NortheastChinaSoybeanYield20m: an annual soybean yield dataset at 20 m in Northeast China from 2019 to 2023

Jingyuan Xu[1,2], Xin Du[1,2], Taifeng Dong[3], Qiangzi Li[1,2], Yuan Zhang[1,2], Hongyan Wang[1,2], Jing Xiao[1,2], Jiashu Zhang[1,4], Yunqi Shen[1,2], Yong Dong[1,2]

[1]Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China
[2]University of Chinese Academy of Sciences, Beijing 100190, China
[3]National Wildlife Research Centre, Environment and Climate Change Canada, 1125 Colonel By Drive, Ottawa, ON K1A0H3, Canada
[4]School of Science, China University of Geosciences (Beijing), Beijing 100083, China

Correspondence to: Xin Du (duxin@aircas.ac.cn)

Abstract. Accurate monitoring of crop yield is critical~~important~~ for ensuring food security. While various~~However, exiting~~ yield datasets covering Northeast China exist, they were produced at a ~~with a~~ coarse spatial resolution and remain~~are~~ inadequate for capturing small~~-~~-scale spatial heterogeneity. Current yield estimation methods, such as machine learning models and~~or~~ the assimilation of remotely sensed biophysical variables into crop growth models, are~~depend~~ heavily reliant on ground observations and ~~involve significant~~ computationally expensive ~~costs~~. To address~~solve~~ these~~-~~ limitations~~problems~~, we propose a hybrid framework that couples~~ing~~ the World Food Studies Simulation Model (WOFOST) and ~~the~~ a Gated Recurrent Unit ~~model~~ (GRU) model~~-~~ ~~was proposed~~ to generate a high-resolution (20 m) soybean yield dataset in Northeast China from 2019 to 2023 (NortheastChindaSoybeanYield20m). First, to generate a comprehensive training dataset, WOFOST was employed to simulate diverse~~A~~ soybean growth scenarios by ~~dataset was first generated based on the WOFOST that simulated various production scenarios~~ accounting for variations in ~~(~~climates, crop varieties, soil types and agro-managements practices~~)~~. The GRU model was then trained to establish~~for characterizing~~ relationships between model simulated leaf area index (LAI) and soybean yield. The trained model was ~~then~~ applied to estimate ~~for~~ soybean yield~~estimation~~ in Northeast China using time-series LAI derived from Sentinel-2 at key ~~of different~~ growth stages~~derived from Sentinel-2~~. The accuracy of estimates~~the dataset~~ was evaluated using~~by~~ in-situ measurements~~measured~~ and government statistical data. The overall accuracy was 287.44 kg ha^{-1} and 272.36 kg ha^{-1} in the root mean squared error (RMSE) for field and regional scale, respectively. The model exhibited consistent interannual stability, with ~~Stable results were achieved through the years with~~ mean relative error (MRE) ~~on~~ averaging~~e of~~ 11.46 % and 7.94% at the ~~in~~ municipal scale and ~~7.94 %~~ the ~~in~~ provincial scale, respectively. The dataset effectively captured spatiotemporal yield variability, offering ~~Results demonstrated that the model was able to capture spatial-temporal variation of soybean yield~~potentials~~. The NortheastChinaSoybeanYield20m was able to capture spatial temporal variation of soybean yield, which can be applied~~ for optimizing soybean production, ~~distribution and~~guiding precise

agriculture practices, ~~agricultural decision-making~~and informing agricultural policy. The NortheastChinaSoybeanYield20m dataset is publicly available at ~~can be downloaded from~~ https://doi.org/10.5281/zenodo.14263103 (Xu et al., 2024).

1 Introduction

Soybean is a crucial~~an important~~ crop for both food and oil production, providing~~supplying~~ more than a quarter of the world's edible protein (Graham and Vance, 2003). Global demand for soybean~~With rapid population growth, the demand for food continues to rise. It~~ is projected to ~~that global demand for soybeans will~~ increase by 46 % by 2050, driven by rapid population growth (Falcon et al., 2022). As an major~~important~~ traded agricultural commodity, soybean production in key exporting nations has wide-reaching effects on ~~the impact of soybean production in major producing countries can be felt globally through~~ international markets, and can significantly influence agricultural economies worldwide~~trade~~ (Qiao et al., 2023) ~~and thus affecting agricultural economies around the world~~. Notably, China is the world's largest consumer of soybeans (FAOSTAT, 2022), and its. ~~The~~ soybean demand relies~~in China is~~ heavily ~~dependent~~ on international trade (Zhao et al., 2023). Consequently, ~~In essence,~~ accurate monitoring of soybean yield is vital for promoting~~fostering~~ sustainable agriculture, ensuring food security, and maintaining economic stability from regional to ~~on a~~ global scale. Moreover, effective yield monitoring and mapping supports farmers by informing ~~In addition, it is also beneficial for farmers for managing~~ field management practices, bolstering ~~supporting~~ agricultural insurance and enhancing ~~improving~~ poverty alleviation initiatives~~efforts~~ (Zhuo et al., 2022).

Remote sensing data provides ~~has provided~~ time-series observations~~data support~~ for crop yield estimation across multiple ~~at different~~ scales (e.g., field, regional and national~~country~~) (Dong et al., 2020; Hunt et al., 2019; Zhao et al., 2023b). Current ~~The~~ methodologies for ~~crop~~ yield estimation ~~based on remote sensing data~~ can be broadly categorized as ~~further divided into~~ data-driven or ~~methods and~~ knowledge-driven approaches~~methods~~.

Data-driven methods leverage satellite-derived variables such as leaf area index (LAI), fraction of absorbed photosynthetically active radiation (FAPAR), and vegetation indices (VIs) to establish~~can explore the~~ linear or nonlinear relationships with measured ~~between satellite-derived variables and~~ crop yield (Ang et al., 2022; Xie et al., 2019). ~~Variables retrieved from remote sensing data include leaf area index (LAI), fraction of absorbed photosynthetically active radiation (FAPAR), leaf chlorophyll content (LCC) and vegetation indexes (VIs) (Ang et al., 2022; Xie et al., 2019). Due to their capability in handling large volumes of input data and nonlinear tasks, machine learning algorithms~~ Machine learning algorithms such as Random Forest (RF), and Artificial Neural Networks (ANN), due to their ability to process large dataset and model complex nonlinear interactions, have been widely applied in crop yield estimations (Pang et al., 2022; Tian et al., 2021; Yildirim et al., 2022). ~~In reality, it is difficult to meet the demand of high-precision yield estimation using a single type of indicator. Some literatures combined soil moisture (SM), evapotranspiration (ET) with crop physiological parameters to participate in modeling (Islam et al., 2023; Ji et al., 2022).~~ These methods can extract effective information from multi-source ~~a large amount of~~ structured or ~~/~~unstructured data without manual intervention ~~to meet the needs of multi-source data~~. However, they are heavily reliant on extensive ground-truth training data, which is challenging to collect over large areas and high time

intervals~~one disadvantage of data-driven approaches is that they dependent heavily on large training datasets~~ (Cao et al., 2021). ~~It is often challenging to obtain sufficient ground samples in large area application. Outliers and noise of the obtained data can also degrade the performance of the model and increase uncertainty in the prediction (Taylor et al., 2007). In addition,~~ Additionally, these models often overlook the impacts of environmental factors on crop growth, such as the influence of early-season soil moisture on root establishment or the effect of high temperatures during flowering on pod set, and are lack of interpretability, as they cannot~~due to the lack of theoretical support, data-driven models cannot~~ explain the causal relationship between input features and ~~predicted~~ outputs, leading to~~ which results in~~ poor spatial-temporal generalization ~~capability~~ (Gevaert, 2022).

In contrast, ~~Different from data-driven methods,~~ knowledge-driven crop growth models simulate crop development from sowing to harvest based on agronomic mechanisms (Kaur and Singh, 2020). ~~can characterize the entire evolution of crops from sowing to harvesting (Huang et al., 2019).~~ Common model types include light-use efficiency models (e.g., SAFY (Duchemin et al., 2008)), soil-driven models (e.g., AquaCrop (Steduto et al., 2009)), and atmospheric-driven models (e.g., WOFOST (Diepen et al., 1989)). These models integrate ~~were built on the basis of agronomic mechanism knowledge. The growth of crops was simulated by combining~~ environmental factors (e.g., climate conditions and soil characteristics) with crop physiological ~~growth~~ processes ~~(e.g., photosynthesis, respiration and transpiration)~~ (Gaso et al., 2024). Climate variables like temperature, precipitation, and solar radiation are critical in regulating essential physiological processes such as photosynthesis, respiration and transpiration, which influence the rate and duration of crop growth stages (Misaal et al., 2023). Climate anomalies during specific growth stages may disrupt biochemical processes, ultimately affecting yield formation. Similarly, soil properties influence crop productivity by regulating water retention, aeration, and nutrient uptake (Muhuri et al., 2023). ~~According to knowledge-driven types, crop growth models can be further categorized into light-use efficiency models (e.g., SAFY (Duchemin et al., 2008)), soil-driven models (e.g., AquaCrop (Steduto et al., 2009)), atmospheric-driven models (e.g., WOFOST (Diepen et al., 1989)) and so on.~~ Despite their mechanistic rigor, applications of crop models over large area are typically constrained by (1) insufficient spatial-temporal input data, and (2) parameter uncertainty, which can propagate errors into yield estimations ~~However, the application of crop growth models in large area is often constrained by the lack of available model input data which varies with space and time~~ (Dokoohaki et al., 2021). To overcome these challenges, data assimilation techniques to integrate remote sensing observations (e.g., LAI) into ~~The uncertainty of model parameters will further increase the bias in crop yield estimation. Since remote sensing data can provide spatial inputs for crop growth models, data assimilation methods that combining spatial-temporal monitoring of remote sensing data with the simulation of~~ crop growth models have been developed to enhance spatial representativity (Huang et al., 2024). However, ~~the~~ high ~~spatial~~ resolution ~~of~~ remote sensing data drastically~~significantly~~ increases ~~the~~ computational cost, limiting the scalability of these approaches for regional or national mappings efforts ~~of the data assimilation process~~ (Huang et al., 2019). ~~As a result, crop growth models have been hampered in large area applications.~~

Given the limitations~~difficulties~~ above, integrating ~~the strengths of~~ data-driven and knowledge-driven models has emerged as a critical strategy to enhance spatial-temporal generalization and mitigate ~~capabilities of the model and to address~~

sparse training data challengessample issues becomes a critical focus in crop yield estimations. Hybrid frameworksA hybrid method coupling crop growth model with machine learning algorithm, such as those proposed and evaluated by has raising increasing interest (Ren et al., (2023b) and Xie and Huang, (2021), are gaining tractions. These approaches system utilized simulated outputs results from crop growth models (e.g., meteorological, soil, crop physiological, and management factors) as inputs for machine learning, reducing reliance on limited ground observations. Input features of the predicted model include meteorological factors (Isia et al., 2022), soil characteristic factors (Saravi et al., 2020), crop growth factors (Paudel et al., 2021), management factors (Ren et al., 2023) and observation geometry for remote sensing data (Chen et al., 2022). Many studies have demonstrated the ability of hybrid methods are able to enhancein crop yield estimation due to three benefits (Feng et al., 2020; Xie and Huang, 2021; Yang et al., 2021). The simulations from On the one hand, the application of crop growth model can provide biophysical constraints toforfor machine learning, ensuring agronomic plausibility. The crop growth models generate synthetic training datasets to address data scarcity. Finally, the modeling and provide a sufficient simulation dataset for training the machine learning model. On the other hand, the combination of machine learning improves the computational efficiency of yield estimation at the regional scale compared to traditional with data assimilation techniques method (Xie and Huang, 2021). However, exiting studies generally extracted input features (e.g., LAI, and soil moisture) across the extraction of characteristic factors was generally based on the entire growth cycle or on coarse temporal scales, increasing computational costs of model calculation and obscuring stage-specific physiological responsestage of the crop in existing studies (Pinke and Lövei, 2017; Wang et al., 2015). Additionally, while deep learning models, such as This increased the cost of model calculation and might not capture the influence of characteristic factors to a specific stage. Several studies obtained characteristic factors on a monthly scale or based on field observation dates (Everingham et al., 2016; Kern et al., 2018). However, these studies did not consider the specific growth stages of crops. As for the model use, the deep learning models, such as Long Short-Term Memory (LSTM) and GRU model excel at modelling temporal dependencies, their integration into hybrid frameworks have not been widely explored., have a better ability to capture time series information but have not been widely used in hybrid modeling. According to surveys, there is currently no high resolution soybean yield dataset available in the main production regions of China for studying spatiotemporal patterns of soybean production. Therefore, it is urgent to leverage the advantages of hybrid modeling method to create a high resolution soybean yield dataset to further guide agricultural practices.

Critically, the primary soybean-producing regions of China lack a publicly available high-resolution yield dataset to analyse spatiotemporal production patterns, hindering precision agriculture and policy optimization. To address this, we developed a hybrid model coupling the World Food Studies (WOFOST) crop growth model with a GRU deep learning method to estimate soybean yield in Northeast China. The objectives include: (1) Design a hybrid framework integrating WOFOST-simulated growth scenarios with GRU-based temporal feature extraction; (2) Generate a high-resolution (20 m) soybean yield dataset in Northeast China (NortheastChinaSoybeanYield20m) from 2019 to 2023; (3) Evaluate the accuracy of the dataset across field, municipal, and provincial scales using in situ and statistical benchmarks. The WOFOST model first simulated a multi-scenario soybean growth (varying climate, soil, crop varieties and management conditions) to train the GRU model. The time series Sentinel-2 data, capturing soybean growth development, were then input into the GRU model to estimate yield.

This approach prioritizes stage-specific physiological dynamics which balancing computational efficiency and spatial granularity, providing a critical advancement for scalable agricultural monitoring.

~~This study developed a hybrid model coupling data-driven and knowledge-driven models for estimation of soybean yield in Northeast China. The WOFOST model was first adopted to simulate a multi-scenario soybean growth dataset to train the GRU model. The time series Sentinel-2 data of different soybean growth stages were then input into the GRU model to estimate soybean yield. Our study aims to address the following objectives: (1) Designing a hybrid model coupling crop growth model and deep learning model for soybean yield estimation; (2) Generating a high resolution (20 m) soybean yield dataset in Northeast China (NortheastChinaSoybeanYield20m) from 2019 to 2023; (3) Exploring the accuracy of the datasetat multi-scale applications.~~

2 Data preparation and preprocessing

2.1 Study areas

The study was conducted ~~carried out~~ in Northeast China (38°40′ N to 53°34′ N, 115°05′ E to 135°02′ E), encompassing ~~covering~~ Heilongjiang, Jilin, Liaoning province, as well as the eastern parts of the Inner Mongolia Autonomous Region (IMAR) (Fig. 1). The study area includes~~It comprises~~ 40 cities and spans ~~its total area is~~ approximately 1.24 million km^2. The region is characterized by a continental monsoon climate, with an~~. The~~ annual accumulated temperature (≥ 10 ℃) ranginges from 2200 to 3600 ℃ (Pu et al., 2019), and a~~the~~ frost-free period of ~~is between~~ 140 to 170 days (Tan et al., 2014). The average annual precipitation exhibits a strong east-west gradient, decreasinges from 1000 mm in the east ~~(1000 mm)~~ to 350 mm in the west~~west (350 mm)~~ (Zhao et al., 2011). The predominant ~~main~~ soil types ~~in the study area~~ include brown coniferous forest soil, dark brown forest soil, forest steppe chernozem and meadow grassland chernozem soil (Pu et al., 2019). Soybean is one of the three main crops in the region, primarily ~~study area. It is mainly~~ cultivated in the northern parts of the Songliao plain in rotation with ~~rotated with~~ maize. Notably, this region contributes around 64 % of China's total annual soybean production (National Bureau of Statistics of China (NBSC), 2023). Approximately ~~Around~~ 97 % of the soybean in the region is rainfed (Guo et al., 2022; Yu et al., 2020), with growing season typically spanning~~. The growth period generally lasts~~ from May to late September (Zhao et al., 2021). ~~The soybean yield accounts for 64 % of the total yearly soybean yield in China (National Bureau of Statistics of China (NBSC), 2023).~~

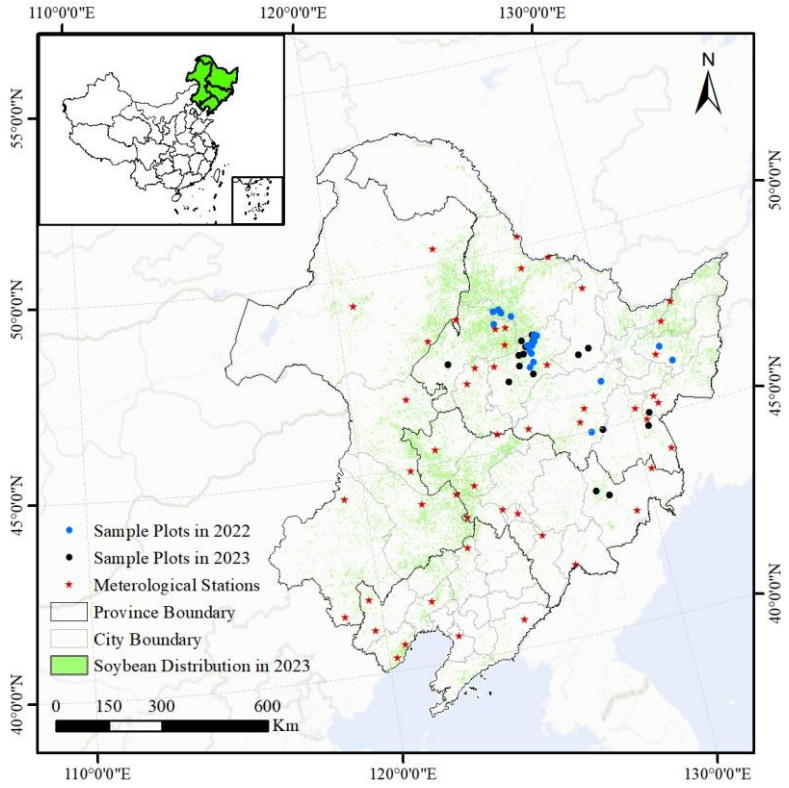


Figure 1: Location of the study area and the distribution of sample plots in two years (2022 and 2023) and selected meteorological stations. **The soybean distribution map was obtained from Zhao et al., (2022) using a moment-preserving segmentation method, achieving an overall accuracy over 90% for soybean in 2023 (Details are provided in Section 2.2.5).**

2.2 Data collections

2.2.1 In-situ measurement data

Due to limitations of resources and personnel, in-situ measurements were not available during the earlier years (from 2019 to 2021). Field-scale yield data was separately collected through field investigation in September 2022 and 2023. In each year, a total of 21 and 18 sample plots were selected, respectively (Fig. 1). Within each sample plot that was around 100 m × 100 m in area, nine quadrats with area of 1 m × 1 m were selected randomly for destructive sampling of yield in soybean. The central location of each quadrat was recorded using a GPS device with accuracy of 1 m. The harvested beans were then oven-dried about 72 hours in Hailun Agricultural Ecology Experimental Station, Chinese Academy of Sciences~~Sihua Agricultural Research Institute, Hailun~~ to determine the yield. Finally, the average yield for the selected nine quadrats represents the soybean yield of the sample plot. In addition, soybean planting dates for different regions were collected through field surveys, providing agro-management data for this study.

