating the soybean yield compared to two calendar-interval based designs (i.e., 8-day and monthly temporal aggregations). Across the phenological stages and yield predictor groups, the satellite-based EVI2 predictor group is the most important in soybean yield estimation for most phenological stages. The importance of EVI2 is due to its direct characterization of crop growth vigor and greenness status to infer the plant biomass as well as crop yield. The importance of vegetation index in crop yield estimation has also been suggested by previous studies (Dadsetan et al., 2020; Raun et al., 2002; Yao et al., 2012). At different phenological stages, the final crop yield responses to changes in meteorological conditions may vary largely. The phenological stages 3–5, corresponding to soybean blooming to dropping leaves phenological stages, exert the most critical roles in soybean yield estimation (Diao, 2020). It indicates that the final soybean yield may heavily be affected by the meteorological conditions (i.e., heat stress and water stress) during this period. The extreme weather events within these stages, particularly the water stress during the setting pod stage, would potentially cause significant soybean yield reduction. The construction of Phenology Imagery largely facilitates the

phenology-based predictor importance analytics for crop yield estimation.

To accommodate the heterogeneity of yield predictors, the multi-stream deep learning architecture is devised in PB-CNN. This multi-stream architecture enables the characterization of the soil–plant-atmospheric continuum using a diverse set of yield predictors, with each sub-network optimized to account for the heterogeneous nature of those predictors. It facilitates the flexible design of customized sub-network structures to learn unique feature representations for each predictor group, and to comprehensively model the yield response to diverse crop growth and environmental conditions of varying characteristics. By comparison, the shared modeling architecture for all yield predictors in the single-stream CNN can hardly accommodate the predictor heterogeneity. Therefore, PB-CNN consistently outperforms conventional single-stream CNN models, especially when predictors with distinct characteristics and resolutions (e.g., daily meteorological predictors vs one-off soil predictors) are employed as the inputs. Also, given the independence of each stream, more highly customizable sub-network structures (e.g., other deep learning structures, including ANN and LSTM) can be integrated into the multi-stream architecture to enhance the modeling flexibility and can further be investigated in future studies.

With Bayesian modeling and specially designed modeling outputs, our PB-CNN model can not only provide yield estimation but also quantify predictive uncertainties for assessing the estimation reliability. Taking the variations from both input observations and model parameters into account, PB-CNN can simultaneously estimate the aleatoric and epistemic uncertainties. The aleatoric uncertainty stems from inherent randomness and variations in modeling inputs (e.g., sensor noise and fluctuations of meteorological predictors), while the epistemic uncertainty arises from incomplete knowledge of yield-predictor empirical relationships possibly caused by the lack of training data. In our study, aleatoric uncertainty is the dominant yield estimation uncertainty source, particularly in the southwest of the study site (e.g., Kansas and Missouri) (Wright & Wimberly, 2013). According to the novel uncertainty permutation analysis, soil predictors are the major contributors to aleatoric uncertainty among all the yield predictors. Although the incorporation of soil predictors improves the yield estimation accuracy, the unchanged nature of soil predictors across phenological stages and years per county may also bring more uncertainties into the modeling estimation. Among the phenological stages, the yield predictor inputs during the soybean early vegetative stage tend to bring more uncertainties into the yield estimation, partly due to the phenological characterization noise from the MODIS Land Cover Dynamics product, as well as the weaker association between the yield predictors in the early growth stage and the final soybean yield. The relatively large aleatoric uncertainties in the southwest of the study site (e.g., the Great Plain) may be caused by the fluctuations in environmental conditions of the region, such as low precipitation and humidity, sudden temperature changes, as well as varying irrigation applications (Grassini et al., 2015; Kukal & Irmak, 2018; Wright & Wimberly, 2013). These unique environmental settings also increase the epistemic uncertainty in the region of Great Plains as the samples in this region tend to be underrepresented in the training dataset. Relatively large epistemic uncertainty also appears in the high-yield counties of central Illinois, as the soybean yield of this area in the testing year (2018) tends to be higher than the historical yields of the study site in the training years (2008 to 2017). Our varying historical yield scenarios further indicate that both the aleatoric and epistemic uncertainties tend to increase with more limited historical yield records. The decrease in the number of samples in the training dataset makes it more difficult to learn appropriate modeling structures for estimating the soybean yield under the testing dataset conditions. The more limited historical datasets may also make the model less robust to the noise and variations in the yield input observations, resulting in larger data uncertainties. The development of PB-CNN opens up unique opportunities to investigate the predictive uncertainty in crop yield estimation and to enhance our understanding of the complex mechanisms underlying crop growth and yield in response to varying environmental predictors.

Owing to innovative Phenology Imagery design, multi-stream deep learning architecture, and Bayesian modeling, the proposed PB-CNN framework outperforms three advanced benchmark models (i.e., SVR, RF, and LSTM), which have been widely used in crop yield estimation studies (Ashapure et al., 2020; Jiang et al., 2020b; Kang et al., 2020). Compared to these benchmark models (an average RMSE of 5.281 bu/ac), PB-CNN achieves superior performance in county-level soybean yield estimation of the US Corn Belt, with an average RMSE of 4.622 bu/ac for the testing years 2014–2018. This improved performance is also evident in the context of relevant previous studies, which typically report an RMSE range of 5.0 to 7.0 bu/ac for county-level soybean yield estimation in US (Johnson, 2014; Russello, 2018; You et al., 2017). It is partly due to the strength of PB-CNN in accommodating spatio-temporal crop phenological patterns, tackling the heterogeneity of yield predictors, and modeling complex yield-predictor relationships. The integration of Bayesian modeling further enables the evaluation of both aleatoric and epistemic uncertainties, which have seldom been explored in previous soybean yield estimation studies.