Field measured LAI data of soybean was obtained from the Common Aapplication Ssupport Pplatform for Lland Oobservation Ssatellite (CAPLOS, https://124.16.188.131:9699/web/server3/build/#/Guide), an open data portal providing ~~.~~

~~It provides~~ in-situ ~~measured~~ biophysical variables (e.g., LAI and vegetation cover) for validating remote sensing products and refining retrieval algorithms.~~from ground stations for the validation of authenticity and theoretical research on remote sensing product retrieval algorithms.~~ LAI measurements were collected ~~was measured~~ using ~~a~~the LICOR LAI-2200 plant canopy analyzer following a standardized protocol. At each site, the instrument was positioned above the canopy to obtain a reference reading of incoming solar radiation, and then positioned about 5 cm above ~~the~~ ground to collect~~get~~ six readings of radiation transmitting beneath the soybean canopy~~by canopy and was positioned above the canopy to get a reading for incoming radiation~~. The raw data was taken rigorous quality control ~~cleaned~~ to remove outliers, missing or duplicate values. After processing, ~~Finally, a total of~~ 94 LAI observations were retained, spanning~~of LAI for the~~ three soybean growing seasons (2021 – 2023) ~~from 2021 to 2023 were collected~~.

2.2.2 Meteorological data

In this study, two different climate datasets were used.

The meteorological station data used in this study came from the meteorological stations of the National Meteorological Information Center (http://data.cma.cn). There are 238 meteorological stations within the study area. Here 51 of the meteorological stations that located within 1 km buffer zone of the soybean cultivation areas were selected (Fig. 1). The meteorological datasets generally include insolation duration (h), minimum temperature (℃), maximum temperature (℃), daily average temperature (℃), average water vapor pressure (kPa), average wind speed (m sec^{-1}), precipitation (mm) and snow-depth (cm). Observed data from 1980 to 2021 of the 51 selected stations were collected. Missing values and outliers in the data were filtered out. The data were then directly~~They were~~ used for setting~~as~~ input climate parameters of~~for~~ the WOFOST model to drive simulations.

The climate reanalysis data was obtained from the ERA5-land Daily Aggregated - ECMWF Climate Reanalysis Product. ~~The spatiotemporal distribution of meteorological data was obtained from the ERA5 land Daily Aggregated – ECMWF Climate Reanalysis Product.~~ The data was only used to calculate soybean phenology for preparation of yield estimations. It was ~~The ERA5 land is~~ a global climate reanalysis product that provides continuous climate data at a resolution of 0.1° × 0.1° (e.g., air temperature and atmospheric pressure) starting from 1950. The ~~It provides a~~ daily aggregated air temperature data at 2 m above the surface of land measured in kelvin (K) during~~. All the selected data for~~ the soybean growth periods from 2019 to 2023 was~~ere~~ collected in this study ~~obtained~~ from the Google Earth Engine (http://earthengine.google.com). The product was resampled to 20 m using bilinear interpolation model ~~to~~ to match with the resolution of satellite imagery data~~and was then used to calculate the soybean phenology for preperation of yield estimations~~.

2.2.3 Soil data

Soil data was obtained from the 1:1000,000 Chinese soil database, established by the Institute of Soil Science, Chinese Academy of Sciences (Shi et al., 2004) ~~downloading from Geographic Data Sharing Infrastructure, global resources data cloud (www.gis5g.com)~~. The dataset consisted of two parts: soil spatial data (digital soil maps) and soil attribute data. In this study,

the 1:1000,000 soil spatial data was obtained. The spatial database was developed by digitizing, mosaicking, and reassembling sheets from the 1:1,000,000 Soil Map of the People's Republic of China (National Soil Survey Office, 1995), with the Genetic Soil Classification of China (GSCC) soil families as the fundamental mapping units. The final dataset includes 909 soil types and over 94,000 polygons. ~~The dataset coversthe distribution of various soil types and their main chemical characteristics across the country.~~ The dataset was utilized to determine the dominant soil types within the study area, serving as the basis for assigning soil parameter settings according to literatures.

2.2.4 Satellite imagery data

Two satellite data includinge: 1) Sentinel-2 Multi-Spectral Instrument (MSI) ~~images (~~Level ~~-~~ ~~L~~2A Surface reflectance product (10 – 60 m spatial resolution, 5-day revisit), and 2) the Moderate Resolution Imaging Spectroradiometer (MODIS) Leaf Area Index (LAI) / Fraction of Photosynthetically Active Radiation (FPAR) Level 4 product (MCD15A3H, v061, 500 m spatial resolution, 4-day period)~~Product (MCD15A3H.061)~~ were used ~~in this stud~~to generate yield maps.~~y.~~ All ~~of the~~ data spanning ~~collected in the four~~ soybean growth periods (2019 – 2023) were accessed and pre-processed via the~~(2019-2023) were downloaded from~~ Google Earth Engine (GEE, http://earthengine.google.com).

The MSI aboard Sentinel-2A/B satellites provides 10 m (visible and near-infrared bands), 20 m (red-edge and shortwave infrared bands) and 60 m (atmospheric bands) bands at 5-day revisit. The Level-2A data, which are geometrically and atmospherically corrected via the Sen2Cor, were masked for clouds and shadows using the Quality Assurance (QA) band. The 60 m band was excluded due to their low spatial resolution and limited relevance for yield estimation and the 10 m (B2: Blue, B3: Green, B4: Red, B8: Near-Infrared) and 20 m (B5–B7: Red-edge, B8A: Near-Infrared, B11–B12: Shortwave Infrared) bands were retained. To harmonize spatial resolution, the 10 m bands were resampled to 20 m using bilinear interpolation model.

The MODIS MCD15A3H (Collection 6.1, Level 4) provides 4-day composite LAI and FAPAR at 500 m derived from Terra and Aqua satellite sensors LAI/FAPAR are primarily inverted via a 3D radiative transfer model-based look-up-table (LUT) algorithm (Knyazikhin et al., 2018). When the primary algorithm fails, they are estimated using an empirical NDVI-LAI model. The LAI data was similarly reprojected to WGS -84 to ensure spatial alignment with Sentinel-2 imagery. These coarse-resolution LAI data were used to generate 500 m yield maps. The coarse-resolution yield maps were then used to bias-correct the 20 m Sentinel-2 yield maps, improving their regional consistency. Details about the bias correction are present in following 3.3.2 Section. ~~The Multi-Spectral Instrument (MSI) on the Sentinel-2 satellites (including Sentinel-2A and Sentinel-2B) provides high spatial resolution imagery with resolutions of 10-m, 20-m and 60-m, and a temporal resolution of 5-days. Only cloud-free images were selected. The band with 60-m was excluded in the analysis, and the bands of 10 and 20-m were used. To match the red-edge bands, the four 10-m bands including B2 / B3 / B4 / B8 was resampled to 20-m using bilinear interpolation model.~~

~~The MCD15A3H (Collection 6.1, Level 4) provides 500-m LAI and FAPAR product with 4-day composite. The LAI product was chosen in this study. The optimal pixel from data of MODIS sensors on NASA's Terra and Aqua satellites within~~

~~the 4 days is selected for inversion of LAI. The retrieval algorithm is based on a look-up-table (LUT) derived from a 3D radiative transfer model (Knyazikhin et al., 2018). It estimates LAI by using the ratio of red to near infrared reflectance. An empirical model based on relationships between LAI and NDVI is served as an alternative choice when the main algorithm fails. As the product provides a higher temporal resolution than Sentinel-2 satellites, yield estimations were also done based on LAI product to better characterize the rapid changes in soybean growth stage. The MODIS yield maps were then applied for bias correction of Sentinel-2 yield maps.~~

2.2.5 Crop distribution data

The soybean distribution maps for the study area (2019 – 2023) ~~from 2019 to 2023~~ were obtained from Zhao et al., (2022), which employed a novel methodology for crop type identification. The study proposed an~~the~~ optimal identification feature (OIF) knowledge graph coupled~~ing~~ with a moment-preserving segmentation method to classify~~for~~ crop types without ground-truth data~~identification through the crop growth season~~. The method achieved ~~The~~ overall accuracy above 90% and ~~the~~ producer's accuracy exceeding 93% for maize, soybean and rice, ~~was higher than 90 % and 93 %, respectively~~with a Kappa coefficient greater than 0.90.

2.2.6 Statistical data

Crop yield records (1980-2022) were obtained from the Statistical Yearbooks published by the Statistic Bureau of Heilongjiang (http://tjj.hlj.gov.cn), Jilin (http://tjj.jl.gov.cn), Liaoning (https://tjj.ln.gov.cn) and Inner Mongolia Autonomous Region (https://tj.nmg.gov.cn) to validate the crop yield estimates. Because the 2022 Statistical Yearbook was not fully released, yield records for that year cover only a subset of cities. The statistical data served two main purposes, model simulation validation and regional-scale accuracy evaluation in this study. To ensure the multi-scenario soybean growth dataset capture the full range of production conditions that across multi-years meteorological data, various soil types, multiple soybean varieties and different agro-managements, the yield records from 1980 to 2022 along with published yield data and field samples were used to assess the reasonableness of simulated yields. For the spatial validation, regionally aggregated statistical yield data (2019 – 2022) were applied to evaluate the accuracy of the hybrid framework at municipal and provincial scales.

~~The crop yield records (1980-2022) of the Statistical Yearbook provided by Statistic Bureau of Heilongjiang (http://tjj.hlj.gov.cn), Jilin (http://tjj.jl.gov.cn), Liaoning (https://tjj.ln.gov.cn) as well as IMAR (https://tj.nmg.gov.cn) were collected for the validation of crop yield estimation. For the statistical yield in 2022, the data covered a limited number of cities as the Statistical Yearbook in 2022 has not been fully released. The data from 1980-2022 were all used to validate the reasonability of model simulations for construction of soybean growth dataset. For further yield estimates at regional scale, data from 2019 to 2022 were applied for accuracy evaluation.~~

3 Methodology

Our proposed hybrid model utilizes both the advantages of machine learning in data mining and the mechanism advantages of crop growth model. Figure 2 presents the flowchart of the hybrid methodology for soybean yield estimation. It mainly includes 1) Generating a training dataset based on the WOFOST model that simulate multi-scenario soybean growth and yields under various climates, soil, cultivars and agro-management practices, 2) Training a GRU model to identify the relationships between simulated LAI and yield, 3) Producing soybean yield maps under ~~Constructing multi-scenario soybean growth dataset, 2) Modeling for relationships between soybean yield and inputs including meteorological data, soil types, crop varieties and agro-managements, and 3) Estimating soybean yield in~~ multi-scale using~~from~~ LAI derived from MODIS and Sentinel-2 remote sensing data.

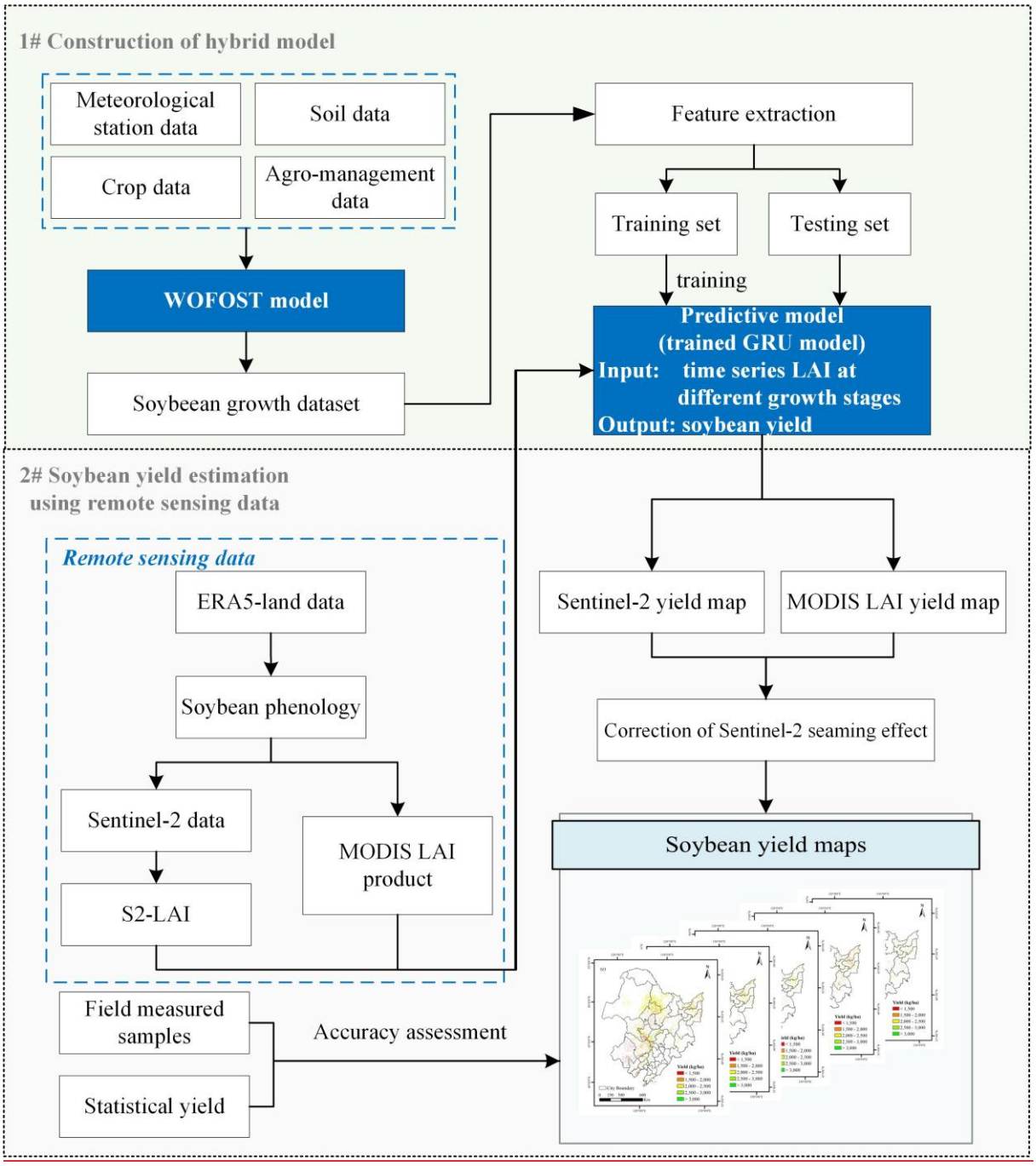


Figure 2: The flowchart of the overall yield estimation methodology in this study.

3.1 Construction of multi-scenario soybean growth dataset

The soybean growth dataset used in this study was a multi-scenario, knowledge-based dataset derived from crop growth model simulations. The dataset provided a quantitative description of yield formation by simulating soybean development under diverse agricultural production scenarios including variations in meteorology (temperature, precipitation, solar radiation), soil types (texture, organic matter), crop varieties (phenology, thermal requirements) and agro-managements (sowing date).

To generate the dataset, we employed the World Food Studies Simulation Model (WOFOST) (Diepen et al., 1989), implemented via the Python Crop Simulation Environment framework (PCSE, v5.5). The WOFOST model is well-suited for large-scale simulations and has been extensively validated (Huang et al., 2015). Given that soybean cultivation in the study region is predominantly rainfed, we adopted the water-limited mode (Wofost72_WLP_CWB) for simulations. Driven by daily weather data, soil profiles, and cultivar parameters, the water-limited model simulated CO2-driven photosynthesis, water deficits, and biomass partitioning, outputting daily LAI, daily biomass accumulation, and final grain yield. Crop development is modelled through development stages (DVS) : 0 for emergence, 1 for anthesis and 2 for maturity (Diepen et al., 1989). This dataset mechanistically captures yield-limiting processes (e.g., drought during critical growth phases) while enabling scalable scenario analysis across Northeast China's rainfed soybean systems.

~~Soybean growth dataset was a multi-scenario dataset based on simulations of crop growth model. It was a quantitative description of the formation process of soybean yield, considering the association between soybean yield and growth state as well as environmental conditions. It was constructed by comprehensively and systematically simulating the crop growth under various agricultural production scenarios such as different meteorological conditions, soil types, crop varieties and agro-managements.~~

~~In this study, the World Food Studies Simulation Model (WOFOST) model (Diepen et al., 1989) was employed to generate knowledge based soybean growth dataset. Detailed information about the model could be found in Huang et al., (2015). The Python Crop Simulation Environment (PCSE) framework version 5.5 was used in this study. The generic process description of the model was well suited for conducting large-scale simulations (Zhuo et al., 2023). Based on the theory of CO_2, the WOFOST was developed to simulate daily changes of biomass and final grain yield for specific crop types. The crop growth process was regulated by crop development stages (DVS) in WOFOST with 0 indicating emergence, 1 indicating anthesis and 2 indicating maturity (Diepen et al., 1989). Through the crop growth simulation, the model can calculate and predict the crop yield.~~

3.1.1 Preparation of model input parameters

In this study, parameters of WOFOST were not calibrated or optimized using in-situ data, as our goal was to generate a scenario-based dataset representing diverse agricultural conditions rather than global optimal value. ~~The WOFOST model employs a range of parameters including meteorology, soil, crop and agro management for crop simulation (Diepen et al., 1989). None of advanced optimization was utilized in this study, since our objective is not to determine the globally optimal~~

value of model parameters. The model parameters were sourced from set to accommodate various agricultural production scenarios during the soybean growth period without calibration with in-situ data. The values of the parameters were determined from four sources: ground observation, online database, peer-reviewed literatures and default WOFSOT valuesthe initial values provided in WOFOST.

(1) Meteorological parameters

The meteorological parameters required in WOFOST is shown in Table 1. To capture regional climate variability (e.g., temperature extremes, rainfall patterns), account for various meteorological data ofconditions, the meteorological data collected from the selected 51 meteorological stations spanningover a period of 42 years (1980-2021) were compiled. These data – including daily temperature, precipitation, and solar radiation – were preprocessed into the model's required input format (e.g., daily time steps, unit conversions) to ensure compatibility. was all utilized to provide values of these parameters. The meteorological inputs were then processed into the format recognized by the model.

Table 1 Meteorological parameters required in WOFOST.

Parameter	Description	Units
IRRAD	Incoming global shortware radiation	KJ m^{-2} d^{-1}
TMIN	Daily minimum temperature	℃
TMAX	Daily maximum temperature	℃
VAP	Daily mean vapour pressure	kPa
WIND	Daily mean windspeed at 2 m above the surface	m s^{-1}
RAIN	Daily rainfall	mm
SNOWDEPTH	Snow depth	cm

(2) Soil parameters

The soil parameters in the WOFOST mainly include soil moisture content at wilting point (SMW), field capacity (SMFCF) and saturation (SM0) as well as hydraulic conductivity of saturated soil (K0). Based on the 1:1,000,000 Chinese soil database, the study area predominantly comprises loam soil that is further classified into sandy, light, medium and heavy loam. The parameters for sandy, loam and medium loam were sourced from Du et al., (2025), while the parameters for heavy loam came from Sun et al., (2022). All soil parameter values, summarized in Table 2, were integrated into the model to evaluate the influence of soil variability on soybean yield (Du et al., 2025; Sun et al., 2022).

In our study, they were acquired from the 1:1000,000 Chinese soil database. The soil texture in study area is predominantly loam soil. The loam soil is further divided into sandy loam, light loam, medium loam and heavy loam. The value settings of soil parameters for different soil types in the study area was presented in Table 2.

Table 2 Values of main soil parameters in WOFOST.