In future studies, the further addition of heterogenous yield predictors using the multi-stream modeling architecture can be evaluated. For instance, the county-level average historical yield may potentially be added as another predictor group under the multi-stream structures to account for the spatial dependency in soybean yield, which may further improve the model accuracy and reduce the predictive uncertainty. Incorporating yield predictors characteristic of agricultural management advances as well as yield detrending methods may also help refine the model for more accurate crop yield estimations. Beyond the varying historical yield scenarios, the PB-CNN can further be evaluated over extended regions of varying environmental conditions as well as abnormal years when heat or water stress is more severe in the future. With the important role of crop phenology in constructing the Phenology Imagery and characterizing the phenological stage-specific yield responses, the retrieval of crop phenology, particularly the early vegetative stages, may further be explored in future studies using the phenology matching models (e.g., hybrid phenology matching model) (Diao, 2020; Diao et al., 2021).

7. Conclusion

The continuing increase in global population and living standards dramatically boosts the agricultural production demand, resulting in food security challenges. Accurate estimation of crop yield is essential for the optimization of farm management practices as well as the evaluation of agricultural decision-making under future climate change. In this study, we develop a PB-CNN framework for county-level soybean yield estimation and uncertainty quantification. The PB-CNN framework mainly encompasses three components: Phenology Imagery construction, multi-stream Bayesian-CNN modeling, as well as feature and uncertainty analysis. By modeling crop yield responses to a heterogenous set of yield predictors across crop phenological stages, PB-CNN outperforms three advanced machine learning models (i.e., SVR, RF, and LSTM) in estimating the county-level soybean yield of the US Corn Belt for the testing years 2014–2018. It achieves an average RMSE of 4.622 bu/ac, an average R^2 of 0.709, and an average bias of −2.057 bu/ac. With the comprehensive feature importance and predictive uncertainty analysis, we found that the EVI2 predictor group as well as the soybean blooming to dropping leaves phenological stages is more critical in estimating the soybean yield. The soil predictor group as well as the early growing stages can improve the model estimation accuracy, yet potentially brings more uncertainties into the yield estimation. The further uncertainty disentanglement facilitates the evaluation of both aleatoric and epistemic uncertainties as well as corresponding distributions over the study site. Overall, the PB-CNN framework enables the modeling of complex crop yield responses to varying phenological stage-

specific yield predictors and associated uncertainties, which further helps optimize farm management strategies to support the building of more sustainable agricultural systems.

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 study is supported in part by the National Science Foundation under Grant 2048068, in part by the National Aeronautics and Space Administration under Grant 80NSSC21K0946, and in part by the United States Department of Agriculture under Grant 2021-67021-33446.

Appendix

A-1. Pre-processing of EVI2 time series

Despite the rigorous protocols for generating the MCD43A4, this MODIS product may still be subject to snow-contamination and residual atmospheric interference, resulting in abnormal and spurious surface reflectance values. To eliminate the influence of those abnormal observations, we leverage three vegetation indices (EVI2, Normalized Difference Vegetation Index [NDVI], and Normalized Difference Water Index [NDWI]) by following two criteria (Zhang et al., 2020). The first criterion is that EVI2 values that are smaller than the co-located corresponding NDWI values should be excluded. NDWI values are sensitive to contaminations from the residual cloud, land surface moisture, and snow cover. These contaminations can be detected when NDWI values are larger than the corresponding EVI2 values. The second criterion is that abnormally high EVI2 values caused by inaccurate atmospheric correction and other factors should be excluded. As NDVI values are less sensitive to inaccurate atmospheric correction, we remove the spurious EVI2 values that are larger than 90 % of the corresponding NDVI values with reference to previous global comparison study of MODIS NDVI and EVI2 time series (Zhang et al., 2018). We further remove the abnormal EVI2 values that are larger than 110 % of the maximum EVI2 value during the corresponding previous and succeeding one-month periods.

The pre-processed EVI2 time series is further interpolated, with the missing EVI2 on a specific day replaced by the mean EVI2 of its previous and following 15 days. The interpolated time series is then composited by the average EVI2 value on 3-day intervals, which can preserve the phenological dynamic patterns, reduce the subtle fluctuations, as well as decreasing the data volume (Zhang, 2015; Zhang et al., 2020).

A-2. Permutation feature importance analysis

The permutation feature importance is measured as the decrease of the model accuracy by the random permutation of the feature in the model. Specifically, we first construct the PB-CNN using the original training data $\{X, Y\}$. With the permutation importance feature selection method, we then calculate the model prediction error of PB-CNN using the permuted data $\{X(j^{'}), Y\}$, where $X(j^{'})$ represents the *j-th* input variable being randomly shuffled. The permutation feature importance is calculated as the difference between the model prediction error using the permuted data $\{X(j^{'}), Y\}$ and that using the original data $\{X, Y\}$ to investigate the importance of the *j-th* input variable. Given the similar characteristics of the predictors within a group associated with the yield, we mainly assess the relative importance of each predictor group in soybean yield estimation, instead of each individual yield predictor. For instance, we randomly shuffle all the predictors (i.e., four types of aggregated EVI2 predictors) in the satellite-based predictor group to assess the importance and role of this predictor group.

Besides the importance of each predictor group, we also investigate the contributions of each phenological stage in our framework using the permutation-based feature selection method. Here we test the model performance of the PB-CNN using the permuted data $\{X(pheno_j^{'}), Y\}$ where $X(pheno_j^{'})$ denotes that all the yield predictors in the *j-th* phenological stage are randomly shuffled. The contribution of the yield predictors of this phenological stage can be measured as the difference between the model prediction error using the permuted data $\{X(pheno_j^{'}), Y\}$ and that using the original data $\{X, Y\}$. Comparably, we further evaluate the importance of each predictor group in each phenological stage with the permutation feature importance method to comprehensively unravel the roles of diverse types of predictors throughout the soybean growing cycle in the yield estimations.