Soil type	SMW	SMFCF	SM0	K0

	($cm^3\ cm^{-3}$)	($cm^3\ cm^{-3}$)	($cm^3\ cm^{-3}$)	($cm\ d^{-1}$)
Sandy loam	0.060	0.280	0.350	22.6
Light loam	0.090	0.280	0.340	19.3
Medium loam	0.110	0.280	0.340	18.1
Heavy loam	0.194	0.355	0.356	34.6

(3) Crop-specific parameters

~~In this study, five different soybean varieties were considered in the study area to enhance the diversity of cultivars in the simulation, named early maturity, medium-early maturity, intermediate maturity, medium-late maturity and late maturity. They were designed for planting in the five thermal zones in Heilongjiang Province (Qu et al., 2023) (Table 3). To determine the distribution of soybean varieties in the study area, the ten thermal zones divided in Northeast China by Wang et al., (2022) were acquired. The thermal zones were calculated with a temperature difference of 200 °C·d based on the thermal zones in Heilongjiang Province using historical meteorological data. The soybean varieties suitable for various thermal zones was then determined based on the demand of accumulated temperature (Table 3).~~

~~Table 3 The division standard of the thermal zones and soybean varieties in Northeast China.~~

~~Thermal zones~~	~~Annual accumulated temperature ≥ 10 °C (°C·d)~~	~~Corresponding thermal zones in Heilongjiang Province~~	~~Soybean varieties~~
~~1st thermal zone~~	~~≥ 3500~~		
~~2nd thermal zone~~	~~3300 – 3500~~		
~~3rd thermal zone~~	~~3100 – 3300~~	~~1st thermal zone~~	~~late maturity~~
~~4th thermal zone~~	~~2900 – 3100~~		
~~5th thermal zone~~	~~2700 – 2900~~		
~~6th thermal zone~~	~~2500 – 2700~~	~~2nd thermal zone~~	~~medium-late maturity~~
~~7th thermal zone~~	~~2300 – 2500~~	~~3rd thermal zone~~	~~intermediate maturity~~
~~8th thermal zone~~	~~2100 – 2300~~	~~4th thermal zone~~	~~medium-early maturity~~
~~9th thermal zone~~	~~1900 – 2100~~	~~5th thermal zone~~	~~early maturity~~
~~10th thermal zone~~	~~≤ 1900~~		

In this study, the soybeans were classified into five types including early, medium-early, intermediate, medium-late and late maturity according to Qu et al., (2023). In the WOFOST model, soybean phenology is governed by temperature-driven parameters: the minimum (TBASEM) and maximum (TEFFMX) threshold temperature for emergence, and accumulated thermal time (TSUMEM: sowing to emergence; TSUM1: emergence to anthesis; TSUM2:anthesis to maturity). These thermal parameters are cultivar-sensitive and were set based on historical meteorological data and field phenology records, validated against field observations (Qu et al., 2023). Remaining crop parameters (e.g., SLATB: specific leaf area) were assigned default values or optimal values from Sun et al., (2022). Full parameter specifications are provided in Table A1.

In the WOFOST model, soybean growth stages are mainly determined by temperature-related parameters including the minimum and maximum threshold temperature for emergence (TBASEM, TEFFMX, respectively), accumulated temperature (T_e) from sowing to emergence (TSUMEM), from emergence to anthesis (TSUM1) and from anthesis to maturity (TSUM2). The accumulated temperature for different growth stage is sensitive to crop varieties according to the study of Qu et al., (2023). The values of main crop parameters for different soybean varieties were shown in Table 4. They were set according to the historical meteorological data and observation data of soybean and had been validated using actual development periods (Qu et al., 2023). Other crop parameters used the default values of the WOFOST model or the optimal values from the study of Sun et al., (2022).

Table 4 Values of main crop parameters in WOFOST.

Crop type	TBASEM (°C)	TEFFMX (°C)	TSUMEM (°C d)	TSUM1 (°C d)	TSUM2 (°C d)
Early maturity	8	22	70	450	660
Medium-early maturity	8	22	70	480	770
Intermediate maturity	8	22	70	520	870
Medium-late maturity	8	22	70	540	960
Late maturity	8	22	70	580	1000

(4) Agro-management parameters

Planting date is the major agro-management factors for soybean in the study area. The difference of planting date can significantly impact on soybean growth development, pod count, and biomass accumulation (Urda et al., 2024). Soybean growing in the study area is mainly rainfed, and is received fertilizer management practices from the local government. Therefore, the water-limited mode of the WOFOST was employed for soybean simulation. The water-limited model mainly requires the planting date for starting growth simulation. Soybeans in study area is typically sown between late April and late May and are seldom subjected to nutrient stress. FFour planting dates 20 April, 30 April, 10 May, and 20 May to reflect the typical sowing window (late April to late May) of the study area were set for model simulation according to Mei et al., (2024).

to represent different agro-management scenarios, were set for model simulation.

3.1.2 Multi-scenarios crop simulations

Following parameter preparation, the four parameter categories, including meteorological (51 stations × 42 years), soil (4 types), crop-specific (5 varieties) and agro-management (4 planting dates), were systematically combined to create 171,360 unique scenarios (Table 3). These scenarios were executed in the WOFOST simulations, yielding a dataset of 171,360 various simulations that quantify yield responses to diverse agricultural production conditions.After parameter preparation, a soybean growth dataset was constructed through model simulations which accounted for the multi-scenarios in agricultural production. The four different types of model parameters were arranged and combined to generate various simulation scenarios. The

scenarios were then put into the model for simulation. Finally, a dataset containing more than 8,000 available simulations were generated.

Table 3 Scenarios for WOFOST simulations

Parameters	Number of categories	Details
Meteorological parameters	51×42	Meteorological data from 51 stations over 42 years (1980 – 2021)
Soil parameters	4	Sandy loam, light loam, medium loam and heavy loam
Crop-specific parameters	5	Early maturity, medium-early maturity, intermediate maturity, medium-late maturity and late maturity
Agro-management parameters	4	Four planting dates 20 April, 30 April, 10 May, and 20 May

3.2 Development of the Grated Recurrent Unit model (GRU)

A GRU (Grated Recurrent Unit) model, a streamlined variant of recurrent neural networks (RNNs), was employed to be trained using the multi- scenarios simulated dataset for large-scale soybean yield estimation. Unlike LSTM (Long short-term memory), GRU simplifies gating mechanisms to two adaptive gates, update and reset gates (Cho et al., 2014). The update gate retains the past information for future calculations. The reset gate aims to remove irrelevant historical context for simplifying the new candidate hidden states. Using the two gates together is benefit to balance long-term dependency capture and computational efficiency (Peng and Yili, 2022; Zhang et al., 2022). This design mitigates vanishing gradient issues while accelerating model training, making GRU particularly effective for time-series yield estimation (Gopi and Karthikeyan, 2023; Ren et al., 2023b).The GRU was trained to estimate soybean yields in this study. The structure of a GRU cell is shown in Fig. 3. GRU controls the flow of information through update and reset gates (Cho et al., 2014). The update gate aims to control how much of the past information that are retrained and will be used in the future calculation. The reset gate aims to evaluate whether the remained previous information can be ignored in the new candidate hidden state. The use of two gates maintains the balance between retaining the hold hidden state and incorporating new information (Peng and Yili, 2022; Zhang et al., 2022). This improves the training speed of the model and helps mitigate the vanishing gradient problem during training. Since GRU can effectively capture long-term dependencies in time series data, it has achieved good performance in applications of crop yield estimation (Gopi and Karthikeyan, 2023; Ren et al., 2023b). The computation of a GRU unit can be summarized by the following equations:

$$R_t = \sigma(W_r \cdot [X_t, H_{t-1}] + b_r) \quad (1)$$

~~$Z_t = \sigma(W_z \cdot [X_t, H_{t-1}] + b_z)$ (2)~~

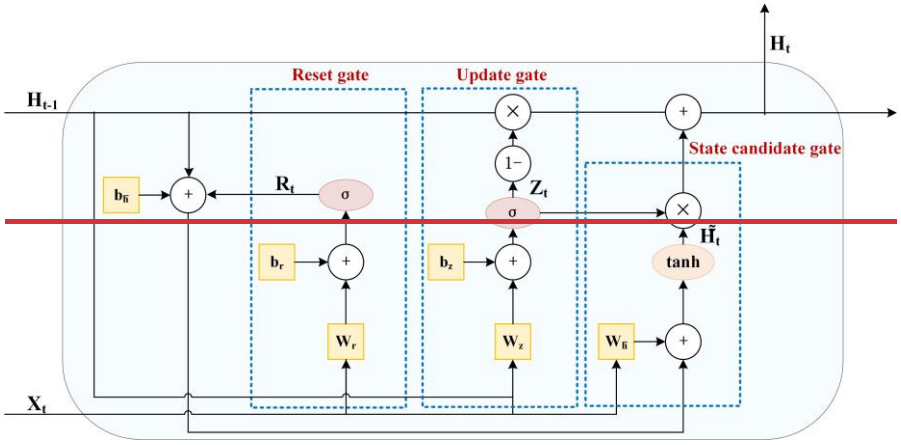
~~$\widetilde{H_t} = \tanh(W_{\tilde{h}} \cdot X_t + W_{\tilde{h}} \cdot (R_t \odot H_{t-1}) + b_{\tilde{h}})$ (3)~~

~~$H_t = (1 - Z_t) \odot H_{t-1} + Z_t \odot \widetilde{H_t}$ (4)~~

~~where R_t and Z_t represents the activation vector of reset and update gates, respectively; $\widetilde{H_t}$ represents the potential update vector; H_t and X_t represent the hidden state output and the input at time t; σ and tanh are the sigmoid function and the hyperbolic tangent function; W_r, W_z and $W_{\tilde{h}}$ represent the reset gate weight, the update gate weight and the update candidate weight, respectively; b is the bias vector of the parameters.~~

~~The GRU layer is connected with a fully connected layer. The output of the network at time t (Y_t) is finally determined by multiplying the hidden states of all cells in the GRU layer by the weights of the fully connected layer:~~

~~$Y_t = W_y \cdot H_t + b_y$ (5)~~



~~Figure 3: Structure of a GRU cell.~~

Trained on the multi-scenarios simulated dataset, the GRU constructed based on TensorFlow 2.6 linked simulated environmental inputs to yield outputs. Accounting for the computational efficiency of the model in large areas, two key features include LAI_{mean1} (mean LAI during vegetative growth: emergence to flowering) and LAI_{mean2} (mean LAI during reproductive growth: flowering to maturity), were calculated to reflect photosynthetic capacity and yield potential. These two LAI metrics served as inputs, while simulated yields acted as outputs. The multi-scenarios simulated dataset was partitioned using 10-fold cross-validation, with hyperparameters (e.g. learning rate and batch size) optimized using a grid search to achieve minimal root mean squared error (RMSE, Eq. (5)) (Açikkar, 2024).

Once trained, the GRU model taken Sentinel-2-derived LAI time series as inputs to generate 20 m yield maps.

~~In this study, the GRU model was constructed based on TensorFlow 2.6. LAI was selected as the input feature of the model. As a crucial state variable in WOFOST, LAI signifies the photosynthetic capability of crops and can effectively characterize the potential yield (Huang et al., 2015). Accounting for the computational efficiency of the model in large areas, the average value of LAI was used as the input feature of GRU. To better capture the growth dynamics of soybean, the mean LAI at vegetative (from emergence to flowering) and reproductive (from flowering to maturity) growth period in the soybean~~

growth dataset were calculated, represented as LAI_{mean1} and LAI_{mean2}. The two LAI values derived from the simulations, serving as temporal input features, were then combined with model simulated soybean yield in knowledge base to create the simulated dataset for the GRU model. The simulated dataset was splits into training and testing datasets using 10-fold cross-validation. The hyperparameters of GRU was optimized using a grid search method (Açikkar, 2024). The root mean squared error (RMSE, Eq. (10)) was applied to assess the predictive performance of different set of hyperparameters. After optimization of each fold, the hyperparameters that yielded the smallest predictive error were selected as the optimal ones.

3.3 Generation of NortheastChinaSoybeanYield20m

3.3.1 Determination of soybean phenology

SDue to the spatial variability of soybean phenology of the study area exhibits significant spatial variability due to climatic and varietal differencesat the regional scale (Gaso et al., 2024). To address this, soybean phenology maps were generated from daily thermal time by integrating thermal zone divisions and regional adapted cultivars (Fig. A1). Soybean phenology (including emergence, anthesis and maturity) were calculated using daily aggregated air temperature data from ERA5-land dataset and a thermal time model (T_e) (Eq. (1)):, it is necessary to extract phenological stages of soybean before yield estimation. Based on the soybean varieties suitable for different thermal zones (Table 3) and the division of thermal zones in the study area, spatial distribution of soybean planting types in the study area was produced (Fig. A1). Daily aggregated air temperature data ERA5-land was applied for calculation of soybean phenology for different variaties regionally based on the T_e (Eq. (6)):

$$T_e = \begin{cases} 0, (T_{mean} \le T_{base}) \\ T_{mean} - T_{base}, (T_{base} < T_{mean} < T_{max}) \\ T_{max} - T_{base}, (T_{mean} \ge T_{max}) \end{cases} \quad (16)$$

where T_{mean} represents the is daily mean temperature, T_{base} (8 °C) and T_{max} (37 °C) represent the minimum and the maximum temperature for soybean development, respectively (Allen et al., 1997; Choi et al., 2016). Soybean growth proceeds from a growth stage to the next stage when accumulated T_e reached the threshold of accumulated temperature required for growth according to the setting of crop parameters of WOFOST model (Table A1). In this study, T_{base} was set to 8 °C and T_{max} was set to 37 °C (Allen et al., 1997; Choi et al., 2016).

Based on field surveys and literatures, the planting dates of soybean were fixed uniformly set as 5 May for Heilongjiang Province and Inner Mongolia Autonomous Region, and 1 May for Jilin and Liaoning Province (Huang and Liu, 2024), with emergence constrained to before. The emergence date of soybean in Northeast China should not exceed 1 June at the latest (Mei et al., 2024), and maturity to . In addition, soybeans generally mature before 1 October (Huang and Liu, 2024). Under the constraints of T_e (Table A1Table 2) and the agro-management, phenological dates (emergence, anthesis and maturity) were computed for we calculated the emergence, anthesis and maturation time of soybeans in each Sentinel-2 pixel (2019 – 2023)of planting areas from 2019 to 2023. Finally, the phenological maps were Based on changes in agricultural phenology and the revisit cycle of remote sensing satellites, the phenological calculation results for each year were further clustered into 10 phenology classes using K-means clustering method (Jain and Dubes, 1988), aligning with Sentinel-2's revisit cycle to

optimize imagery selection for yield estimation. ~~Remote sensing imagery were then obtained for different classes accordingly for yield estimation.~~

3.3.2 Model estimations of soybean yield

The ~~The estimation of soybean yield was mainly based on Sentinel-2 data. r~~Red-edge normalized difference vegetation index ($NDVI_{RE}$) (Gitelson and Merzlyak, 1994) was employed for LAI mapping~~utilized to derived the model input feature, LAI~~ (Eq. (2~~7~~)).

$$NDVI_{RE} = \frac{B8A NIR - B5 RE}{B8A NIR + B5 RE} \quad (2\text{~~7~~})$$

where B8A (near-infrared) and B5 (red-edge) are Sentinel-2 bands.~~NIR and RE represent the B8A and B5 band of Sentinel-2 data, respectively.~~

The ~~LAI of~~ soybean LAI was estimated from~~characterized by a linear fit with~~ $NDVI_{RE}$ using a linear regression (Eq. (3~~8~~)) validated across multiple crops (R^2 = 0.732, RMSE = 0.69) (Pasqualotto et al., 2019). ~~The linear fitting has been validated using a multi-crop dataset with R^2 of 0.732 and RMSE of 0.69 (Pasqualotto et al., 2019) which showed a unified potential for LAI estimation in various crops.~~

$$LAI = 5.405 \cdot NDVI_{RE} - 0.114 \quad (3\text{~~8~~})$$

Mean~~The average value of~~ LAI values for vegetative (LAI_{mean1}) and reproductive (LAI_{mean2}) growth stages were computed from the time-series S2-derived LAI. These LAI values were ~~for different growth stages of soybean (LAI_{mean1} and LAI_{mean2}) was calculated. The two LAI was then~~ input into the GRU model for yield prediction. For pixels with missing values during these stages, ~~which had no observation either in vegetation or reproductive stage,~~ LAI values were~~was~~ replaced by the average of ~~surrounding~~ eight neighbouring pixels ~~instead~~. Final~~Finally, the yield maps with~~ 20 m yield maps ~~spatial resolution~~ were marked using ~~by~~ soybean distribution maps to exclude non-soybean areas.

For large area estimations, a total of 194 Sentinel-2 tiles were required to fully cover the study area. Affected by cloud cover, the frequency of available data varied across each tile. Therefore, the yield maps often exhibited discontinuities along the edges of different tiles ("seaming effects"). This seaming effect could obscure real yield variations. To address this issue,~~minimize the Sentinel-2 seaming effect,~~ a bias correction method proposed by ~~was employed following~~ Azzari et al., (2017) was applied. The overall framework is to use yield estimation based on MODIS LAI to correct the yield estimation from Sentinel-2. MCD15A3H generally provided more continuous estimation results of LAI due to its higher temporal resolution (4-day composites) and broader coverage. Yield maps were generated from the trained GRU taking MCD15A3LAI products as inputs. Sentinel-2 yield maps were adjusted by adding the difference between MODIS-derived mean yield and initial Sentinel-2 mean yield for each tile. This process minimized seams while preserving fine-scale yield variability within tiles.~~The main idea of the bias correction approach should be present in here. In this framework, MODIS imagery was used for intercalibration. Due to its higher temporal resolution and broader image coverage, MODIS generally provided more continuous estimation results. For correction, yield maps were also generated using MODIS LAI products. We utilized the~~

~~estimation results from MODIS to calibrate the mean yield for each Sentinel-2 tile. The difference between the mean value of the yield derived from MODIS for the region cover the tile and the initial Sentinel-2 estimations was then added to the Sentinel-2 values. Through this correction, the seams within yield maps were alleviated while the sub-tile variation of yields were preserved.~~

3.4 Accuracy evaluation

The accuracy of generated NortheastChinaSoybeanYield20m (2019-2023) was evaluated on multiple scales. For field scale, in-situ yield data in 2022 and 2023 was used for assessment. For regional scale, the mean soybean yield for each city and province were separately calculated for each year, and compared with the statistical data. Accuracy evaluation was based on the coefficient of determination (R^2, Eq. (~~4~~9)), the root mean squared error (RMSE, Eq. (~~5~~10)) and mean relative error (MRE, Eq. (~~6~~11)).

$$R^2 = 1 - \frac{\sum_i (y_{o,i} - y_{m,i})^2}{\sum_i (y_{o,i} - \overline{y_o})^2} \quad (~~4~~9)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (y_{o,i} - y_{m,i})^2}{n}} \quad (~~5~~10)$$

$$MRE = \frac{\sum_{i=1}^{n} |y_{o,i} - y_{m,i}|}{n \cdot y_{o,i}} \quad (~~6~~11)$$

where $y_{o,i}$ and $y_{m,i}$ represent the actual yield (observed or statistical yield) and model estimated yield, respectively, $\overline{y_0}$ is the mean value of the actual yield.