A-3. PB-CNN model setting and training

Before training the PB-CNN, we first tune the hyper-parameters in the model, including the number of convolutional blocks, the number of convolutional filters, the convolutional kernel size, and the dropout rate. Through experimental analysis with reference to previous studies (Russello, 2018; Sun et al., 2020; You et al., 2017), the ranges of the hyper-parameters are selected as follows: the number of convolutional blocks is within the values of [1, 2, 4, 6, 8]; the number of filters for the convolutional layers is within the values of [4, 8, 16, 32, 64, 128]; the kernel size of the convolutional filters is set as 3; the dropout rate is set within the values of [0.25, 0.5]. The combination of those hyper-parameter values of selected ranges is then used to build the hyper-parameter dictionary, and the random search is performed on the hyper-parameter dictionary based on the tenfold cross-validation to find a tuple of hyper-parameters that yields an optimal model that minimizes the loss function.

During the model training, the initial learning rate is set to 0.0001, and the Adam optimizer is applied to minimize the loss. The batch size is set as 64, and the maximum number of iterations is set as 1500. Early stopping based on the validation loss is used in the learning process to prevent the network from overfitting when the validation loss stops decreasing. The data acquisition and pre-processing are conducted on the Google Earth Engine (GEE) platform (Gorelick et al., 2017), and the PB-CNN model is constructed and trained using the Keras 2.8 (the Python deep learning library with Python version 3.7) (Nguyen et al., 2019).

A-4. Benchmark models

In this study, we compare the performance of PB-CNN with that of three advanced benchmark models, including SVR, RF, and LSTM. As a widely used supervised learning model, SVR employs a kernel function to project the input data into a higher dimensional feature space and optimizes the

hyperplane in the new feature space to fit all the data points via a margin of tolerance. Because the kernels can accommodate non-linear relationships, SVR's computational complexity does not depend on the dimensionality of the input space and has good generalization capability (Awad & Khanna, 2015). RF is an ensemble learning method that constructs a multitude of decision trees to produce ensemble predictions. Because each decision tree is learned using random subsets of features and bootstrap samples, RF can reduce the overfitting of data and improve accuracy by integrating different trees. As a type of Recurrent Neural Network (RNN), LSTM can capture long-range temporal dependencies underlying sequential data and is also considered in this study. By utilizing a combination of forget, input, and output gates, LSTM can selectively update or remove the memory in the cell state to learn the temporally evolving features. This gating mechanism controls the propagation of temporally evolving information and enables the information to be stored in memory for a more extended period. LSTM thus can overcome the vanishing gradient problems in conventional RNNs.

The SVR and RF models are implemented with the scikit-learn library (Python version 3.7) (Pedregosa et al., 2011), and the LSTM model is constructed using Keras 2.8 (Python version 3.7). The input predictors in PB-CNN are used in these benchmark models for a fair comparison. Four Phenology Imageries are stacked together and transformed into six vectors according to phenological stages as the input for LSTM, and one vector as the input for SVR and RF. All model hyper-parameters are tuned based on the ten-fold cross-validation.

A-5. Evaluation of satellite-based Phenology product

The satellite-based phenology product provides pixel-level estimations of phenology stage transition dates, which are estimated using the threshold-based method to characterize the timing of plant greenness and vigor change throughout its phenological cycle. Specifically, the greenup, midgreenup, and maturity transition dates correspond to the timing when 15 %, 50 %, and 90 % of the amplitude of EVI2 time series are reached in the upward direction, respectively. The peak greenness date is estimated via the timing of maximum EVI2 of the phenological cycle. The senescence, midgreendown, and dormancy dates correspond to the timing when 90 %, 50 %, and 15 % of the amplitude of EVI2 time series are reached in the downward direction, respectively. We evaluate the ability of the satellite-based vegetation phenology product in characterizing the soybean growing stages using the ground-based CPRs. The CPRs document the proportion of soybean reaching certain phenological growing stages on a weekly basis throughout the season at the state level. Specifically, we compare the medians of satellite-derived phenology stage transition dates (i.e., onset of greenup, greenup midpoint, maturity, peak greenness, senescence, greendown midpoint, and dormancy) with the CPR-documented median dates of soybean growing stages (i.e., emerged, blooming, setting pods, dropping leaves, and harvest stages) from 2008 to 2018 at the state-level for our study site.

The state-level median dates of satellite-based phenological metrics and CPR-based soybean phenological measures throughout the study area are both shown in Fig. 7. All the satellite-based phenological metrics are represented by solid lines, and CPR-based crop phenological measures are represented by dashed lines in the figure. Overall, the satellite- and corresponding CPR-based median phenological transition dates exhibit comparable variation patterns across years and states. The median dates of those phenological measures vary from year to year, with relatively earlier dates in 2010 and 2012. The median dates also differ from state to state, with relatively earlier dates in Minnesota and South Dakota as well as later dates in North Dakota. Across years and locations, the variations in the timing of soybean entering the phenological stages may be caused by a diversity of weather conditions, soil conditions, and farming practices.

The CPR-documented soybean growing stages can be approximated by satellite-based phenological measures (Table 2). Specifically, the satellite-based greenup/mid-greenup metrics can be connected to soybean's emerged stage. The satellite-based maturity metric has implications for the blooming stage of soybean. The timing of peak greenness metric aligns with the soybean setting pods stage. The mid-greendown metric can be connected to the dropping leaves stage of soybean. The soybean harvest stage can be approximated by the satellite-based dormancy metric. For the whole study area, the mean and standard deviation of the median phenological dates for both satellite-based and CPR-based measures from 2008 to 2018 are summarized in Table 2. Except for the soybean emerged stage, the median transition dates of all the other phenology stages can be aligned by the corresponding satellite-derived metrics, with the mean difference less than 5 days and standard deviation less than 10 days.