4 Results and analysis

4.1 Simulations of the WOFOST model

Since LAI was used as~~As~~ the input feature, the accuracy of WOFSOT-simulated LAI ~~simulated by WOFOST~~ directly influenced the reliability ~~impacted the predictive capability~~ of the GRU model for soybean yield prediction. To validate simulated LAI, f~~F~~ield-measured LAI from 2021 to 2023 were compared against mean LAI curve calculated from value of ~~was used to validate the reasonability of model simulated LAI.~~ 5,000 simulated LAI curves ~~were~~ randomly selected from the multi-scenarios simulated dataset~~and the mean value was calculated for comparison with ground observation data~~ (Fig. ~~3~~4). The results showed that simulated LAI trends aligned closely with observed field variations, capturing~~indicated that the changes of simulated LAI were generally aligned with the observed variations from field measurements. The range of simulated results encompassed more than~~ 88 % of the field-measured sample sites (n = 83) within the simulated range.~~,~~ This demonstrated robust agreement between model outputs and ground truth, confirming the WOFOST simulations' ability to represent realistic LAI dynamics for GRU training.~~indicating a high level of confidence.~~

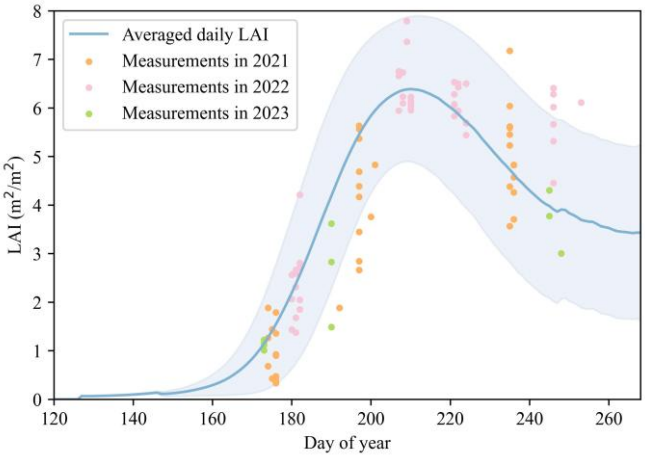


Figure 34: Comparison of averaged daily LAI randomly selected from model simulations (n = 5000) with field-measured LAI in 2021 (n = 38), 2022 (n = 46) and 2023 (n = 10). The gray shading represents one standard deviation, indicating the uncertainty in LAI simulation.

Figure 4 (a)~~5~~ displays the histogram distribution of simulated soybean yields, revealing an approximately normal distribution (mean = 2675.66 kg ha^{-1}). The ~~result indicates that the~~ multi-scenario soybean growth dataset ~~developed in this study~~ effectively captured wide range of production conditions, spanning ~~various production scenarios of soybeans, encompassing both~~ low to ~~and~~ high yield extremes ~~simulations~~. Fig. 4 (b) shows a box plot for comparing the simulated yield with historical~~The simulations of soybean yield followed a normal distribution. The mean value of simulated soybean yield is 2675.66 kg ha^{-1}. The box plot illustrates the distribution of simulated data alongside~~ statistics from 1980 to 2022, published yield data from literatures, and field measurements from~~in~~ 2022 and 2023. T~~Compared with other yield data, t~~he simulated dataset exhibited the widest value range, demonstrating the comprehensiveness~~for this study had the widest range of values which reflected the credibility~~ of the multi-scenario ~~in the~~ knowledge base and the robustness~~effectiveness~~ of the simulation outcomes~~results~~.

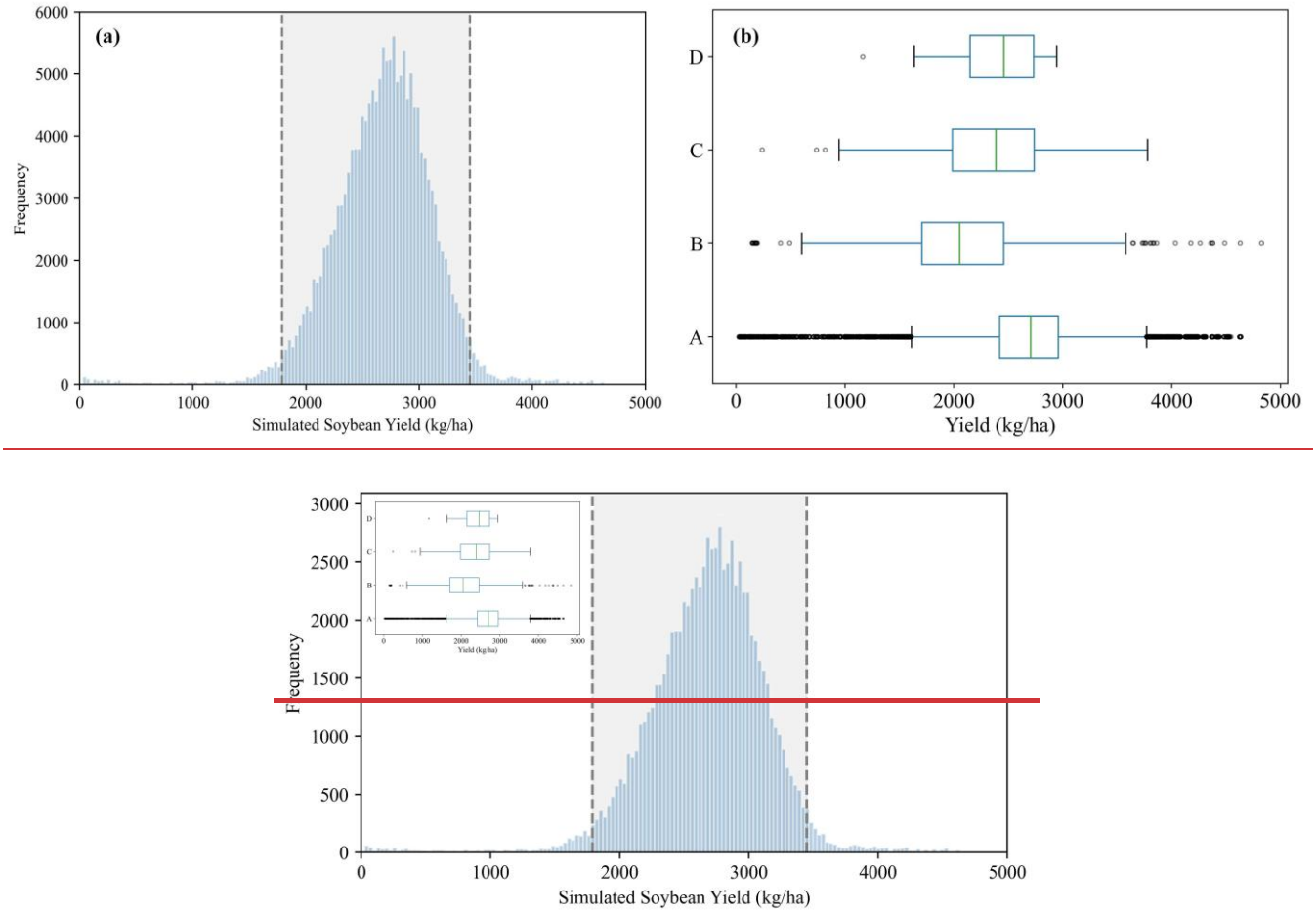


Figure 45: (a) Histogram statistics of simulated soybean yield where t~~The gray area in the histogram represents 95 % confidence intervals; .(b) d~~Distribution of simulated soybean yield compared with other datasets where . ~~The gray area in the histogram represents 95 % confidence intervals.~~ A represents simulated yield in this study (n = 171,360~~83,972~~), B represents statistical yield from 1980 to 2022 (n = 961), C represents specific measurements from the literature (Chen et al., 2011; Fan et al., 2012; Liu et al., 2005, 2008; Liu and Herbert, 2002; Wang et al., 2020, 2024; Zheng and Zhang, 2021) (n = 138) and D represents measurements in 2022 and 2023 carried by this study (n = 39).

4.2 Yield estimation at field scale

The field-scale performance of NortheastChinaSoybeanYield20m ~~at field scale~~ was validated against in-situ~~using field~~ measurement from 2022 and 2023 ~~(Fig. 6)~~, demonstrating strong accuracy in capturing spatial yield variability (Fig. 5). The estimated yields showed strong agreement with observed yield, with~~comparison between estimated and observed yield showed a great equality with~~ $R^2 > 0.65$ ($p < 0.01$) ~~in both of the two years~~. The error-bars indicated more consistent performance in fields with uniform yields, while higher uncertainties appear in fields with larger estimation deviations. Overall accuracy across both years reached~~The overall accuracy in estimations for both of the two years was~~ 0.73 in R^2 ($p < 0.01$), 287.44 kg ha^{-1} in RMSE and 10.02 % in MRE (Fig. A2). Notably, higher ~~and the model achieved higher yield estimation~~ accuracy in 2023 with

RMSE of 271.07 kg ha^{-1} and MRE of 8.57 % (Fig. 56b) was achieved. The results indicated that the dataset well captured the spatial variation of soybean yield.~~This assessment indicated that the dataset well captured the spatial variation of soybean yield.~~

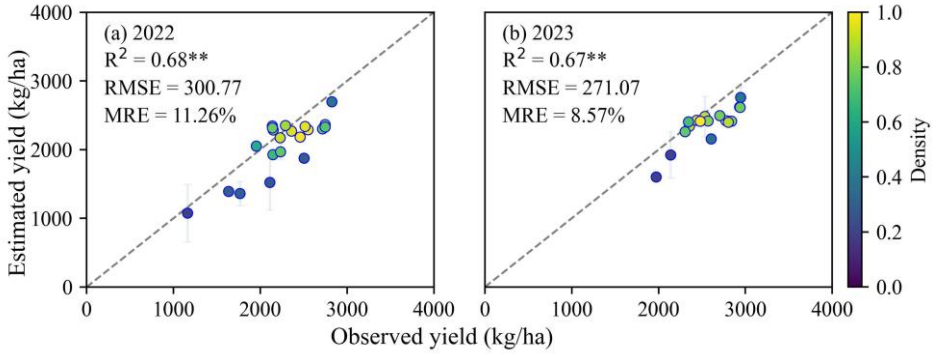


Figure 56: Scatterplots between estimated and observed soybean yield in 2022 and 2023, respectively. The error-bars represent one standard deviation indicating the uncertainty of yield estimations. Dashed line represents 1:1 line. ** denotes statistical significance at p < 0.01.

4.3 Yield estimation at regional scale

4.3.1 Variability of accuracy through years

The NortheastChinaSoybeanYield20m was validated at the municipal scale (2019 to 2022) by aggregating yield maps to match statistical data (Fig. 6). Compared to the field-scale validation, the municipal-scale estimates exhibited greater uncertainty, likely reflecting increased heterogeneity of soybean yields over larger areas. The estimates maintained stable interannual performance, with correlation between estimated and statistical yields consistently exceeding 0.60 ($p < 0.01$). The overall accuracy, pooled across 2019- 2022, for municipal-scale achieved $R^2 = 0.62$ ($p < 0.01$), RMSE = 272.36 kg ha^{-1}, and MRE = 12.08 % (Fig. 11a). Annual accuracy metrics ranged from 221.69 kg ha^{-1} to 310.66 kg ha^{-1} for RMSE and from 8.24 % to 14.40 % for MRE, with the 2022 year achieving the highest accuracy (MRE < 10%, Fig. 6d).

~~The accuracy assessment of NortheastChinaSoybeanYield20m at regional scale from 2019 to 2022 was presented in Fig. 7. The yield maps were aggregated to the municipal scale for comparison with statistics. It could be observed that the dataset achieved stable performance in different years. The correlation between annual estimated soybean yields and statistical yields were all above 0.60 ($p < 0.01$). The RMSE ranged from 221.69 kg ha^{-1} to 310.66 kg ha^{-1} and MRE from 8.24 % to 14.40 %. The result in 2022 yielded the highest accuracy with simulation error lower than 10 % (Fig. 7d). The overall accuracy in estimations through the four years was 0.62 in R^2 ($p < 0.01$), 272.36 kg ha^{-1} in RMSE and 12.08 % in MRE (Fig. 12a).~~

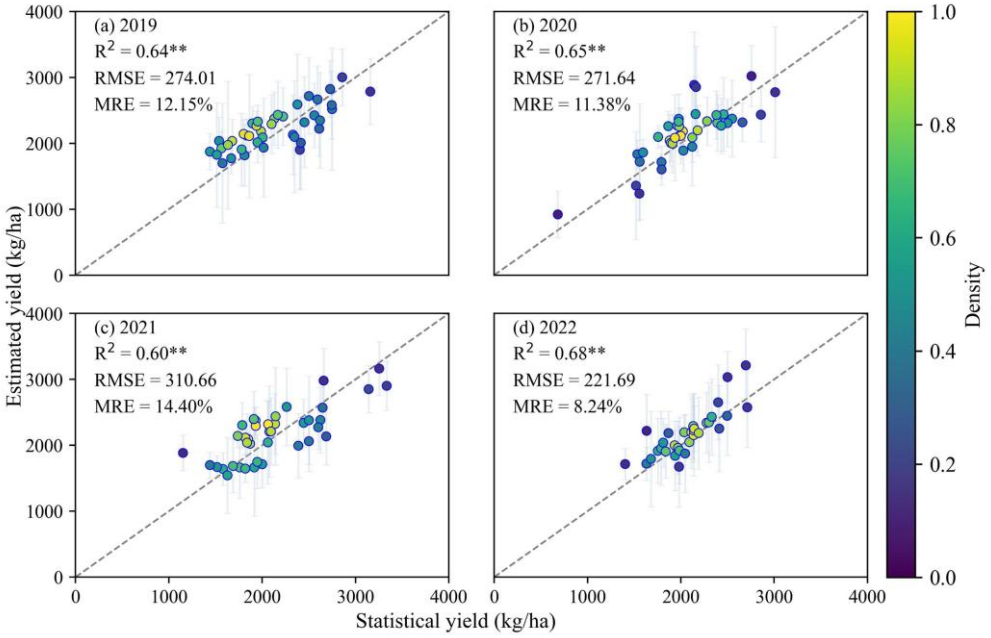


Figure ~~6~~7: Scatterplots between estimated soybean yield from Sentinel-2 and municipal statistical yields for 2019 – 2022 (excluding 2023 for which no statistical data was not available from the government). The error-bars represent one standard deviation indicating the uncertainty of yield estimations. Dashed line represents 1:1 line. ** denotes statistical significance at p < 0.01.

For temporal performance at the municipal scale,~~of the NortheastChinaSoybeanYield20m at regional scale,~~ the RMSE between estimated ~~model estimates~~ and statistical yields from~~in~~ 2019 to ~~-~~2022 remained below~~was all lower than~~ 500 kg ha^{-1}, with 80 % of cities exhibiting~~having~~ RMSE under~~below~~ 350 kg ha^{-1} (Fig. 7~~8~~a). Spatially, large~~Within the whole study area, the estimation~~ error~~s~~ were concentrated ~~of soybean yield was larger~~ in the northern part of Northeast China especially for the Greater Khingan Mountains area, while the flatter~~.~~ ~~Estimates in the relatively flat~~ region, central regions of Northeast China, showed smaller deviations~~less error~~. The spatial distribution pattern of MRE closely mirrored~~was similar to~~ that of RMSE (Fig. 7~~8~~b), averaging value~~.~~ ~~The MRE was on average~~ of 11.46 % across~~for~~ all cities over~~through~~ the four-~~-~~year periods~~.~~ These findings underscore the model's robust capacity to capture interannual variability of soybean yield.~~, indicating the great performance for the model to capture the interannual variability of soybean yield.~~

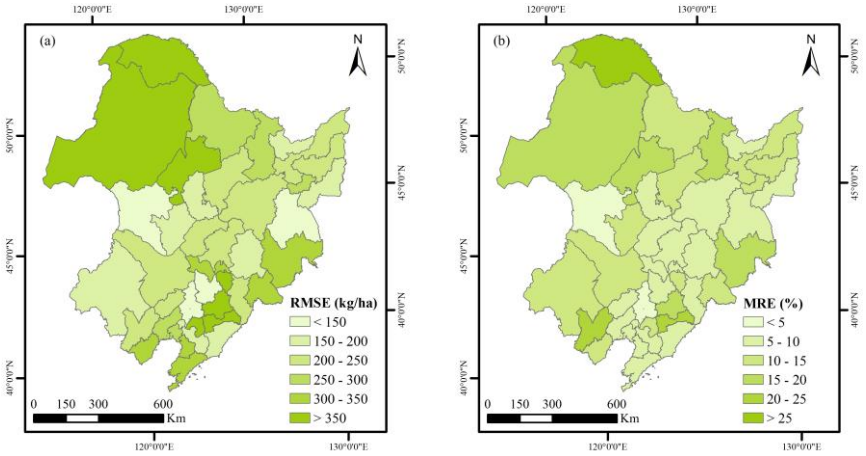


Figure 7~~8~~: Spatial patterns of the mean value of the root mean squared error and mean relative error between model estimated yields from Sentinel-2 and statistical yields from 2019 to 2022 (excluding 2023 for which no statistical data was not available from the government), (a) and (b), respectively. For years from 2019 to 2021, a total of 40 cities were calculated. For 2022, 32 cities were calculated due to missing statistics.

4.3.2 Spatial-temporal dynamics of soybean yield

To examine~~analyze the~~ spatial patterns of soybean yield across~~in~~ Northeast China, ~~the~~ yield distribution maps for ~~of soybean yield during~~ 2019 - 2023 were generated (Fig. 8~~9~~a-e). After bias correction with estimated yield derived from MODIS LAI products, the Sentinel-2 striping artifacts were markedly reduced (Fig. 8 vs. uncorrected estimates in Fig. A3), resulting in seamless 20 m yield surfaces with strong spatial continuity. Detailed yield estimations can be found in Fig. 9. Across five-year estimates, soybean yield in Northeast China predominantly ranged between 1500 and 2500 kg ha^{-1}, with higher yield concentrated in the central part plains where a region characterized by flat terrain and factorable agroclimatic conditions. The predicted yield is consistent with the municipal-scale statistical data (Table 4). Spatial variability, quantified by the coefficient of variation (CV), ranged from 17.51 % to 29.65 % over the study period, reflecting both inter and intra-annual heterogeneity in soybean productivity (Table 4). ~~After calibration with MODIS LAI products, the Sentinel-2 seaming effect has been alleviated (Fig. 10) compared with the estimates before calibration (Fig. A3). The results showed that the estimation results had good spatial continuity. The soybean yield in Northeast China was mainly concentrated between 1500 and 2500 kg ha^{-1}. The soybean yield was generally higher in the central part of Northeast China where the terrain was relatively flat. The predicted yield is consistent with the statistical values (Table 5). Spatial variability could be observed (Fig. 9), characterized by the coefficient of variation (CV) ranging from 17.51 % to 29.65 % (Table 5).~~

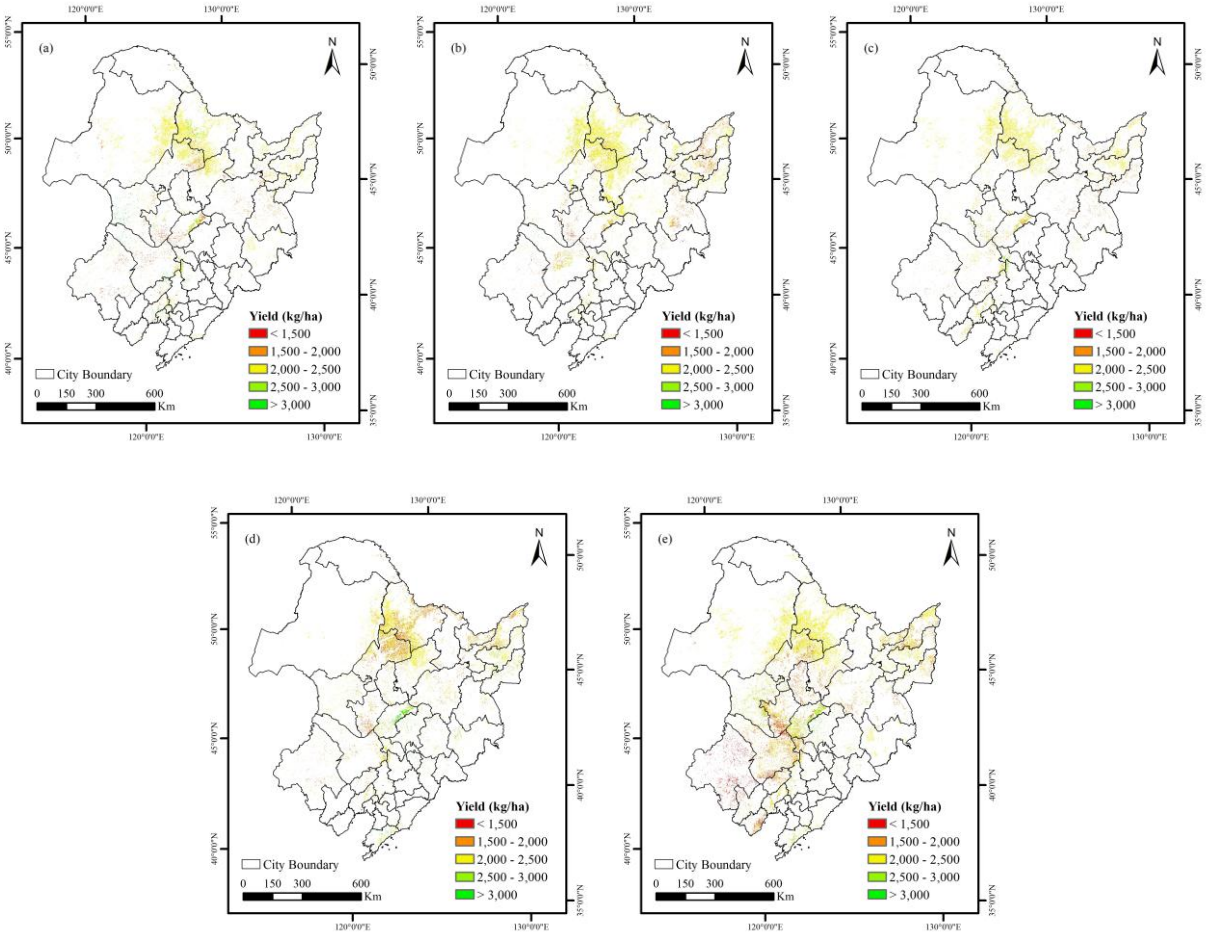


Figure 89: Spatial distribution of annual soybean yield derived from Sentinel-2 after calibration in Northeast China from 2019 to 2023.

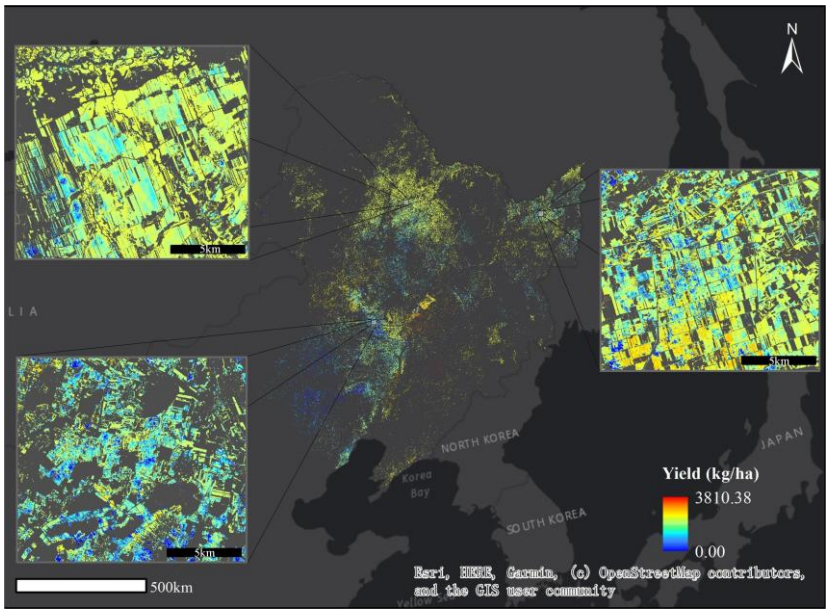


Figure 9~~10~~: An example of yield estimation result (for the year 2023) used to showcase detailed local estimates

Table ~~4~~5 Mean values of statistical soybean yield at municipal scale in Northeast China compared with mean values, standard deviation (STD) and coefficient of variation (CV) for estimated soybean yield in Northeast China.

Year	Statistics (kg ha^{-1})	Mean (kg ha^{-1})	STD (kg ha^{-1})	CV (%)
2019	2137.24	2150.02	504.61	23.47
2020	2069.08	2125.49	372.21	17.51
2021	2115.57	2136.65	374.58	17.53
2022	2073.68	2036.89	465.26	22.84
2023	——	2035.34	603.43	29.65

We further analysed the spatial-temporal variation of soybean yield at the provincial scale (Fig. 10). On average, provincial-scale estimates achieved a mean relative error (MRE) of 7.94 % (Fig. 10b), with the highest accuracy observed in 2022 (Fig. 10b), mirroring municipal-level results (Fig. 6d). Over 2019 – 2022, Liaoning Province consistently exhibited the highest yields, whereas Heilongjiang Province, despite having the largest soybean planting area, recorded the lowest yields (Fig. 10a). This disparity likely due to Heilongjiang's cooler climate, where later planting date result in shorter soybean development length. Across the four provinces, yields remained relatively stable, except in Jilin Province showed greatest interannual fluctuation exhibiting distinct decline followed by recovery. These results underscore the capacity of the proposed hybrid framework to capture spatial-temporal variations in soybean production.

~~For further study, we analysed the spatial temporal variation of soybean yield at the provincial scale (Fig. 11). The accuracy of soybean yield estimation at provincial scale was on average of 7.94 % in MRE (Fig. 11b). The accuracy of yield estimation at the provincial scale achieved the highest in 2022 (Fig. 11b), which was consistent with the results at the municipal~~

~~scale (Fig. 7d). The soybean yield was highest in Liaoning Province through the four years. Although Heilongjiang Province had the largest soybean planting area in Northeast China, its soybean yield was at its lowest among different years (Fig. 11a). This could be attributed to the climate in Heilongjiang Province. The low temperatures may make it difficult to accumulate the required heat for normal development of soybean. Results showed that the soybean yield in the four provinces remained stable between 2019-2022, while in Jilin Province showed the most noticeable fluctuation, exhibiting a trend of initially decreasing and then increasing. The predicted yield further demonstrated the effectiveness of the method proposed in this study for capturing spatial-temporal variations in soybean production.~~

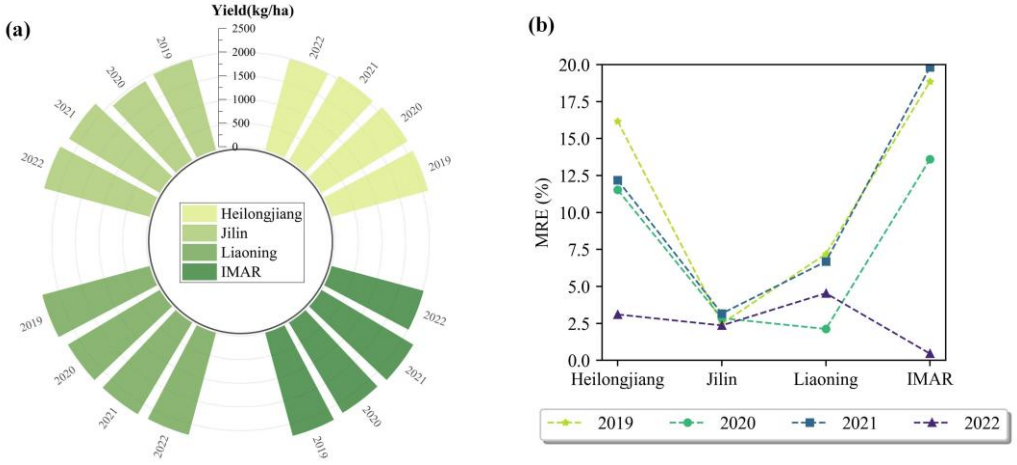


Figure 10~~11~~: Accuracy of the soybean yield estimation at provincial scale in Northeast China from 2019 to 2022. (a) represents the change in estimated yield for each province through the years; (b) represents MRE of results compared with statistical yield for each province

5 Discussion

5.1 The complementarity between MODIS and Sentinel-2

This study generated soybean yield estimates using both MODIS LAI (500 m) products and S2 derived LAI (20 m) data. Over 2019 – 2022, the MODIS-based estimates achieved an overall R^2 of 0.58 ($p < 0.01$), an RMSE of 272.36 kg ha^{-1} and an MRE of 12.08 % (Fig. 11b), slightly lower than the Sentinel-2 based results (Fig. 11a). The uncertainty of MODIS based estimates was higher than that the Sentinel-2 based estimates, likely reflecting MODIS's coarser resolution. However, the Sentinel-2 based estimates exhibit inherent seaming effects caused by cloud-affected tile edges. We additionally used MODIS LAI to bias-correct Sentinel-2 yield maps, effectively minimizing the striping ("seaming") effects in the 20 m products (Fig. 9), while preserving pixel-level detail through tile-based calibration (Fig. 13). Despite difference in spatial resolution, both MODIS and Sentinel-2 satellite data demonstrated comparable ability to capture spatiotemporal variation in soybean yield (Fig. 12), achieving correlations with statistical data > 0.55 and overall errors < 13 % across all years.

~~In this study, both MODIS and Sentinel-2 satellite data were used forgenerating the soybean yield dataset. For further comparison, accuracy assessment was done for the estimations based on MODIS LAI at municipal scale. The overall accuracy in estimations for the four years was 0.58 in R^2 ($p < 0.01$), 272.36 kg ha^{-1} in RMSE and 12.08 % in MRE (Fig. 12b), which was slightly lower than that using Sentinel-2. The uncertainty of yield estimation at municipal scale using MODIS data was higher than that using Sentinel-2. This might be due to that Sentinel-2 has a higher spatial resolution. In addition, in this study, the seaming effect of Sentinel-2 was corrected using MODIS LAI products. The seaming effect of Sentinel-2 estimates had been greatly minimized (Fig. 10). Since calibration was done on a tile-by-tile basis rather than at each pixel, the original estimation details of yield maps were well preserved (Fig. 14). The results showed that the two data sources had the same ability to capture the spatial-temporal variation in soybean yield, yielding stable and similar prediction results across different years (Fig. 13). The estimation results in different years showed good correlation with statistical data, and the overall estimation errors were less than 13 %.~~

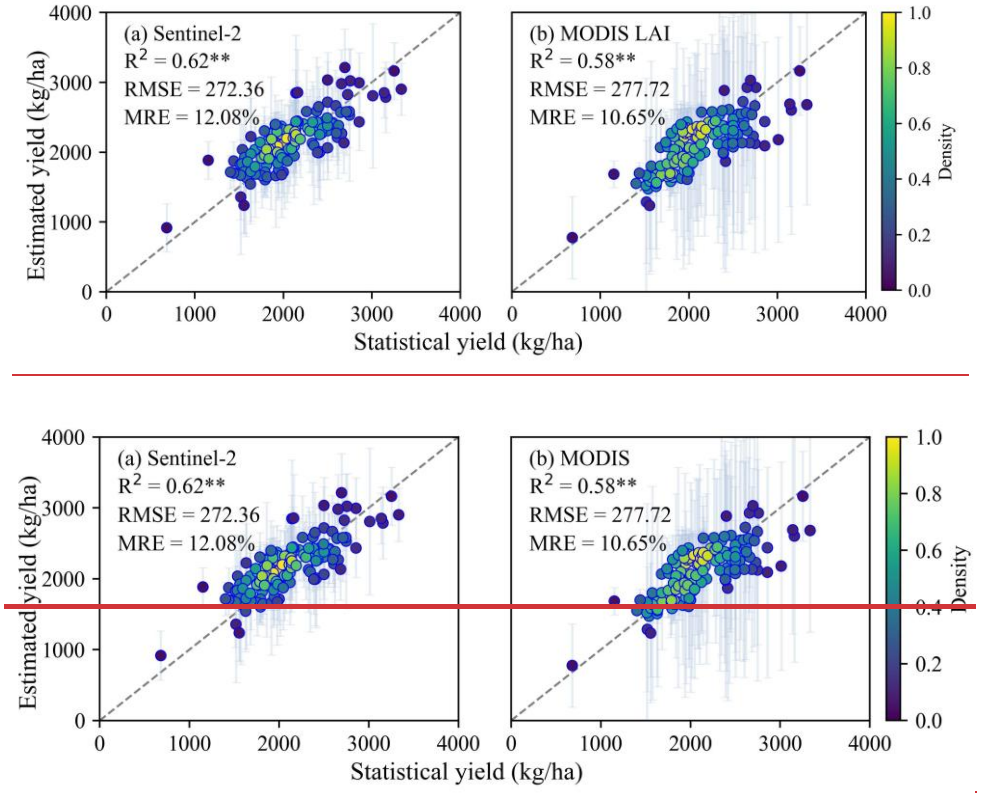


Figure 11~~12~~: Comparison between estimated and statistical yield for 2019 – 2022 using Sentinel-2 (a) and MODIS LAI (b), respectively (excluding 2023 for which no statistical data was reported). The error-bars represent one standard deviation indicating the uncertainty of yield estimations. Dashed line represents 1:1 line. ** denotes statistical significance at p < 0.01.

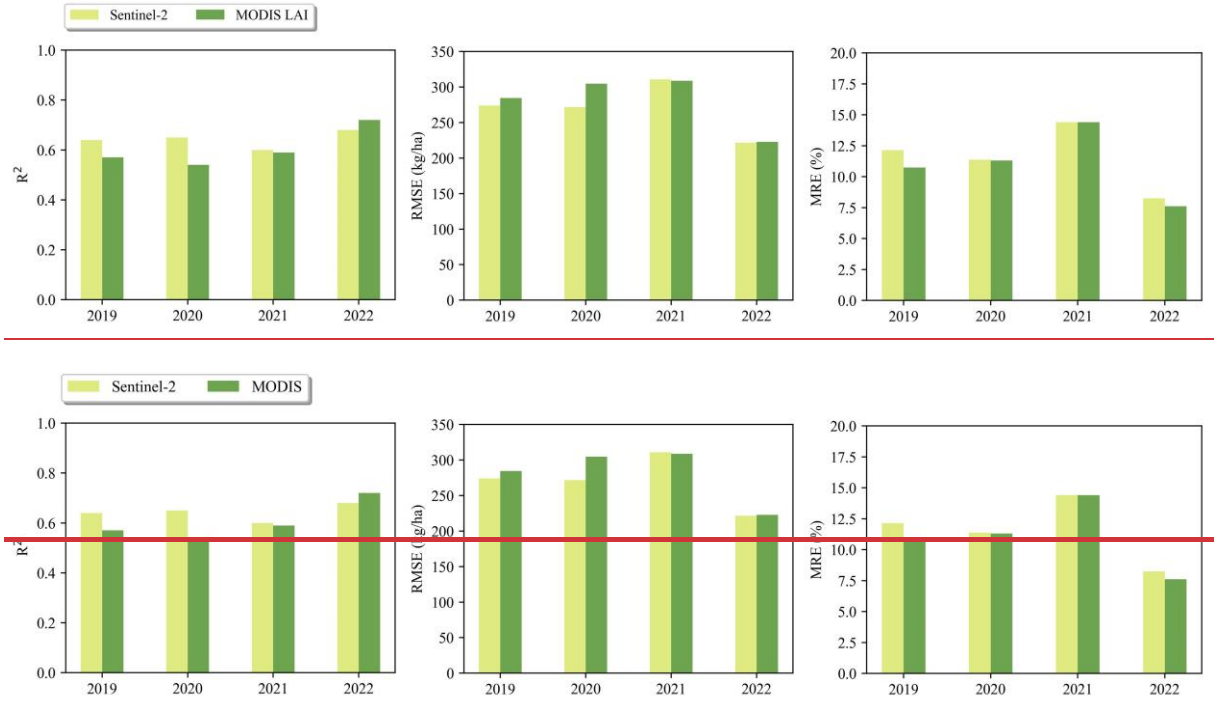


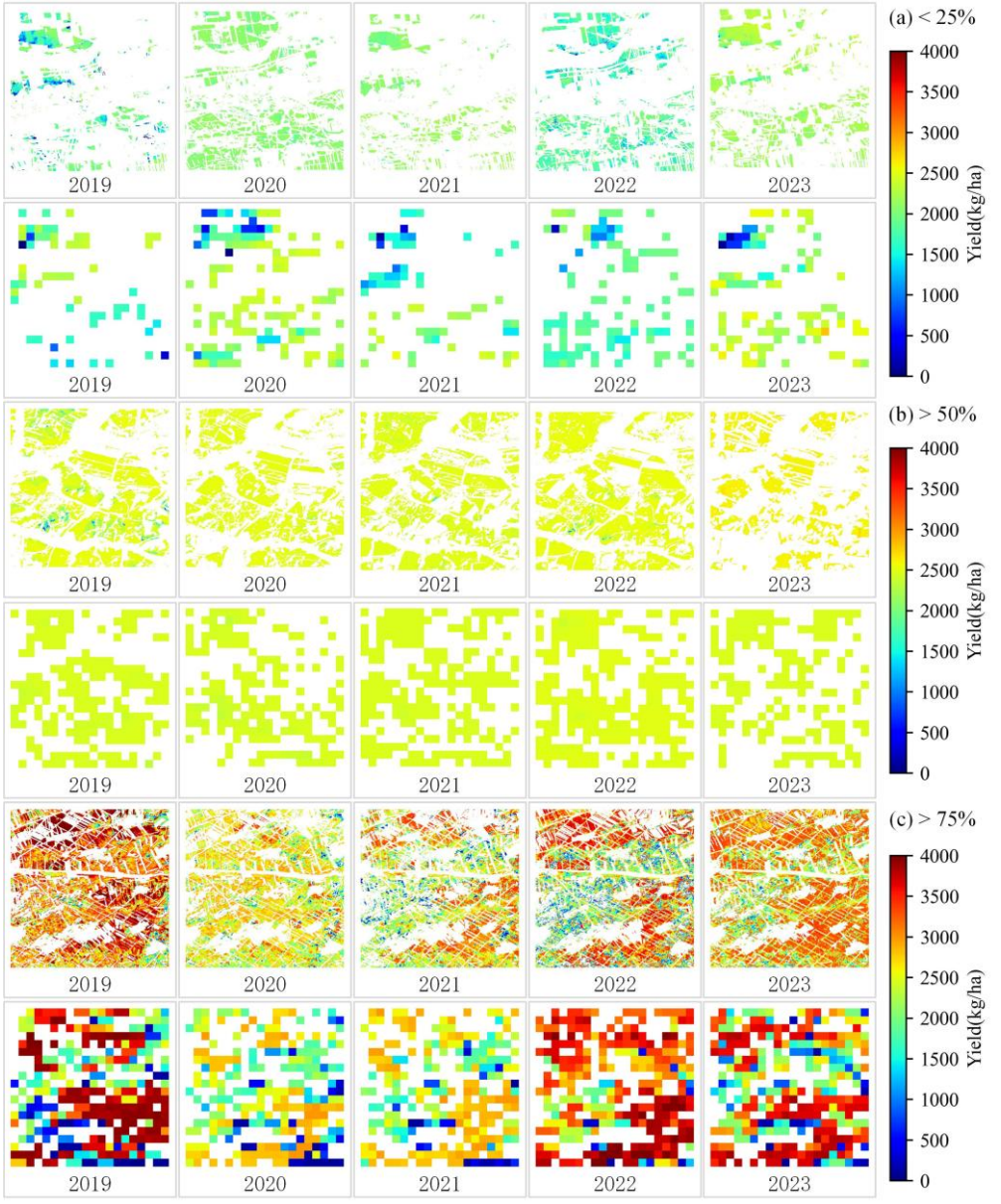
Figure 12~~13~~: Comparison of accuracy evaluation results for soybean yield estimation in 2019 – 2022 (excluding 2023, for which no statistical data was reported) using Sentinel-2 and MODIS LAI data, respectively.