References

Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., Fieguth, P., Cao, X., Khosravi, A., Acharya, U.R., Makarenkov, V., Nahavandi, S., 2021. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. Information Fusion 76, 243–297. https://doi.org/10.1016/j.inffus.2021.05.008.

Anderson, M.C., Hain, C.R., Jurecka, F., Trnka, M., Hlavinka, P., Dulaney, W., Otkin, J. A., Johnson, D., Gao, F., 2016. Relationships between the evaporative stress index and winter wheat and spring barley yield anomalies in the Czech Republic. Climate Res. 70 (2–3), 215–230. https://doi.org/10.3354/cr01411.

Ashapure, A., Jung, J., Chang, A., Oh, S., Yeom, J., Maeda, M., Maeda, A., Dube, N., Landivar, J., Hague, S., Smith, W., 2020. Developing a machine learning based cotton yield estimation framework using multi-temporal UAS data. ISPRS J. Photogramm. Remote Sens. 169, 180–194. https://doi.org/10.1016/j.isprsjprs.2020.09.015.

Awad, M., Khanna, R., 2015. Support Vector Regression. Efficient Learning Machines 67–80. https://doi.org/10.1007/978-1-4302-5990-9_4.

Barbosa, A., Trevisan, R., Hovakimyan, N., Martin, N.F., 2020. Modeling yield response to crop management using convolutional neural networks. Comput. Electron. Agric. 170, 105197 https://doi.org/10.1016/j.compag.2019.105197.

Basso, B., Liu, L., 2019. Seasonal crop yield forecast: Methods, applications, and accuracies. In: Advances in Agronomy, Vol. 154. Academic Press Inc., pp. 201–255. https://doi.org/10.1016/bs.agron.2018.11.002

Bassu, S., Brisson, N., Durand, J.L., Boote, K., Lizaso, J., Jones, J.W., Rosenzweig, C., Ruane, A.C., Adam, M., Baron, C., Basso, B., Biernath, C., Boogaard, H., Conijn, S., Corbeels, M., Deryng, D., De Sanctis, G., Gayler, S., Grassini, P., Waha, K., 2014. How do various maize crop models vary in their responses to climate change factors? Glob. Chang. Biol. 20 (7), 2301–2320. https://doi.org/10.1111/gcb.12520.

Becker-Reshef, I., Vermote, E., Lindeman, M., Justice, C., 2010. A generalized regression-based model for forecasting winter wheat yields in Kansas and Ukraine using MODIS data. Remote Sens. Environ. 114 (6), 1312–1323. https://doi.org/10.1016/j.rse.2010.01.010.

Benedetti, R., Rossini, P., 1993. On the use of NDVI profiles as a tool for agricultural statistics: The case study of wheat yield estimate and forecast in Emilia Romagna. Remote Sens. Environ. 45 (3), 311–326. https://doi.org/10.1016/0034-4257(93)90113-C.

Bolton, D.K., Friedl, M.A., 2013. Forecasting crop yield using remotely sensed vegetation indices and crop phenology metrics. Agric. For. Meteorol. 173, 74–84. https://doi.org/10.1016/j.agrformet.2013.01.007.

Boryan, C., Yang, Z., Mueller, R., & Craig, M. (2011). Monitoring US agriculture: the US Department of Agriculture, National Agricultural Statistics Service, Cropland Data Layer Program. Http://Dx.Doi.Org/10.1080/10106049.2011.562309, *26*(5), 341–358. 10.1080/10106049.2011.562309.

Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32. https://doi.org/10.1023/A:1010933404324.

Butts-Wilmsmeyer, C.J., Seebauer, J.R., Singleton, L., Below, F.E., 2019. Weather during key growth stages explains grain quality and yield of maize. Agronomy 9 (1). https://doi.org/10.3390/agronomy9010016.

Carletto, C., Jolliffe, D., Banerjee, R., 2015. From Tragedy to Renaissance: Improving Agricultural Data for Better Policies. https://doi.org/10.1080/00220388.2014.968140, 51(2), 133-148. 10(1080/00220388), pp. 968140, 2014.

Challinor, A.J., Watson, J., Lobell, D.B., Howden, S.M., Smith, D.R., Chhetri, N., 2014. A meta-analysis of crop yield under climate change and adaptation. Nat. Clim. Chang. 4 (4), 287–291. https://doi.org/10.1038/nclimate2153.

Chlingaryan, A., Sukkarieh, S., Whelan, B., 2018. Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review. Comput. Electron. Agric. 151 (June), 61–69. https://doi.org/10.1016/j.compag.2018.05.012.

Cohen, I., Zandalinas, S.I., Fritschi, F.B., Sengupta, S., Fichman, Y., Azad, R.K., Mittler, R., 2021. The impact of water deficit and heat stress combination on the molecular response, physiology, and seed production of soybean. Physiol. Plant. 172 (1), 41–52. https://doi.org/10.1111/PPL.13269.

Crane-Droesch, A., 2018. Machine learning methods for crop yield prediction and climate change impact assessment in agriculture. Environ. Res. Lett. 13 (11), 114003 https://doi.org/10.1088/1748-9326/aae159.

Dadsetan, S., Rose, G., Hovakimyan, N., & Hobbs, J. (2020). *Detection and Prediction of Nutrient Deficiency Stress using Longitudinal Aerial Imagery*. 10.48550/arxiv.2012.09654.

Depeweg, S., Hernández-Lobato, J. M., Doshi-Velez, F., & Udluft, S. (2017). *Uncertainty Decomposition in Bayesian Neural Networks with Latent Variables*. https://arxiv.org/abs/1706.08495v2.

Diao, C., Li, G., 2022. Near-Surface and High-Resolution Satellite Time Series for Detecting Crop Phenology. Remote Sens. (Basel) 14 (9). https://doi.org/10.3390/RS14091957.