In practical applications, balancing both temporal and spatial resolution is critical for achieving robust yield prediction results (Azzari et al., 2017). Figure 13 compares the Sentinel-2 yield maps and the MODIS LAI yield maps within a 10 km grid under different soybean coverage. Thanks to 4-day revisit, MODIS LAI provides more cloud-free observations during the critical growth stages, improving the reliability of two LAI metrics (LAI_{mean1} and LAI_{mean2}). Its coarser spatial resolution also accelerates spatial processing over large areas. However, Sentinel-2's finer more effectively resolves intra-field yield heterogeneity (Fig. 13). MODIS-derived maps occasionally underestimated yields due to mixed pixels containing non-crop features (e.g., infrastructure), whereas Sentinel-2 minimized such errors.

While this study prioritized high-resolution mapping (using MODIS solely for Sentinel-2 seam correction), combing high spatial data (e.g., Sentinel-2 or UAV imagery) with high temporal frequency satellites (e.g., geostationary sensors or radar) could provide an optimal data source for crop-yield modelling (Gao and Anderson, 2019; He et al., 2018).

~~In practical applications, better temporal resolution and spatial resolution are equally important to obtain ideal prediction results (Azzari et al., 2017). Since MODIS has a higher temporal resolution, it provides more available images during the crop growing season to calculate the model input features (LAI_{mean1} and LAI_{mean2}) more accurately. In addition, coarser spatial resolution can speed up spatial processing. However, our results showed that Sentinel-2 data with higher spatial resolution could be better to capture the spatial heterogeneity of yield among fields (Fig. 14). As the aim of this study was to generate yield dataset with high spatial resolution, MODIS LAI was only used to adjust the seaming effects of LAI derived from~~

~~Sentinel-2 satellite. However, further integration of the information from remote sensing data with high spatial and temporal resolution may form a more ideal data source for crop yield estimation (Gao and Anderson, 2019; He et al., 2018).~~


(a) < 25%
(b) > 50%
(c) > 75%
2019
2020
2021
2022
2023
Yield(kg/ha)
4000
3500
3000
2500
2000
1500
1000
500
0

675

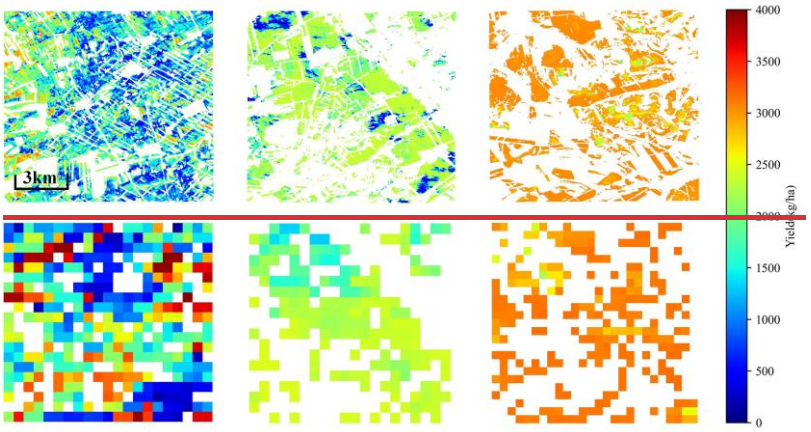


Figure 13~~14~~: Comparisons ~~Examples~~ of soybean yield estimation within a 10 km grid~~at 3 km scale~~ under different soybean coverage using Sentinel-2 (20 m) and MODIS LAI (500 m) data, where~~, respectively.~~ (a), (b), (c) represent soybean coverage less than 25%, more than 50% and more than 75%, respectively.

5.2 Advancements in this study~~Limitations and future developments~~

Accurate monitoring of soybean yield is crucial for food policy decision-making and security assessment. While previous studies have primarily explored the impact of environmental factors such as climate on soybean productivity (Guo et al., 2022; Zhao et al., 2023a), few efforts have focused on producing high-resolution soybean yield dataset for China's major soybean-producing regions. To address this gap, our study produced the NortheastChinaSoybeanYield20m dataset, a 20-meter resolution dataset generated through a hybrid framework integrating the mechanistic WOFOST crop growth model and a GRU deep learning algorithm. Unlike purely data-driven approaches that rely on extensive ground data, our approach leveraged both data mining capabilities and mechanistic modelling, which improve the model's interpretability and enhances its potential for transferability across regions. The integration of the WOFOST model ensured the simulation of diverse production scenarios under varying climate, soil, crop variety and management conditions, providing a robust synthetic training data for the GRU network. This combination allowed the model to generate well, even in areas with limited observational data, therefore overcoming common limitations related to data scarcity and high computational costs. Accuracy assessments using both in-situ and statistical yield data confirmed that the generated NortheastChinaSoybeanYield20m dataset delivered reliable yield estimates across field and regional scales (Fig. 5 and 6). The results also verified the model's stability across time and space, reinforcing its potential for large-scale agricultural monitoring and strategic planning.

When compared to previous studies using integrated remote sensing data and process-based model to estimate soybean yield, for instance, Baup et al., (2015) reported estimation error ranging from 2% to 18%, our method achieved comparable levels of accuracy. It also outperformed existing field-scale studies (e.g., RMSE = 400.946 kg ha^{-1} in Ren et al., (2023) and MRE of 29.73% in Du et al., (2014)) and municipal-scale models (e.g., RMSE = 16 % in Von Bloh et al., (2023)). Furthermore, the NortheastChinaSoybeanYield20m dataset showed improved performance relative to similar high-resolution soybean yield

products from other countries (e.g., annual 30 m soybean yield mapping in Brazil, with R^2 values between 0.31 and 0.71 and RMSEs ranging from 275 to 740 kg ha^{-1} (Song et al., 2022)).

Although studies based on UAV and RGB data have demonstrated even higher soybean yield estimation accuracy (Li et al., 2021, 2024), such methods are often constrained by high costs and limited spatial coverage, making them impractical for large-scale applications. In contrast, the method developed in this study offers a well-balanced solution that combines computational efficiency, high spatial resolution, and strong predictive accuracy. Our approach offers scalable and practical solution for producing high-resolution, large-scale crop yield datasets.

~~Accurate monitoring of soybean yield is crucial for food policy decision-making and security assessment. Previous studies have primarily focused on the impact of various factors (e.g., climate) on soybean yield (Guo et al., 2022; Zhao et al., 2023a). To our knowledge, high-resolution soybean yield dataset is currently unavailable in the main production regions of China. The study combined crop growth model with deep learning to construct a hybrid model driven by data and knowledge simultaneously for soybean yield estimation. The model retained its data mining capabilities while incorporating mechanistic constraints, thereby enhancing the model's interpretability and transferability. Accuracy verification based on in-situ and statistical data showed that the NortheastChinaSoybeanYield20m generated in this study accurately estimated soybean yield at both field and regional scales (Fig. 6 and 7).~~

5.3 Limitations and future developments

In this study, a multi-scenario soybean growth dataset was developed by simulating various combinations input parameters within the WOFOST model. These diverse scenarios were designed to reflect different environmental and management conditions, ultimately serving as training data for the yield estimation model. One advantage of the model is its scalability, it can be readily applied to other regions and countries that lack sufficient ground observation data, such as parts of Africa and India, thus offering a promising tool for global agricultural monitoring.

However, the validation results revealed some notable limitations. Specifically, the model exhibited a tendency to produce large uncertainty in low- or high- yielding areas, introducing error into the overall yield estimation (Fig. 5 and 6). This pattern suggests a systematic bias in the model's predictions, particularly in regions with extreme yield values. Additionally, spatial analysis showed that estimation errors were more pronounced in the northern region, where is characterized by complex terrain, compared to the relatively flat central region (Fig. 7). These discrepancies highlight the need to refine parameterization for extreme yield conditions and integrate higher-resolution environmental drivers (e.g., terrain, localized weather).

On the one hand, the estimation errors may be attributed to the inherent limitations of the WOFOST model. As a process-based model, WOFOST simplifies its calculations for simulating physiological processes, which can hinder its ability to fully replicate the complex realities of soybean in the field. Factors, such as pest infestations, diseases, and abiotic stresses are either oversimplified or excluded (Gaso et al., 2024). These omissions can lead to systematic simulation errors, particularly under stress conditions that significantly affect crop yield. Moreover, the parameterization of the WOFOST model in this study purely relied on values from literature and existing dataset rather than local optimization. As a result, local variability because

of farming practices, soil properties, and environmental conditions may not have been adequately captured. This lacks local optimization likely result in higher estimation error, especially in complex landscapes with spare ground observations. To address these issues, future works incorporating field-specific parameters or advanced data assimilation techniques could help reduce bias and improve model accuracy across heterogeneous landscapes. Given the spatial variability in soybean growth within the study area, constructing ecological zones based on factors like climate, elevation, and management practices might provide a more targeted model approach. For instance, Huang et al., (2023) defined the ecological zones through using Theissen polygons derived from meteorological station locations. This zoning strategy could enhance the representativeness of the training data and reduce yield estimation uncertainties.

On the other hand, the estimation errors may stem from the overfitting of the GRU model. The GRU was trained on the multi-scenarios simulated dataset, a large number of simulations that included all available combinations (e.g., all meteorological data), which introduced a significant amount of redundant information. The redundancy not only potentially reduce the dataset's representativeness, but also increase the computational burden during model training. As a result, the trained GRU model may have become overly turned to specific temporal patterns in certain years, limiting its ability to generalize to other time period or regions with different growth conditions. This overfitting effect might result in large yield estimation errors across different years and regions, particularly in areas where soybean yields deviated significantly from the norm. To address these issues, refining the structure and composition of the training dataset, and removing redundant information would enhance the diversity and quality of the training inputs. One potential approach to reduce redundancy is through spatiotemporal clustering of various environmental (e.g., meteorological station data), which could filter out stations with highly similar information. Moreover, monitoring the validation error throughout the training process, and implementing regularization techniques (e.g., L2 weight regularization) could help to prevent overfitting and improve the GRU model's generalization capability, leading to improve soybean estimation across varying conditions.

~~However, there is still several work worth of advancement to further improve the accuracy in yield estimations.~~

~~a. Improvements in construction of soybean growth dataset~~

~~In this study, a multi-scenario soybean growth dataset was constructed by setting various simulation scenarios for the input parameters of the WOFOST to provide training data for the yield estimation modeling. The model can easily be expanded to regions and countries lacking for ground observation data (such as Africa and India) to provide agricultural monitoring in the future. However, due to the extensive information in the simulation dataset, precise yield estimation came at the cost of high computational demand. In addition, as the setting of model parameters mainly came from literatures in this study, the construction of dataset might not fully account for all scenarios related to crop yield, such as pests, diseases and abiotic stresses (Gaso et al., 2024).~~

~~In future research, it will be beneficial to improve the construction of soybean growth dataset to remove redundant information and to better simulate various growing scenarios. For example, all of the meteorological data collected in this study was used for simulation, which brought a lot of redundant information to the soybean growth dataset. In future study, spatiotemporal clustering of meteorological station data can be carried out to further remove stations with redundant data. As~~

~~the growth of soybeans exhibit spatial variability in the study area, it is recommended to construct different dataset for different ecological zones if more ground observation data is available in the future. The method of establishing ecological zones includes establishing Tyson polygons based on meteorological stations (Huang et al., 2023).~~

~~b. Improvements in model estimations~~

Finally, a~~A~~ccurate estimation of soybean yield depended on the quality of the input data. The spatial resolution of remote sensing imagery could limit the model's ability to predict spatial variability in yield. In this study, ERA5-land dataset was applied to obtain the spatial-temporal distribution of soybean phenology in the study area. To be consistent with Sentinel-2 data, all datasets were resampled to a 20 m resolution. Downscaling the coarse spatial data could increase the uncertainty of inputs to the model.

Moreover, the issue of mixed pixels led to a loss of detailed surface information (Zhao et al., 2023), particularly in heterogeneous or complex environments. With the advent of higher spatial-temporal resolution remote sensing data, the estimation accuracy of crop yield is expected to be further improved. The mean values of LAI at two key soybean growth stages were used as the primary remote sensing-based input features for yield estimation. However, errors in LAI retrieval from remote sensing data also contributed uncertainty in yield predictions. Integrating agronomic knowledge with remote sensing mechanisms has emerged as a promising way to reduce the uncertainty and improve model reliability (Chen et al., 2022; Hu et al., 2024). Coupling radiative transfer model such as PROSAIL (Jacquemoud et al., 2009)) with crop growth model can enhance the simulation of leaf and canopy characteristics and provide additional constraints for more accurate modelling (Ntakos et al., 2024). In addition, the combination of IoT, blockchain, and precision agriculture with machine learning and biophysical models can offer a powerful framework for sustainable agricultural monitoring, addressing challenges in data heterogeneity, model scalability, and decision-making processes. These technologies can facilitate real-time data collection, ensure data security and transparency. Precision agriculture techniques, combined with advanced sensing technologies, can effectively improve the accuracy and timeliness of input data, addressing current limitations in model calibration, validation and prediction.

~~In addition, the presence of mixed pixels led to the loss of part of the surface information (Zhao et al., 2023) and impaired the prediction ability of the model especially in the complex environment (Fig. 8). With the advent of higher spatial–temporal resolution remote sensing data, the estimation accuracy of crop yield is expected to be further improved. The mean values of LAI at two growth stages were used as model input features for regional soybean yield estimation. The errors in LAI inversion from remote sensing data were introduced into the model prediction, thus increasing the uncertainty of yield estimation. The coupling of agronomic knowledge and remote sensing mechanism has become a new research focus (Chen et al., 2022; Hu et al., 2024). By coupling radiative transfer model such as PROSAIL (Jacquemoud et al., 2009)) with crop growth model, remote sensing data can further provide additional constraints and enhance the simulation ability of crop growth models (Ntakos et al., 2024).~~

~~For example, the time series of crop biophysical variables simulated by crop growth models can serve as input for radiative transfer models to generate corresponding spectral reflectance. Subsequently, these spectral reflectance data can further be used as input features for deep learning models to estimate crop yield.~~

6 Data availability

The soybean yield dataset for Northeast China (NortheastChinaSoybeanYield20m) during the 2019-2023 period is available at https://doi.org/10.5281/zenodo.14263103 (Xu et al., 2024).

7 Conclusions

This study generated a high-resolution (20 m) soybean yield dataset for~~in~~ Northeast Chinda from 2019 to 2023 (~~, called~~ NortheastChinaSoybeanYield20m) using a hybrid framework that ~~method~~ couple~~ing~~ the WOFOST crop growth model ~~(WOFOST)~~ with a Gated Recurrent Unit (GRU) deep learning algorithm ~~(GRU)~~. The framework leveraged a comprehensive soybean growth dataset simulated by WOFOST, which accounted for diverse production scenarios, including variations in ~~The construction of the hybrid method was based on a huge soybean growth dataset simulated by WOFOST accounting for various~~ climates, crop varieties, soil types and agro-managements practices. This approach ~~The method~~ effectively reduces reliance~~the dependence~~ on ground observation data, which demonstrating enhanced spatiotemporal generalization capabilities.~~and had better spatial-temporal generalization.~~

The ~~soybean yield~~ dataset was conducted~~generated~~ using multi-source remote sensing data, with Sentinel-2 derived time-series LAI as the primary input. Yield estimations showed robust~~, over the years 2019-2023. The results showed that the~~performance ~~of the NortheastChinaSoybeanYield20m~~ at both field and municipal~~regional~~ scales, achieving RMSE of ~~was encouraging. The yield estimations were highly consistent with in-situ measured and municipal statistical data ($p < 0.01$). The overall accuracy was~~ 287.44 kg ha^{-1} and 272.36 kg ha^{-1} ~~in RMSE~~, respectively. To address spatial discontinuities in Sentinel-2 data, corrections using MODIS LAI-derived yield maps effectively mitigated seam effects, achieving complementary benefits in temporal and spatial resolution. ~~It was worth emphasizing that after correction with yield maps derived from MODIS LAI products, the seaming effect of Sentinel-2 was mitigated. The combined use of multi-source remote sensing data realized the complementarity of temporal resolution and spatial resolution. The results of estimation showed great spatial continuity.~~ The final dataset exhibits high temporal stability and spatial continuity, with mean relative errors (MRE) averaging of ~~had stable yield estimation performance through different years, and could effectively capture the spatial-temporal variation of soybean yield (MRE on average of~~ 11.46 % at the ~~for~~ municipal scale and 7.94 % at the ~~for~~ provincial scale~~). Our proposed soybean yield dataset is helpful for optimizing soybean production distribution and ensuring food security~~.

The NortheastChinaSoybeanYield20m dataset successfully captures fine-scale spatiotemporal variations in soybean yield, offering potentials for optimizing production strategies, guiding precision agriculture, and enhancing food security and policy.

Authorship contributions

JX (first author) and QL – conceptualization; JX (first author), XD, YZ, HW, JX, YS and YD – data curation; JX (first author), XD, TD – methodology; JX (first author), XD, JX and JZ – investigation; TD and QL – supervision; HW, JX (first author) and JZ – validation; YZ, HW and JZ – visualization; JX (first author) – original draft preparation; XD, TD and YZ – reviewing and editing the manuscript.

Competing interests

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 research was funded by the National Key R&D Program of China (2021YFD1500103), the Strategic Priority Research Program of the Chinese Academy of Sciences (XDA28070504), the National Science Foundation of China (42371359), and the Key Program of High-resolution Earth Observation System (71-Y50G10-9001-22/23).

Appendix A

Table A1 Values of crop parameters in WOFOST.

Parameter	Description	Units	Value	Source
Crop initial parameters				
TDWI	Initial total crop dry weight	kg ha^{-1}	120	Default value in WOFOST
RGRLAI	Maximum relative increase in LAI	ha ha^{-1} d^{-1}	0.01	Default value in WOFOST
Parameters for emergence				
TBASEM	Minimum threshold temperature for emergence	℃	8.0	Qu et al., (2023)
TEFFMX	Maximum threshold temperature for emergence	℃	22.0	Qu et al., (2023)
TSUMEM	Accumulated temperature from sowing to emergence	℃	70.0	Qu et al., (2023)
Phenological parameters				
DLO	Optimal daylength for development	h	-99	Default value in WOFOST
DLC	Critical daylength	h	-99	Default value in WOFOST

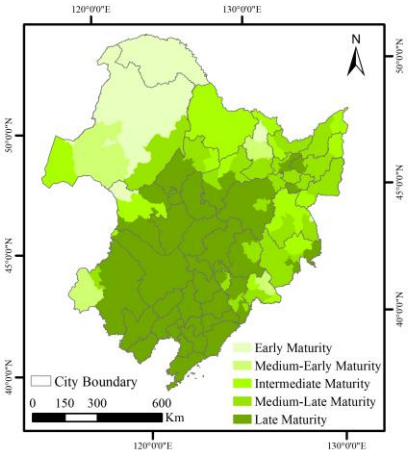
TSUM1	Cumulative temperature from emergence to anthesis	°C	450 (early maturity) 480 (medium-early maturity) 520 (intermediate maturity) 540 (medium-late maturity) 580 (late maturity)	Qu et al., (2023)
TSUM2	Cumulative temperature from anthesis to maturity	°C	660 (early maturity) 770 (medium-early maturity) 870 (intermediate maturity) 960 (medium-late maturity) 1000 (late maturity)	Qu et al., (2023)
Green area parameters				
TBASE	Lower threshold temperature for aging of leaves	°C	7.0	Default value in WOFOST
SPAN	Life span of leaves growing at 35 °C	d	23	Default value in WOFOST
SLATB00	Specific leaf area at DVS = 0.00	ha kg^{-1}	0.00140	Default value in WOFOST
SLATB045	Specific leaf area at DVS = 0.45	ha kg^{-1}	0.00250	Default value in WOFOST
SLATB090	Specific leaf area at DVS = 0.90	ha kg^{-1}	0.00250	Default value in WOFOST
SLATB200	Specific leaf area at DVS = 2.00	ha kg^{-1}	0.00070	Default value in WOFOST
Assimilation parameters				
KDIFTB00	Extinction coefficient for diffuse visible light (DVS = 0)	-	0.80	Default value in WOFOST
KDIFTB200	Extinction coefficient for diffuse visible light (DVS = 2)	-	0.80	Default value in WOFOST
EFFTB0	Light use efficiency of a single leaf (T = 0 °C)	kg ha^{-1} h^{-1} J^{-1} m^2 s^{-1}	0.40	Default value in WOFOST
EFTB40	Light use efficiency of a single leaf (T = 40 °C)	kg ha^{-1} h^{-1} J^{-1} m^2 s^{-1}	0.40	Default value in WOFOST

AMAXTB00	Maximum leaf CO_2 assimilation rate (DVS = 0)	kg ha^{-1} h^{-1}	29.00	Default value in WOFOST
AMAXTB170	Maximum leaf CO_2 assimilation rate (DVS = 1.7)	kg ha^{-1} h^{-1}	25.31	Sun et al., (2022)
AMAXTB200	Maximum leaf CO_2 assimilation rate (DVS = 2)	kg ha^{-1} h^{-1}	0.00	Default value in WOFOST
TMPFTB00	Reduction factor of AMAX (T = 0 °C)	-	0.00	Default value in WOFOST
TMPFTB10	Reduction factor of AMAX (T = 10 °C)	-	0.30	Default value in WOFOST
TMPFTB20	Reduction factor of AMAX (T = 20 °C)	-	0.60	Default value in WOFOST
TMPFTB25	Reduction factor of AMAX (T = 25 °C)	-	0.80	Default value in WOFOST
TMPFTB30	Reduction factor of AMAX (T = 30 °C)	-	1.00	Default value in WOFOST
TMPFTB35	Reduction factor of AMAX (T = 35 °C)	-	1.00	Default value in WOFOST
Conversion of assimilates into biomass				
CVL	Conversion efficiency of assimilates into leaf tissue	kg kg^{-1}	0.72	Default value in WOFOST
CVO	Conversion efficiency of assimilates into storage organs	kg kg^{-1}	0.48	Default value in WOFOST
CVR	Conversion efficiency of assimilates into root tissue	kg kg^{-1}	0.72	Default value in WOFOST
CVS	Conversion efficiency of assimilates into stem tissue	kg kg^{-1}	0.69	Default value in WOFOST
Maintenance respiration parameters				
Q10	Relative change in respiration rate per 10 °C temperature increase	-	2.0	Default value in WOFOST
RML	Ralative maintenance respiration rate of leaves	kg CH_2O kg^{-1} d^{-1}	0.03	Default value in WOFOST

RMO	Ralative maintenance respiration rate of storage organs	kg CH_2O kg^{-1} d^{-1}	0.017	Default value in WOFOST
RMR	Ralative maintenance respiration rate of toots	kg CH_2O kg^{-1} d^{-1}	0.01	Default value in WOFOST
RMS	Ralative maintenance respiration rate of stems	kg CH_2O kg^{-1} d^{-1}	0.015	Default value in WOFOST
Partitioning parameters				
FRTB00	Fraction of total dry matter to roots at DVS = 0	kg kg^{-1}	0.62	Sun et al., (2022)
FRTB075	Fraction of total dry matter to roots at DVS = 0.75	kg kg^{-1}	0.35	Default value in WOFOST
FRTB100	Fraction of total dry matter to roots at DVS = 1	kg kg^{-1}	0.15	Default value in WOFOST
FRTB150	Fraction of total dry matter to roots at DVS = 1.5	kg kg^{-1}	0.00	Default value in WOFOST
FRTB200	Fraction of total dry matter to roots at DVS = 2.0	kg kg^{-1}	0.00	Default value in WOFOST
FLTB00	Fraction of total dry matter to leaves at DVS = 0	kg kg^{-1}	0.70	Default value in WOFOST
FLTB100	Fraction of total dry matter to leaves at DVS = 1.0	kg kg^{-1}	0.70	Default value in WOFOST
FLTB115	Fraction of total dry matter to leaves at DVS = 1.15	kg kg^{-1}	0.60	Default value in WOFOST
FLTB130	Fraction of total dry matter to leaves at DVS = 1.3	kg kg^{-1}	0.43	Default value in WOFOST
FLTB150	Fraction of total dry matter to leaves at DVS = 1.5	kg kg^{-1}	0.15	Default value in WOFOST
FLTB200	Fraction of total dry matter to leaves at DVS = 2.0	kg kg^{-1}	0.00	Default value in WOFOST
FSTB00	Fraction of total dry matter to stems at DVS = 0	kg kg^{-1}	0.30	Default value in WOFOST
FSTB100	Fraction of total dry matter to stems at DVS = 1.0	kg kg^{-1}	0.30	Default value in WOFOST

FSTB115	Fraction of total dry matter to stems at DVS = 1.15	kg kg^{-1}	0.25	Default value in WOFOST
FSTB130	Fraction of total dry matter to stems at DVS = 1.3	kg kg^{-1}	0.10	Default value in WOFOST
FSTB150	Fraction of total dry matter to stems at DVS = 1.5	kg kg^{-1}	0.10	Default value in WOFOST
FSTB200	Fraction of total dry matter to stems at DVS = 2.0	kg kg^{-1}	0.00	Default value in WOFOST
FOTB00	Fraction of total dry matter to storage organs at DVS = 0	kg kg^{-1}	0.00	Default value in WOFOST
FOTB100	Fraction of total dry matter to storage organs at DVS = 1.0	kg kg^{-1}	0.00	Default value in WOFOST
FOTB115	Fraction of total dry matter to storage organs at DVS = 1.15	kg kg^{-1}	0.15	Default value in WOFOST
FOTB130	Fraction of total dry matter to storage organs at DVS = 1.3	kg kg^{-1}	0.47	Default value in WOFOST
FOTB150	Fraction of total dry matter to storage organs at DVS = 1.5	kg kg^{-1}	0.75	Default value in WOFOST
FOTB200	Fraction of total dry matter to storage organs at DVS = 2.0	kg kg^{-1}	1.00	Default value in WOFOST
Death rate parameters				
PERDL	Maximum relative death rate of leaves due to water stress	kg kg^{-1} d^{-1}	0.03	Default value in WOFOST
RDRRTB00	Relative death rate of roots at DVS = 0	kg kg^{-1} d^{-1}	0.00	Default value in WOFOST
RDRRTB150	Relative death rate of roots at DVS = 1.5	kg kg^{-1} d^{-1}	0.00	Default value in WOFOST
RDRRTB151	Relative death rate of roots at DVS = 1.51	kg kg^{-1} d^{-1}	0.02	Default value in WOFOST
RDRRTB200	Relative death rate of roots at DVS = 2.0	kg kg^{-1} d^{-1}	0.02	Default value in WOFOST
RDRSTB00	Relative death rate of stems at DVS = 0	kg kg^{-1} d^{-1}	0.00	Default value in WOFOST

RDRSTB150	Relative death rate of stems at DVS = 1.5	kg kg^{-1} d^{-1}	0.00	Default value in WOFOST
RDRSTB151	Relative death rate of stems at DVS = 1.51	kg kg^{-1} d^{-1}	0.02	Default value in WOFOST
RDRSTB200	Relative death rate of stems at DVS = 2.0	kg kg^{-1} d^{-1}	0.02	Default value in WOFOST
Water use parameters				
CFET	Correction factor transpiration rate	-	1.0	Default value in WOFOST
DEPNR	Crop group number for soil water depletion	-	5.0	Default value in WOFOST
IAIRDU	Air ducts in roots present (=1) or not (=0)	-	0	Default value in WOFOST
IOX	Oxygen stress effect enabled (=1) or not (=0)	-	0	Default value in WOFOST
Rooting parameters				
RDI	Initial rooting depth	cm	10	Default value in WOFOST
RRI	Maximum daily increase in rooting depth	cm d^{-1}	1.2	Default value in WOFOST
RDMCR	Maximum rooting depth	cm	120	Default value in WOFOST



845 **Figure A1: Spatial distribution of soybean types in Northeast China.**

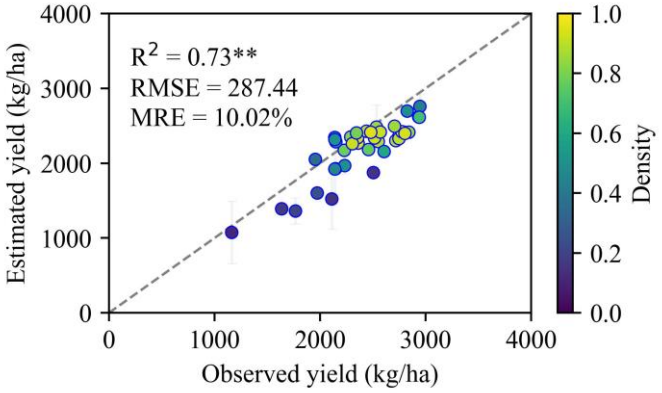


Figure A2: Comparison between estimated and observed yield for both of 2022 and 2023. The error-bars represent one standard deviation indicating the uncertainty of yield estimations. Dashed lines represent 1:1 line. ** denotes statistical significance at p < 0.01.

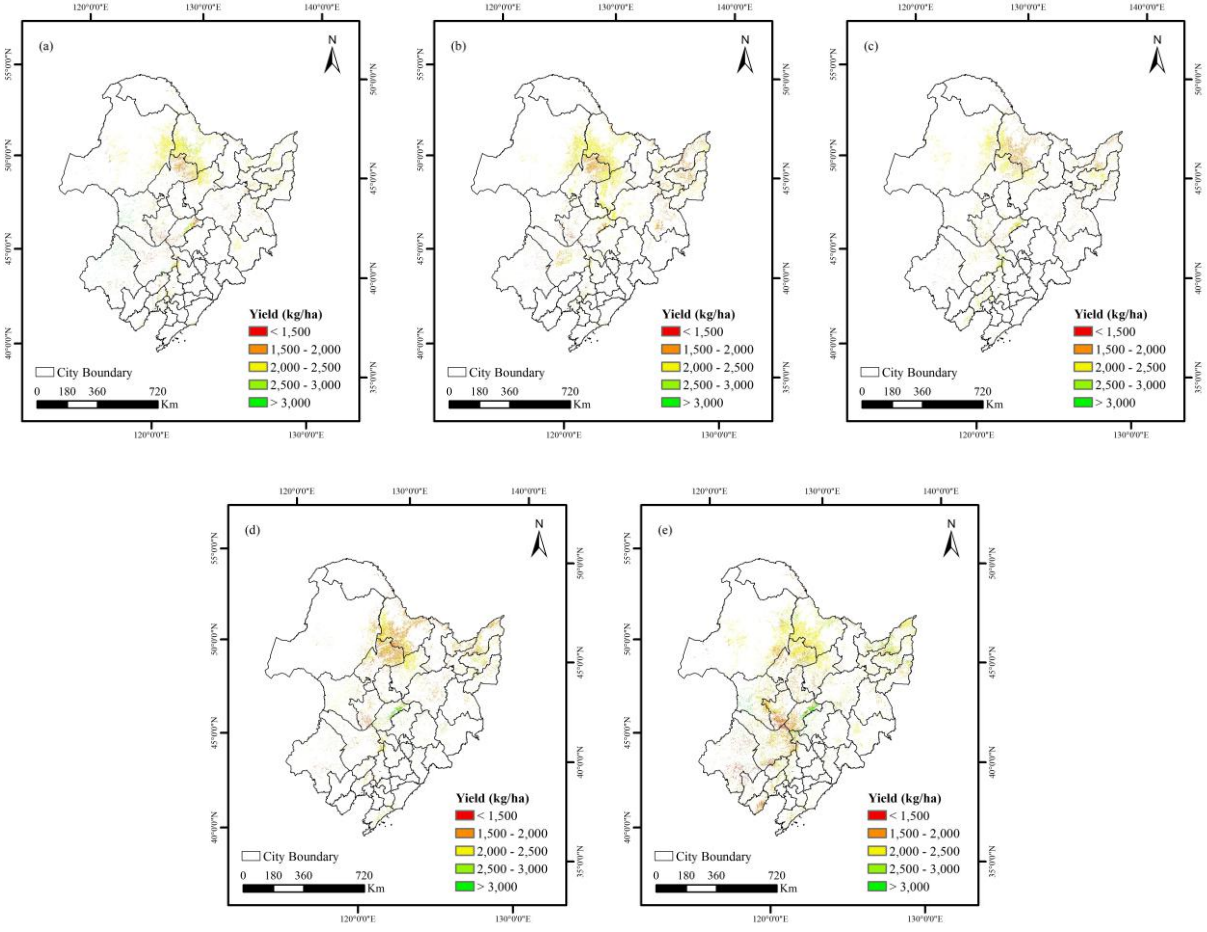


Figure A3: Spatial distribution of annual soybean yield derived from Sentinel-2 before calibration in Northeast China from 2019 to 2023.

References

Açikkar, M.: Fast grid search: A grid search-inspired algorithm for optimizing hyperparameters of support vector regression, Turkish Journal of Electrical Engineering and Computer Sciences, 32, 68–92, https://doi.org/10.55730/1300-0632.4056, 2024.

Allen, L. H., Kirkham, M. B., Olszyk, D. M., Whitman, C. E., and Pickering, N. B.: Plant Modeling: Advances and Gaps in Our Capability to Predict Future Crop Growth and Yield in Response to global Climate Change, Advances in Carbon Dioxide Effects Research, 1997.

Ang, Y., Shafri, H. Z. M., Lee, Y. P., Abidin, H., Bakar, S. A., Hashim, S. J., Che'Ya, N. N., Hassan, M. R., Lim, H. S., and Abdullah, R.: A novel ensemble machine learning and time series approach for oil palm yield prediction using Landsat time series imagery based on NDVI, Geocarto International, 37, 9865–9896, https://doi.org/10.1080/10106049.2022.2025920, 2022.

Azzari, G., Jain, M., and Lobell, D. B.: Towards fine resolution global maps of crop yields: Testing multiple methods and satellites in three countries, Remote Sensing of Environment, 202, 129–141, https://doi.org/10.1016/j.rse.2017.04.014, 2017.

Baup, F., Fieuzal, R., and Betbeder, J.: Estimation of soybean yield from assimilated optical and radar data into a simplified agrometeorological model, in: 2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), IGARSS 2015 - 2015 IEEE International Geoscience and Remote Sensing Symposium, Milan, Italy, 3961–3964, https://doi.org/10.1109/IGARSS.2015.7326692, 2015.

Cao, J., Zhang, Z., Tao, F., Zhang, L., Luo, Y., Zhang, J., Han, J., and Xie, J.: Integrating Multi-Source Data for Rice Yield Prediction across China using Machine Learning and Deep Learning Approaches, Agricultural and Forest Meteorology, 297, 108275, https://doi.org/10.1016/j.agrformet.2020.108275, 2021.

Chen, Q., Zheng, B., Chen, T., and Chapman, S. C.: Integrating a crop growth model and radiative transfer model to improve estimation of crop traits based on deep learning, Journal of Experimental Botany, 73, 6558–6574, https://doi.org/10.1093/jxb/erac291, 2022.

Chen, Y., Liu, S., Li, H., Li, X. F., Song, C. Y., Cruse, R. M., and Zhang, X. Y.: Effects of conservation tillage on corn and soybean yield in the humid continental climate region of Northeast China, Soil and Tillage Research, 115–116, 56–61, https://doi.org/10.1016/j.still.2011.06.007, 2011.

Cho, K., Van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y.: Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 1724–1734, https://doi.org/10.3115/v1/D14-1179, 2014.

Choi, D.-H., Ban, H.-Y., Seo, B.-S., Lee, K.-J., and Lee, B.-W.: Phenology and Seed Yield Performance of Determinate Soybean Cultivars Grown at Elevated Temperatures in a Temperate Region, PLoS ONE, 11, e0165977, https://doi.org/10.1371/journal.pone.0165977, 2016.

Diepen, C. A., Wolf, J., Keulen, H., and Rappoldt, C.: WOFOST: a simulation model of crop production, Soil Use & Management, 5, 16–24, https://doi.org/10.1111/j.1475-2743.1989.tb00755.x, 1989.

Dokoohaki, H., Kivi, M. S., Martinez-Feria, R., Miguez, F. E., and Hoogenboom, G.: A comprehensive uncertainty quantification of large-scale process-based crop modeling frameworks, Environ. Res. Lett., 16, 084010, https://doi.org/10.1088/1748-9326/ac0f26, 2021.

Dong, T., Liu, J., Qian, B., He, L., Liu, J., Wang, R., Jing, Q., Champagne, C., McNairn, H., Powers, J., Shi, Y., Chen, J. M., and Shang, J.: Estimating crop biomass using leaf area index derived from Landsat 8 and Sentinel-2 data, ISPRS Journal of Photogrammetry and Remote Sensing, 168, 236–250, https://doi.org/10.1016/j.isprsjprs.2020.08.003, 2020.

Du, X., Song, F., Wang, H., Huanxuezhang, Meng, J., Li, Q., Liu, J., Ding, L., and Lu, Y.: Soybean yield estimation using HJ-1 CCD data in Northeast China, in: 2014 The Third International Conference on Agro-Geoinformatics, 2014 Third International Conference on Agro-Geoinformatics, Beijing, China, 1–4, https://doi.org/10.1109/Agro-Geoinformatics.2014.6910627, 2014.

Du, X., Zhu, J., Xu, J., Li, Q., Tao, Z., Zhang, Y., Wang, H., and Hu, H.: Remote sensing-based winter wheat yield estimation integrating machine learning and crop growth multi-scenario simulations, International Journal of Digital Earth, 18, 2443470, https://doi.org/10.1080/17538947.2024.2443470, 2025.

Duchemin, B., Maisongrande, P., Boulet, G., and Benhadj, I.: A simple algorithm for yield estimates: Evaluation for semi-arid irrigated winter wheat monitored with green leaf area index, Environmental Modelling & Software, 23, 876–892, https://doi.org/10.1016/j.envsoft.2007.10.003, 2008.

Falcon, W. P., Naylor, R. L., and Shankar, N. D.: Rethinking Global Food Demand for 2050, Population & Development Rev, 48, 921–957, https://doi.org/10.1111/padr.12508, 2022.

Fan, R., Zhang, X., Liang, A., Shi, X., Chen, X., Bao, K., Yang, X., and Jia, S.: Tillage and rotation effects on crop yield and profitability on a Black soil in northeast China, Can. J. Soil. Sci., 92, 463–470, https://doi.org/10.4141/cjss2010-020, 2012.

FAOSTAT: FAO statistical database, 2022.

Feng, P., Wang, B., Liu, D. L., Waters, C., Xiao, D., Shi, L., and Yu, Q.: Dynamic wheat yield forecasts are improved by a hybrid approach using a biophysical model and machine learning technique, Agricultural and Forest Meteorology, 285–286, 107922, https://doi.org/10.1016/j.agrformet.2020.107922, 2020.