Diao, C., Yang, Z.Z., Gao, F., Zhang, X., Yang, Z.Z., 2021. Hybrid phenology matching model for robust crop phenological retrieval. ISPRS J. Photogramm. Remote Sens. 181, 308–326. https://doi.org/10.1016/J.ISPRSJPRS.2021.09.011.

Diao, C., 2019. Innovative pheno-network model in estimating crop phenological stages with satellite time series. ISPRS J. Photogramm. Remote Sens. 153, 96–109. https://doi.org/10.1016/J.ISPRSJPRS.2019.04.012.

Diao, C., 2020. Remote sensing phenological monitoring framework to characterize corn and soybean physiological growing stages. Remote Sens. Environ. 248, 111960 https://doi.org/10.1016/j.rse.2020.111960.

Feng, S., Hao, Z., Zhang, X., Hao, F., 2021. Changes in climate-crop yield relationships affect risks of crop yield reduction. Agric. For. Meteorol. 304–305 https://doi.org/10.1016/j.agrformet.2021.108401.

Fernandez-Beltran, R., Baidar, T., Kang, J., & Pla, F. (2021). Rice-Yield Prediction with Multi-Temporal Sentinel-2 Data and 3D CNN: A Case Study in Nepal. *Remote Sensing 2021, Vol. 13, Page 1391*, *13*(7), 1391. 10.3390/RS13071391.

Gal, Y., & Ghahramani, Z. (2015). *Bayesian Convolutional Neural Networks with Bernoulli Approximate Variational Inference*. http://arxiv.org/abs/1506.02158.

Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. *33rd International Conference on Machine Learning, ICML 2016*, *3*, 1651–1660. https://arxiv.org/abs/1506.02142v6.

Ganguly, S., Friedl, M.A., Tan, B., Zhang, X., Verma, M., 2010. Land surface phenology from MODIS: Characterization of the Collection 5 global land cover dynamics product. Remote Sens. Environ. 114 (8), 1805–1816. https://doi.org/10.1016/j.rse.2010.04.005.

Goparaju, L., Ahmad, F., 2019. Analysis of Seasonal Precipitation, Potential Evapotranspiration, Aridity, Future Precipitation Anomaly and Major Crops at District Level of India. KN - Journal of Cartography and Geographic Information 69 (2), 143–154. https://doi.org/10.1007/S42489-019-00020-4/FIGURES/6.

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., Moore, R., 2017. Google Earth Engine: Planetary-scale geospatial analysis for everyone. Remote Sens. Environ. 202, 18–27. https://doi.org/10.1016/j.rse.2017.06.031.

Grassini, P., Specht, J.E., Tollenaar, M., Ciampitti, I., Cassman, K.G., 2015. High-yield maize–soybean cropping systems in the US Corn Belt. Crop Physiology: Applications for Genetic Improvement and Agronomy: Second Edition 17–41. https://doi.org/10.1016/B978-0-12-417104-6.00002-9.

Guan, K., Wu, J., Kimball, J.S., Anderson, M.C., Frolking, S., Li, B., Hain, C.R., Lobell, D. B., 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. https://doi.org/10.1016/J.RSE.2017.06.043.

Guo, Y., Fu, Y., Hao, F., Zhang, X., Wu, W., Jin, X., Robin Bryant, C., Senthilnath, J., 2021. Integrated phenology and climate in rice yields prediction using machine learning methods. Ecol. Ind. 120, 106935 https://doi.org/10.1016/j.ecolind.2020.106935.

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

Huang, J., Tian, L., Liang, S., Ma, H., Becker-Reshef, I., Huang, Y., Su, W., Zhang, X., Zhu, D., Wu, W., 2015. Improving winter wheat yield estimation by assimilation of the leaf area index from Landsat TM and MODIS data into the WOFOST model. Agric. For. Meteorol. 204, 106–121. https://doi.org/10.1016/j.agrformet.2015.02.001.

Hüllermeier, E., Waegeman, W., 2021. Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. Mach. Learn. 110 (3), 457–506. https://doi.org/10.1007/s10994-021-05946-3.

Jaafar, H., & Mourad, R. (2021). GYMEE: A Global Field-Scale Crop Yield and ET Mapper in Google Earth Engine Based on Landsat, Weather, and Soil Data. *Remote Sensing 2021, Vol. 13, Page 773*, *13*(4), 773. 10.3390/RS13040773.

Jiang, H., Hu, H., Wang, S., Ying, Y., Lin, T., 2020a. Understanding the impact of sub-seasonal meteorological variability on corn yield in the U.S. Corn Belt. Sci. Total Environ. 724, 138235 https://doi.org/10.1016/j.scitotenv.2020.138235.

Jiang, H., Hu, H., Zhong, R., Xu, J.J., Xu, J.J., Huang, J., Wang, S., Ying, Y., Lin, T., 2020b. 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 (3), 1754–1766. https://doi.org/10.1111/gcb.14885.

Jiang, Z., Liu, C., Ganapathysubramanian, B., Hayes, D.J., Sarkar, S., 2020c. Predicting county-scale maize yields with publicly available data. Sci. Rep. 10 (1), 1–12. https://doi.org/10.1038/s41598-020-71898-8.

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. https://doi.org/10.1016/j.rse.2013.10.027.

Jumrani, K., Bhatia, V.S., 2018. Impact of combined stress of high temperature and water deficit on growth and seed yield of soybean. Physiol. Mol. Biol. Plants 24 (1), 37–50. https://doi.org/10.1007/S12298-017-0480-5/FIGURES/8.

Jumrani, K., Bhatia, V.S., 2019. Interactive effect of temperature and water stress on physiological and biochemical processes in soybean. Physiol. Mol. Biol. Plants 25 (3), 667–681. https://doi.org/10.1007/S12298-019-00657-5/FIGURES/7.