Gao, F. and Anderson, M.: Evaluating Yield Variability of Corn and Soybean Using Landsat-8, Sentinel-2 and Modis in Google Earth Engine, in: IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium, IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium, Yokohama, Japan, 7286–7289, https://doi.org/10.1109/IGARSS.2019.8897990, 2019.

Gaso, D. V., Paudel, D., De Wit, A., Puntel, L. A., Mullissa, A., and Kooistra, L.: Beyond assimilation of leaf area index: Leveraging additional spectral information using machine learning for site-specific soybean yield prediction, Agricultural and Forest Meteorology, 351, 110022, https://doi.org/10.1016/j.agrformet.2024.110022, 2024.

Gevaert, C. M.: Explainable AI for earth observation: A review including societal and regulatory perspectives, International Journal of Applied Earth Observation and Geoinformation, 112, 102869, https://doi.org/10.1016/j.jag.2022.102869, 2022.

Gitelson, A. and Merzlyak, M. N.: Spectral Reflectance Changes Associated with Autumn Senescence of Aesculus hippocastanum L. and Acer platanoides L. Leaves. Spectral Features and Relation to Chlorophyll Estimation, Journal of Plant Physiology, 143, 286–292, https://doi.org/10.1016/S0176-1617(11)81633-0, 1994.

Gopi, P. S. S. and Karthikeyan, M.: Red fox optimization with ensemble recurrent neural network for crop recommendation and yield prediction model, Multimed Tools Appl, 83, 13159–13179, https://doi.org/10.1007/s11042-023-16113-2, 2023.

Graham, P. H. and Vance, C. P.: Legumes: Importance and constraints to greater use, Plant physiology, 131, 872–877, 2003.

Guo, S., Guo, E., Zhang, Z., Dong, M., Wang, X., Fu, Z., Guan, K., Zhang, W., Zhang, W., Zhao, J., Liu, Z., Zhao, C., and Yang, X.: Impacts of mean climate and extreme climate indices on soybean yield and yield components in Northeast China, Science of The Total Environment, 838, 156284, https://doi.org/10.1016/j.scitotenv.2022.156284, 2022.

He, M., Kimball, J., Maneta, M., Maxwell, B., Moreno, A., Beguería, S., and Wu, X.: Regional Crop Gross Primary Productivity and Yield Estimation Using Fused Landsat-MODIS Data, Remote Sensing, 10, 372, https://doi.org/10.3390/rs10030372, 2018.

Hu, P., Zheng, B., Chen, Q., Grunefeld, S., Choudhury, M. R., Fernandez, J., Potgieter, A., and Chapman, S. C.: Estimating aboveground biomass dynamics of wheat at small spatial scale by integrating crop growth and radiative transfer models with satellite remote sensing data, Remote Sensing of Environment, 311, 114277, https://doi.org/10.1016/j.rse.2024.114277, 2024.

Huang, H., Huang, J., Wu, Y., Zhuo, W., Song, J., Li, X., Li, L., Su, W., Ma, H., and Liang, S.: The Improved Winter Wheat Yield Estimation by Assimilating GLASS LAI Into a Crop Growth Model With the Proposed Bayesian Posterior-Based Ensemble Kalman Filter, IEEE Trans. Geosci. Remote Sensing, 61, 1–18, https://doi.org/10.1109/TGRS.2023.3259742, 2023.

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

Huang, J., Gómez-Dans, J. L., Huang, H., Ma, H., Wu, Q., Lewis, P. E., Liang, S., Chen, Z., Xue, J.-H., Wu, Y., Zhao, F., Wang, J., and Xie, X.: Assimilation of remote sensing into crop growth models: Current status and perspectives, Agricultural and Forest Meteorology, 276–277, 107609, https://doi.org/10.1016/j.agrformet.2019.06.008, 2019.

Huang, J., Song, J., Huang, H., Zhuo, W., Niu, Q., Wu, S., Ma, H., and Liang, S.: Progress and perspectives in data assimilation algorithms for remote sensing and crop growth model, Science of Remote Sensing, 10, 100146, https://doi.org/10.1016/j.srs.2024.100146, 2024.

Huang, Y. and Liu, Z.: Improving Northeast China's soybean and maize planting structure through subsidy optimization considering climate change and comparative economic benefit, Land Use Policy, 146, 107319, https://doi.org/10.1016/j.landusepol.2024.107319, 2024.

Hunt, M. L., Blackburn, G. A., Carrasco, L., Redhead, J. W., and Rowland, C. S.: High resolution wheat yield mapping using Sentinel-2, Remote Sensing of Environment, 233, 111410, https://doi.org/10.1016/j.rse.2019.111410, 2019.

Jacquemoud, S., Verhoef, W., Baret, F., Bacour, C., Zarco-Tejada, P. J., Asner, G. P., François, C., and Ustin, S. L.: PROSPECT+SAIL models: A review of use for vegetation characterization, Remote Sensing of Environment, 113, S56–S66, https://doi.org/10.1016/j.rse.2008.01.026, 2009.

Jain, A. K. and Dubes, R. C.: Algorithms for clustering data, Technometrics, 32, 227–229, 1988.

Kaur, S. and Singh, M.: Modeling the crop growth - A review, MAUSAM, 71, 103–114, 2020.

Knyazikhin, Y. ;, Glassy, J. ;, Privette, J. L. ;, Tian, Y. ;, and Running, S. W. ; MODIS Leaf Area Index (LAI) and Fraction of Photosynthetically Active Radiation Absorbed by Vegetation (FPAR) Product (MOD15) Algorithm Theoretical Basis Document, 2018.

Li, C., Ma, C., Cui, Y., Lu, G., and Wei, F.: UAV Hyperspectral Remote Sensing Estimation of Soybean Yield Based on Physiological and Ecological Parameter and Meteorological Factor in China, J Indian Soc Remote Sens, 49, 873–886, https://doi.org/10.1007/s12524-020-01269-3, 2021.

Li, X., Chen, M., He, S., Xu, X., He, L., Wang, L., Gao, Y., Tang, F., Gong, T., Wang, W., Xu, M., Liu, C., Yu, L., Liu, W., and Yang, W.: Estimation of soybean yield based on high-throughput phenotyping and machine learning, Front. Plant Sci., 15, 1395760, https://doi.org/10.3389/fpls.2024.1395760, 2024.

Liu, X. and Herbert, S. J.: Fifteen years of research examining cultivation of continuous soybean in northeast China: A review, Field Crops Research, 79, 1–7, https://doi.org/10.1016/S0378-4290(02)00042-4, 2002.

Liu, X., Jin, J., Herbert, S. J., Zhang, Q., and Wang, G.: Yield components, dry matter, LAI and LAD of soybeans in Northeast China, Field Crops Research, 93, 85–93, https://doi.org/10.1016/j.fcr.2004.09.005, 2005.

Liu, X., Jin, J., Wang, G., and Herbert, S. J.: Soybean yield physiology and development of high-yielding practices in Northeast China, Field Crops Research, 105, 157–171, https://doi.org/10.1016/j.fcr.2007.09.003, 2008.

Mei, Q., Zhang, Z., Han, J., Song, J., Dong, J., Wu, H., Xu, J., and Tao, F.: ChinaSoyArea10m: a dataset of soybean-planting areas with a spatial resolution of 10 m across China from 2017 to 2021, Earth Syst. Sci. Data, 16, 3213–3231, https://doi.org/10.5194/essd-16-3213-2024, 2024.

Misaal, M. A., Zahra, S. M., Rasul, F., Imran, M., Noor, R., and Fahad, M.: Influence of Climate Change on Crop Yield and Sustainable Agriculture, in: Climate Change Impacts on Natural Resources, Ecosystems and Agricultural Systems, edited by: Pande, C. B., Moharir, K. N., Singh, S. K., Pham, Q. B., and Elbeltagi, A., Springer International Publishing, Cham, 209–223, https://doi.org/10.1007/978-3-031-19059-9_7, 2023.

Muhuri, A., Goïta, K., Magagi, R., and Wang, H.: Soil Moisture Retrieval During Crop Growth Cycle Using Satellite SAR Time Series, IEEE J. Sel. Top. Appl. Earth Observations Remote Sensing, 16, 9302–9319, https://doi.org/10.1109/JSTARS.2023.3280181, 2023.

National Soil Survey Office: Soil Species of China, China Agriculture Press, Beijing, 924 pp., 1995.

Ntakos, G., Prikaziuk, E., Ten Den, T., Reidsma, P., Vilfan, N., Van Der Wal, T., and Van Der Tol, C.: Coupled WOFOST and SCOPE model for remote sensing-based crop growth simulations, Computers and Electronics in Agriculture, 225, 109238, https://doi.org/10.1016/j.compag.2024.109238, 2024.

Pang, A., Chang, M. W. L., and Chen, Y.: Evaluation of Random Forests (RF) for Regional and Local-Scale Wheat Yield Prediction in Southeast Australia, Sensors, 22, 717, https://doi.org/10.3390/s22030717, 2022.

Pasqualotto, N., Delegido, J., Van Wittenberghe, S., Rinaldi, M., and Moreno, J.: Multi-Crop Green LAI Estimation with a New Simple Sentinel-2 LAI Index (SeLI), Sensors, 19, 904, https://doi.org/10.3390/s19040904, 2019.

Peng, G. and Yili, Z.: Research on Forest Phenology Prediction Based on LSTM and GRU Model, Journal of Resources and Ecology, 14, https://doi.org/10.5814/j.issn.1674-764x.2023.01.003, 2022.

Pinke, Z. and Lövei, G. L.: Increasing temperature cuts back crop yields in Hungary over the last 90 years, Global Change Biology, 23, 5426–5435, https://doi.org/10.1111/gcb.13808, 2017.

Pu, L., Zhang, S., Yang, J., Chang, L., and Bai, S.: Spatio-Temporal Dynamics of Maize Potential Yield and Yield Gaps in Northeast China from 1990 to 2015, IJERPH, 16, 1211, https://doi.org/10.3390/ijerph16071211, 2019.

Qiao, C., Cheng, C., and Ali, T.: How climate change and international trade will shape the future global soybean security pattern, Journal of Cleaner Production, 422, 138603, https://doi.org/10.1016/j.jclepro.2023.138603, 2023.

Qu, H., Li, X., Zhu, H., Wang, L., Qu, B., Wang, Q., Lv, J., Ji, Y., and Jiang, L.: Effects of combination of low temperature and excessive precipitation at seedling stage on soybean yield in high-latitude cold region, Chinese Journal of Ecology, 1–10, 2023.

Ren, P., Li, H., Han, S., Chen, R., Yang, G., Yang, H., Feng, H., and Zhao, C.: Estimation of Soybean Yield by Combining Maturity Group Information and Unmanned Aerial Vehicle Multi-Sensor Data Using Machine Learning, Remote Sensing, 15, 4286, https://doi.org/10.3390/rs15174286, 2023a.

Ren, Y., Li, Q., Du, X., Zhang, Y., Wang, H., Shi, G., and Wei, M.: Analysis of Corn Yield Prediction Potential at Various Growth Phases Using a Process-Based Model and Deep Learning, Plants, 12, 446, https://doi.org/10.3390/plants12030446, 2023b.

Shi, X. Z., Yu, D. S., Warner, E. D., Pan, X. Z., Petersen, G. W., Gong, Z. G., and Weindorf, D. C.: Soil Database of 1:1,000,000 Digital Soil Survey and Reference System of the Chinese Genetic Soil Classification System, Soil Horizons, 45, 129, https://doi.org/10.2136/sh2004.4.0129, 2004.

Song, X.-P., Li, H., Potapov, P., and Hansen, M. C.: Annual 30 m soybean yield mapping in Brazil using long-term satellite observations, climate data and machine learning, Agricultural and Forest Meteorology, 326, 109186, https://doi.org/10.1016/j.agrformet.2022.109186, 2022.

Steduto, P., Hsiao, T. C., Raes, D., and Fereres, E.: AquaCrop—The FAO Crop Model to Simulate Yield Response to Water: I. Concepts and Underlying Principles, Agron.j., 101, 426–437, https://doi.org/10.2134/agronj2008.0139s, 2009.

Sun, X., Li, Q., Qiao, Y., Hu, Z., Zhang, X., and Liu, Y.: Warming and Drought in Hailun of Heilongjiang: Effects on Growth and Development of Soybean, Chinese Agricultural Science Bulletin, 38, 27–33, 2022.

Tan, J., Yang, P., Liu, Z., Wu, W., Zhang, L., Li, Z., You, L., Tang, H., and Li, Z.: Spatio-temporal dynamics of maize cropping system in Northeast China between 1980 and 2010 by using spatial production allocation model, J. Geogr. Sci., 24, 397–410, https://doi.org/10.1007/s11442-014-1096-0, 2014.

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

Urda, C., Rezi, R., Varga, A. G., Negrea, A., Muntean, E., Sopterean, L., and Duda, M. M.: EXPLORING THE IMPACT OF SOWING DATES ON SOYBEAN YIELD, SEED QUALITY AND TRYPSIN INHIBITOR ACTIVITY, AGROLIFE SCIENTIFIC JOURNAL, 13, 223–230, 2024.

Von Bloh, M., Nóia Júnior, R. D. S., Wangerpohl, X., Saltık, A. O., Haller, V., Kaiser, L., and Asseng, S.: Machine learning for soybean yield forecasting in Brazil, Agricultural and Forest Meteorology, 341, 109670, https://doi.org/10.1016/j.agrformet.2023.109670, 2023.

Wang, B., Chen, C., Liu, D., Asseng, S., Yu, Q., and Yang, X.: Effects of climate trends and variability on wheat yield variability in eastern Australia, Clim. Res., 64, 173–186, https://doi.org/10.3354/cr01307, 2015.

Wang, C., Linderholm, H. W., Song, Y., Wang, F., Liu, Y., Tian, J., Xu, J., Song, Y., and Ren, G.: Impacts of Drought on Maize and Soybean Production in Northeast China During the Past Five Decades, IJERPH, 17, 2459, https://doi.org/10.3390/ijerph17072459, 2020.

Wang, X., Zhu, L., Hao, Y., Wang, Z., Xue, L., Ding, K., and Huang, X.: Impacts of aerosol meteorological feedback on China's yield potential of soybean, Meteorological Applications, 31, e2198, https://doi.org/10.1002/met.2198, 2024.

Xie, Q., Dash, J., Huete, A., Jiang, A., Yin, G., Ding, Y., Peng, D., Hall, C. C., Brown, L., Shi, Y., Ye, H., Dong, Y., and Huang, W.: Retrieval of crop biophysical parameters from Sentinel-2 remote sensing imagery, International Journal of Applied Earth Observation and Geoinformation, 80, 187–195, https://doi.org/10.1016/j.jag.2019.04.019, 2019.

Xie, Y. and Huang, J.: Integration of a Crop Growth Model and Deep Learning Methods to Improve Satellite-Based Yield Estimation of Winter Wheat in Henan Province, China, Remote Sensing, 13, 4372, https://doi.org/10.3390/rs13214372, 2021.

Xu, J., Du, X., Dong, T., Li, Q., Zhang, Y., Wang, H., Xiao, J., Zhang, J., Shen, Y., and Dong, Y.: NortheastChinaSoybeanYield20m: an annual soybean yield dataset at 20 m in Northeast China from 2019 to 2023, https://doi.org/10.5281/ZENODO.14263102, 2024.

Yang, S., Hu, L., Wu, H., Ren, H., Qiao, H., Li, P., and Fan, W.: Integration of Crop Growth Model and Random Forest for Winter Wheat Yield Estimation From UAV Hyperspectral Imagery, IEEE J. Sel. Top. Appl. Earth Observations Remote Sensing, 14, 6253–6269, https://doi.org/10.1109/JSTARS.2021.3089203, 2021.

Yildirim, T., Moriasi, D. N., Starks, P. J., and Chakraborty, D.: Using Artificial Neural Network (ANN) for Short-Range Prediction of Cotton Yield in Data-Scarce Regions, Agronomy, 12, 828, https://doi.org/10.3390/agronomy12040828, 2022.

Yu, Q., You, L., Wood-Sichra, U., Ru, Y., Joglekar, A. K. B., Fritz, S., Xiong, W., Lu, M., Wu, W., and Yang, P.: A cultivated planet in 2010 – Part 2: The global gridded agricultural-production maps, Earth Syst. Sci. Data, 12, 3545–3572, https://doi.org/10.5194/essd-12-3545-2020, 2020.

Zhang, Y., Liu, M., Kong, L., Peng, T., Xie, D., Zhang, L., Tian, L., and Zou, X.: Temporal Characteristics of Stress Signals Using GRU Algorithm for Heavy Metal Detection in Rice Based on Sentinel-2 Images, IJERPH, 19, 2567, https://doi.org/10.3390/ijerph19052567, 2022.

Zhao, G., Wang, J., Fan, W., and Ying, T.: Vegetation net primary productivity in Northeast China in 2000-2008: Simulation and seasonal change, Ying yong sheng tai xue bao = The journal of applied ecology / Zhongguo sheng tai xue xue hui, Zhongguo ke xue yuan Shenyang ying yong sheng tai yan jiu suo zhu ban, 22, 621–30, 2011.

Zhao, J., Wang, C., Shi, X., Bo, X., Li, S., Shang, M., Chen, F., and Chu, Q.: Modeling climatically suitable areas for soybean and their shifts across China, Agricultural Systems, 192, 103205, https://doi.org/10.1016/j.agsy.2021.103205, 2021.

Zhao, J., Wang, Y., Zhao, M., Wang, K., Li, S., Gao, Z., Shi, X., and Chu, Q.: Prospects for soybean production increase by closing yield gaps in the Northeast Farming Region, China, Field Crops Research, 293, 108843, https://doi.org/10.1016/j.fcr.2023.108843, 2023a.

Zhao, L., Li, Q., Chang, Q., Shang, J., Du, X., Liu, J., and Dong, T.: In-season crop type identification using optimal feature knowledge graph, ISPRS Journal of Photogrammetry and Remote Sensing, 194, 250–266, https://doi.org/10.1016/j.isprsjprs.2022.10.017, 2022.

Zhao, Y., Han, S., Zheng, J., Xue, H., Li, Z., Meng, Y., Li, X., Yang, X., Li, Z., Cai, S., and Yang, G.: ChinaWheatYield30m: a 30 m annual winter wheat yield dataset from 2016 to 2021 in China, Earth Syst. Sci. Data, 15, 4047–4063, https://doi.org/10.5194/essd-15-4047-2023, 2023b.

Zheng, L. and Zhang, X.: Harvest time monitoring data of Shengyang Station in Liaoning Province from 1998 to 2008, National Ecosystem Science Data Center, https://doi.org/10.12199/nesdc.ecodb.mon.2020.dp2011.sya.004., 2021.

Zhuo, W., Fang, S., Gao, X., Wang, L., Wu, D., Fu, S., Wu, Q., and Huang, J.: Crop yield prediction using MODIS LAI, TIGGE weather forecasts and WOFOST model: A case study for winter wheat in Hebei, China during 2009–2013, International Journal of Applied Earth Observation and Geoinformation, 106, 102668, https://doi.org/10.1016/j.jag.2021.102668, 2022.

Zhuo, W., Huang, H., Gao, X., Li, X., and Huang, J.: An Improved Approach of Winter Wheat Yield Estimation by Jointly Assimilating Remotely Sensed Leaf Area Index and Soil Moisture into the WOFOST Model, Remote Sensing, 15, 1825, https://doi.org/10.3390/rs15071825, 2023.