Kang, Y., Ozdogan, M., Zhu, X., Ye, Z., Hain, C., Anderson, M., 2020. Comparative assessment of environmental variables and machine learning algorithms for maize yield prediction in the US Midwest. Environ. Res. Lett. 15 (6), 064005 https://doi.org/10.1088/1748-9326/ab7df9.

Kattenborn, T., Leitloff, J., Schiefer, F., Hinz, S., 2021. Review on Convolutional Neural Networks (CNN) in vegetation remote sensing. ISPRS J. Photogramm. Remote Sens. 173, 24–49. https://doi.org/10.1016/J.ISPRSJPRS.2020.12.010.

Kaylen, M.S., Koroma, S.S., 1991. Trend, Weather Variables, and the Distribution of U.S. Corn Yields. Rev. Agric. Econ. 13 (2), 249. https://doi.org/10.2307/1349641.

Kendall, A., & Gal, Y. (2017). What uncertainties do we need in Bayesian deep learning for computer vision? *Advances in Neural Information Processing Systems*, *2017-Decem*, 5575–5585.

Kiureghian, A.D., Ditlevsen, O., 2009. Aleatory or epistemic? Does it matter? Struct. Saf. 31 (2), 105–112. https://doi.org/10.1016/J.STRUSAFE.2008.06.020.

Kogan, F., Guo, W., 2016. Strong 2015–2016 El Niño and implication to global ecosystems from space data. Http://Dx.Doi.Org/10.1080/01431161.2016.1259679 38 (1), 161–178. https://doi.org/10.1080/01431161.2016.1259679.

Koirala, A., Walsh, K. B., Wang, Z., & McCarthy, C. (2019). Deep learning – Method overview and review of use for fruit detection and yield estimation. In *Computers and Electronics in Agriculture* (Vol. 162, pp. 219–234). Elsevier B.V. 10.1016/j.compag.2019.04.017.

Kukal, M.S., Irmak, S., 2018. Climate-Driven Crop Yield and Yield Variability and Climate Change Impacts on the U.S. Great Plains Agricultural Production. *Scientific Reports* 8 (1), 1–18. https://doi.org/10.1038/s41598-018-21848-2.

Lencucha, R., Pal, N.E., Appau, A., Thow, A.M., Drope, J., 2020. Government policy and agricultural production: A scoping review to inform research and policy on healthy agricultural commodities. Glob. Health 16 (1), 1–15. https://doi.org/10.1186/S12992-020-0542-2/TABLES/5.

Lin, M., Chen, Q., & Yan, S. (2013). Network In Network. *2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings*. 10.48550/arxiv.1312.4400.

Liu, Y.R., Li, Y.P., Huang, G.H., Zhang, J.L., Fan, Y.R., 2017. A Bayesian-based multilevel factorial analysis method for analyzing parameter uncertainty of hydrological model. J. Hydrol. 553, 750–762. https://doi.org/10.1016/J.JHYDROL.2017.08.048.

Lobell, D.B., Cassman, K.G., Field, C.B., 2009. Crop Yield Gaps: Their Importance, Magnitudes, and Causes. Http://Dx.Doi.Org/10.1146/Annurev.Environ.041008.093740 34, 179–204. https://doi.org/10.1146/ANNUREV.ENVIRON.041008.093740.

Lobell, D.B., Roberts, M.J., Schlenker, W., Braun, N., Little, B.B., Rejesus, R.M., Hammer, G.L., 2014. Greater sensitivity to drought accompanies maize yield increase in the U.S. Midwest. *Science* 344 (6183), 516–519. https://doi.org/10.1126/SCIENCE.1251423/SUPPL_FILE/LOBELLETAL.DATABASES1.ZIP.

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 https://doi.org/10.1016/j.rse.2021.112408.

Maimaitijiang, M., Sagan, V., Sidike, P., Hartling, S., Esposito, F., Fritschi, F.B., 2020. Soybean yield prediction from UAV using multimodal data fusion and deep learning. Remote Sens. Environ. 237, 111599 https://doi.org/10.1016/j.rse.2019.111599.

Meng, T., Carewa, R., Florkowski, W.J., Klepacka, A.M., 2017. Analyzing Temperature and Precipitation Influences on Yield Distributions of Canola and Spring Wheat in Saskatchewan. J. Appl. Meteorol. Climatol. 56 (4), 897–913. https://doi.org/10.1175/JAMC-D-16-0258.1.

Mobiny, A., Yuan, P., Moulik, S.K., Garg, N., Wu, C.C., Van Nguyen, H., 2021. DropConnect is effective in modeling uncertainty of Bayesian deep networks. Sci. Rep. 11 (1), 1–14. https://doi.org/10.1038/s41598-021-84854-x.

Muruganantham, P., Wibowo, S., Grandhi, S., Samrat, N. H., & Islam, N. (2022). A Systematic Literature Review on Crop Yield Prediction with Deep Learning and Remote Sensing. In *Remote Sensing* (Vol. 14, Issue 9, p. 1990). Multidisciplinary Digital Publishing Institute. 10.3390/rs14091990.

Müller, C., Franke, J., Jägermeyr, J., Ruane, A.C., Elliott, J., Moyer, E., Heinke, J., Falloon, P.D., Folberth, C., Francois, L., Hank, T., Izaurralde, R.C., Jacquemin, I., Liu, W., Olin, S., Pugh, T.A.M., Williams, K., Zabel, F., 2021. Exploring uncertainties in global crop yield projections in a large ensemble of crop models and CMIP5 and CMIP6 climate scenarios. Environ. Res. Lett. 16 (3), 034040 https://doi.org/10.1088/1748-9326/ABD8FC.

Naeve, S.L., Miller-Garvin, J., 2019. United States Soybean Quality. American Soybean Association.

NASS CDL, 2016. https://nassgeodata.gmu.edu/CropScape/ (last accessed Oct 1, 2023).

NASS CPR, 2018. https://www.nass.usda.gov/Publications/National_Crop_Progress/ (last accessed Oct 1, 2023).

Nguyen, G., Dlugolinsky, S., Bobák, M., Tran, V., López García, Á., Heredia, I., Malík, P., Hluchý, L., 2019. Machine Learning and Deep Learning frameworks and libraries for large-scale data mining: a survey. Artif. Intell. Rev. 52 (1), 77–124. https://doi.org/10.1007/S10462-018-09679-Z.

Ortez, O.A., McMechan, A.J., Hoegemeyer, T., Ciampitti, I.A., Nielsen, R.L., Thomison, P. R., Abendroth, L.J., Elmore, R.W., 2022. Conditions potentially affecting corn ear formation, yield, and abnormal ears: A review. Crop, Forage & Turfgrass Management 8 (2), e20173.

Pedregosa, F., Weiss, R., Brucher, M., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, É., 2011. Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res. 12, 2825–2830. http://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html%5Cnhttp://arxiv.org/abs/1201.0490.

Pourmohammadali, B., Hosseinifard, S.J., Hassan Salehi, M., Shirani, H., Esfandiarpour Boroujeni, I., 2019. Effects of soil properties, water quality and management practices on pistachio yield in Rafsanjan region, southeast of Iran. Agric Water Manag 213, 894–902. https://doi.org/10.1016/J.AGWAT.2018.12.005.

Rasmussen, M.S., 1992. Assessment of millet yields and production in northern Burkina Faso using integrated NDVI from the AVHRR. Int. J. Remote Sens. 13 (18), 3431–3442. https://doi.org/10.1080/01431169208904132.

Raun, W.R., Solie, J.B., Johnson, G.V., Stone, M.L., Mutten, R.W., Freeman, K.W., Thomason, W.E., Lukina, E.V., 2002. Improving nitrogen use efficiency in cereal grain production with optical sensing and variable rate application. In. Agron. J. Vol. 94 (Issue 4) https://doi.org/10.2134/agronj2002.8150.

Ray, D.K., Gerber, J.S., Macdonald, G.K., West, P.C., Shuai, J., Zhang, Z., Tao, F., Shi, P., Legler, D.M., Bryant, K.J., O'Brien, J.J., Podestá, G.P., Messina, C.D., Grondona, M. O., Magrin, G.O., Barros, V.R., Boninsegna, J.A., Camilloni, I.A., Chidiak, M., Glantz, M.H., 2015. Climate variation explains a third of global crop yield variability. Nat. Commun. 6 (1), 1–9. https://doi.org/10.1038/ncomms6989.

Rodell, M., Houser, P.R., Jambor, U., Gottschalck, J., Mitchell, K., Meng, C.J., Arsenault, K., Cosgrove, B., Radakovich, J., Bosilovich, M., Entin, J.K., Walker, J.P., Lohmann, D., Toll, D., 2004. The Global Land Data Assimilation System. Bull. Am. Meteorol. Soc. 85 (3), 381–394. https://doi.org/10.1175/BAMS-85-3-381.

Russello, H. (2018). Convolutional Neural Networks for Crop Yield Prediction using Satellite Images. *IBM Center for Advanced Studies, January*. https://www.semanticscholar.org/paper/Convolutional-Neural-Networks-for-Crop-Yield-using-Russello-Shang/3fb98e76ffd8ba79e1c22eda4d640da0c037e98a.

Schaaf, C., & Wang, Z. (2015). MCD43A4 MODIS/Terra+ Aqua BRDF/Albedo Nadir BRDF Adjusted Ref Daily L3 Global-500m V006. NASA EOSDIS Land Processes DAAC. *USGS Earth Resources Observation and Science (EROS) Center, Sioux Falls, South Dakota (Https://Lpdaac. Usgs. Gov)*.

Shahhosseini, M., Hu, G., Archontoulis, S.V., 2020. Forecasting Corn Yield With Machine Learning Ensembles. Front. Plant Sci. 11, 1120. https://doi.org/10.3389/fpls.2020.01120.

Shahhosseini, M., Hu, G., Huber, I., Archontoulis, S.V., 2021. Coupling machine learning and crop modeling improves crop yield prediction in the US Corn Belt. Sci. Rep. 11 (1), 1606. https://doi.org/10.1038/s41598-020-80820-1.

Sheng, M., Liu, J., Zhu, A.X., Rossiter, D.G., Liu, H., Liu, Z., Zhu, L., 2019. Comparison of GLUE and DREAM for the estimation of cultivar parameters in the APSIM-maize model. Agric. For. Meteorol. 278, 107659 https://doi.org/10.1016/J.AGRFORMET.2019.107659.

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 (3), 348–363. https://doi.org/10.1108/AFR-05-2013-0021/FULL/XML.

Shirani, H., Habibi, M., Besalatpour, A.A., Esfandiarpour, I., 2015. Determining the features influencing physical quality of calcareous soils in a semiarid region of Iran using a hybrid PSO-DT algorithm. Geoderma 259–260, 1–11. https://doi.org/10.1016/J.GEODERMA.2015.05.002.

Shridhar, K., Laumann, F., & Liwicki, M. (2019). *A Comprehensive guide to Bayesian Convolutional Neural Network with Variational Inference*. 10.48550/arxiv.1901.02731.

Son, N.T., Chen, C.F., Chen, C.R., Minh, V.Q., Trung, N.H., 2014. A comparative analysis of multitemporal MODIS EVI and NDVI data for large-scale rice yield estimation. Agric. For. Meteorol. 197, 52–64. https://doi.org/10.1016/J.AGRFORMET.2014.06.007.

Sun, J., Di, L., Sun, Z., Shen, Y., Lai, Z., 2019. County-level soybean yield prediction using deep CNN-LSTM model. Sensors (switzerland) 19 (20), 1–21. https://doi.org/10.3390/s19204363.

Sun, J., Lai, Z., Di, L., Sun, Z., Tao, J., Shen, Y., 2020. Multilevel Deep Learning Network for County-Level Corn Yield Estimation in the U.S. Corn Belt. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 13, 5048–5060. https://doi.org/10.1109/JSTARS.2020.3019046.

Terliksiz, A. S., & Altylar, D. T. (2019). Use of deep neural networks for crop yield prediction: A case study of soybean yield in lauderdale county, Alabama, USA. *2019 8th International Conference on Agro-Geoinformatics, Agro-Geoinformatics 2019*, 9–12. 10.1109/Agro-Geoinformatics.2019.8820257.

Tian, H., Wang, P., Tansey, K., Zhang, J., Zhang, S., Li, H., 2021. An LSTM neural network for improving wheat yield estimates by integrating remote sensing data and meteorological data in the Guanzhong Plain PR China. *Agric. For. Meteorol.* 310, 108629 https://doi.org/10.1016/J.AGRFORMET.2021.108629.

USDA, 2022. Usda-national agricultural statistics service, cropland data layer. USDA-NASS, Washington, DC.

USDA NASS. (2022). *Farms and Land in Farms Summary 2021*. *February*, 1–17.

van Klompenburg, T., Kassahun, A., Catal, C., 2020. Crop yield prediction using machine learning: A systematic literature review. Comput. Electron. Agric. 177 (August), 105709 https://doi.org/10.1016/j.compag.2020.105709.

Wang, X., Huang, J., Feng, Q., Yin, D., 2020a. Winter wheat yield prediction at county level and uncertainty analysis in main wheat-producing regions of China with deep learning approaches. Remote Sens. (Basel) 12 (11), 1744. https://doi.org/10.3390/rs12111744.

Wang, Y., Zhang, Z., Feng, L., Du, Q., Runge, T., 2020b. Combining multi-source data and machine learning approaches to predict winter wheat yield in the conterminous United States. Remote Sens. (Basel) 12 (8). https://doi.org/10.3390/RS12081232.

Wei, T., Cherry, T.L., Glomrød, S., Zhang, T., 2014. Climate change impacts on crop yield: Evidence from China. Sci. Total Environ. 499, 133–140. https://doi.org/10.1016/J.SCITOTENV.2014.08.035.

Weiss, M., Jacob, F., Duveiller, G., 2020. Remote sensing for agricultural applications: A meta-review. Remote Sens. Environ. 236 (August 2019), 111402 https://doi.org/10.1016/j.rse.2019.111402.

Wright, C.K., Wimberly, M.C., 2013. Recent land use change in the Western Corn Belt threatens grasslands and wetlands. *PNAS* 110 (10), 4134–4139. https://doi.org/10.1073/PNAS.1215404110/SUPPL_FILE/PNAS.201215404SI.PDF.

Yang, H.S., Dobermann, A., Lindquist, J.L., Walters, D.T., Arkebauer, T.J., Cassman, K.G., 2004. Hybrid-maize - A maize simulation model that combines two crop modeling approaches. Field Crop Res 87 (2–3), 131–154. https://doi.org/10.1016/j.fcr.2003.10.003.

Yang, J., Reichert, P., Abbaspour, K.C., Xia, J., Yang, H., 2008. Comparing uncertainty analysis techniques for a SWAT application to the Chaohe Basin in China. J. Hydrol. 358 (1–2), 1–23. https://doi.org/10.1016/J.JHYDROL.2008.05.012.

Yao, Y., Miao, Y., Huang, S., Gao, L., Ma, X., Zhao, G., Jiang, R., Chen, X., Zhang, F., Yu, K., Gnyp, M.L., Bareth, G., Liu, C., Zhao, L., Yang, W., Zhu, H., 2012. Active canopy sensor-based precision N management strategy for rice. Agron. Sustain. Dev. 32 (4), 925–933. https://doi.org/10.1007/s13593-012-0094-9.

You, J., Li, X., Low, M., Lobell, D., & Ermon, S. (2017). Deep Gaussian process for crop yield prediction based on remote sensing data. *31st AAAI Conference on Artificial Intelligence, AAAI 2017*, 4559–4565. www.aaai.org.

Yu, S., Jia, S., Xu, C., 2017. Convolutional neural networks for hyperspectral image classification. Neurocomputing 219, 88–98. https://doi.org/10.1016/j.neucom.2016.09.010.

Zhang, X., Liu, L., Liu, Y., Jayavelu, S., Wang, J., Moon, M., Henebry, G.M., Friedl, M.A., Schaaf, C.B., 2018. Generation and evaluation of the VIIRS land surface phenology product. Remote Sens. Environ. 216, 212–229. https://doi.org/10.1016/j.rse.2018.06.047.

Zhang, Y., Zhao, Y., Feng, L., 2019. Higher contributions of uncertainty from global climate models than crop models in maize-yield simulations under climate change. Meteorol. Appl. 26 (1), 74–82. https://doi.org/10.1002/met.1738.

Zhang, X., Wang, J., Henebry, G.M., Gao, F., 2020. Development and evaluation of a new algorithm for detecting 30 m land surface phenology from VIIRS and HLS time series. ISPRS J. Photogramm. Remote Sens. 161 (August 2019), 37–51. https://doi.org/10.1016/j.isprsjprs.2020.01.012.

Zhang, X., 2015. Reconstruction of a complete global time series of daily vegetation index trajectory from long-term AVHRR data. Remote Sens. Environ. 156, 457–472. https://doi.org/10.1016/J.RSE.2014.10.012.